Information Logistics

A Data Integration Method for Solving Data Quality problems with article information in Large Interorganizational Networks

Bas H.P.J. Vermeer

b.h.p.j.vermeer@tm.tue.nl / bvermeer@bakkenist.nl

Abstract

Poor interpretation plays an important role in data quality problems with article information for large interorganizational networks. In the database literature these problems are solved through data integration. However, traditional data integration approaches such as view integration do not work in large interorganizational networks because (1) it is nearly impossible to integrate hundreds of independent schemas, and (2) data distribution (which means: getting the right article data at the right time at the right place) cannot be solved using the normal automatic replication mechanism, because the replication responsibility lies at each individual participant in the network. This means that the strength of the replication mechanism depends on the weakest link in the network. In this paper, we will discuss three modern data integration approaches, namely, the tight coupling approach, the loose coupling approach and the Context Mediation (CM) approach. Although each approach solves parts of the translation and distribution problems that arise in interorganizational data integration, none of them is completely sufficient. Therefore, we will introduce Information Logistics, a method that extends the CM approach with a distribution mechanism to solve the data distribution problem in large interorganizational networks.

1. Introduction

Poor interpretability plays an important role in explaining interorganizational data quality problems with article information. A short field study in the Dutch food sector, showed that many problems with scanning and EDI resulted from poor quality of the article data. Especially *indistinctnesses about packing units*, and *problems due to differences in internal codes and standard EAN product codes* were mentioned. (Vermeer 1996). These problems are typical interpretation problems.

This finding is supported by Strong, Lee and Wang (1997), who identified poor interpretability as an important root cause for data quality problems. Madnick (1995) also recognizes the importance interpretation problems in global information systems. In the database literature, this type of data quality problems are normally solved through data integration. Data integration generally means the standardization of data definitions and structures through the use of a common conceptual schema (Heimbinger and McLeod 1985, Litwin et. al. 1990). When data is integrated, every user knows how to interpret the data, thus preventing data quality problems.

However, in complex interorganizational networks, where many data suppliers and receivers exchange similar information frequently (for instance, in large food supply networks with hundreds of suppliers for one retailer), two problems exist that prevent good quality data to be available to the user. Firstly, traditional data integration approaches such as the view integration method of Batini et. al. (1986), are not applicable in these complex interorganizational networks, because the integrated data model (which integrates hundreds of individual corporate models) are too complex and in the end have no flexibility whatsoever. Secondly, even if the meaning of the data would be perfectly clear to all users, in a large interorganizational network there is the relatively new problem of *data distribution*. With the data distribution problem we mean: how to get the

right (article) data, at the right time, at the right place. In table 1, we provide an example of why this problem exists using business cards.

Table 1: Business cards example

Similar to an integrated data model, a business card contains highly standardized information that is used by many different users. However, as many have experienced, after some time most cards they possess contain outdated, mostly invalid information. This happens because most people do not send updates, simply because they do not remember who they gave their cards to. Therefore, opposite to what is generally assumed, the existence of a standard in itself is not enough to guarantee integrated data. We also need a mechanism to get the right data at the right time at the right place.

This problem hardly exists in a single organization, where this problem is solved by the automatic replication facilities of most large vendor database packages. This problem also hardly exists in large networks, where human users occasionally seek information at a remote site, and therefore do not need the right information at the right time at the right place, since they get it themselves. However, in large interorganizational networks, where for instance inventory management applications use the same article information frequently to reorder products, this problem becomes manifest.

In this paper, we will introduce Information Logistics as an alternative solution for interorganizational data integration to solve data quality problems. Information Logistics extends the principles in the Context Mediation approach (Goh et. al. 1994) through adding data distribution mechanism. In the first section we will first introduce the problem of data integration. Next, we will present an abstract problem description that explains the structure of data integration problems in multi database situations. In the third section, we will introduce three data integration approaches respectively, explain how they work using the abstract problem description, and discuss the problems of these approaches. In the fourth section, we will present the Information Logistics (IL) approach as an alternative solution for solving the data integration problem for

large interorganizational networks. Finally, in the last section. we will evaluate the IL approach together with the first three approaches.

2. The problem of data integration

Goodhue et. al. (1992) provide a comprehensive model that explains the problem of data-integration (see figure 1).

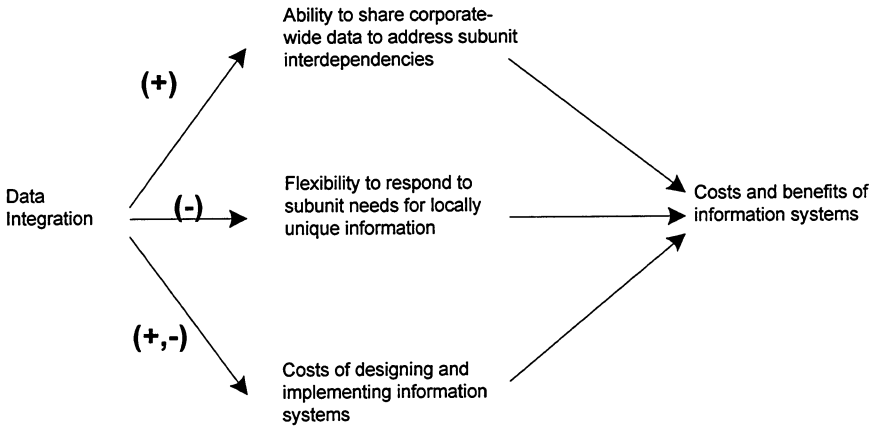

Figure 1: The data integration problem (Goodhue et. al. 1992)

In their paper, Goodhue et al. argue that data integration will have a positive effect in organizational situations where subunits are highly interdependent. Data-integration leads to improved coordination and less costs, because no ambiguous messages between subunits are exchanged. However, data integration has a negative effect on situations where subunit tasks are non-routine or the environment is unstable. Data integration requires that all subunits use the same, standardized agreements on data (the same data model). This decreases the ability of subunits to meet their specific information needs, and therefore it lowers the level of local flexibility of subunits. Finally they argue that data integration may affect the costs of designing and implementing information systems either way. Normally, a more expensive initial design leads to less costs for subsequent modifications. However, *as the number and heterogeneity of subunit information needs*

increase, the difficulty of arriving at acceptable design compromises increases and therefore the initial design costs will increase more than linearly. This same effect will appear in later modifications, thereby increasing the long-term costs.

The importance of Goodhue's model is that it explains why data-integration is a problem, especially in an interorganizational context. Firstly, the use of common field definitions and codes[1] (that is: a common data model) means that no ambiguous messages are exchanged between interdependent locations across an interorgazanitional network. On the other hand, the implementation of a common data model across many interdependent network participants is practically impossible. Firstly, because the construction of such a data model from all the participants models would take many years of work. Secondly, because such a common data model will virtually destroy the flexibility of all participant's organizations to address the needs for locally unique information. Finally, because such complex data models are practically non-maintainable (Pels 1988).

3. Abstract problem description

To understand and compare the different data integration approaches, we will first describe the problem of interorganizational data integration in terms of an abstract problem description. In this description we first make a distinction between data integration and data distribution and we explain that data integration is actually a translation problem. Next, we identify three levels of agreements that are necessary to make a successful translation possible. We will use the distinction between translation on three levels and distribution to describe and evaluate the different data integration approaches.

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

3.1. The distinction between translation and distribution

In a single database situation, real world facts are stored in a single database. This database basically consists of two parts: a database schema, describing the data in the database and the data itself (Date 1990 pp. 38, Elmasri & Navathe 1994, pp.23-28). The schema describes the structure of the data. It describes the real world entities the database recognizes and its attributes. Furthermore, the schema defines which entities are related to each other, what the types of their relationships are and what constraints apply. The data in the database consists of the actual values of the facts that are presented to the database. The single database situation is shown in Figure 2.

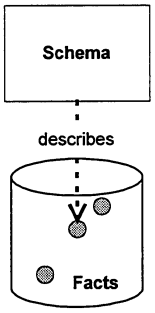

Figure2: Single database situation

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

In a multiple database situation, the same fact is distributed over many different locations, described by different schemas at each location (with distributed we mean that users at different locations either have that

fact in their own database or maintain an active link to the source). Using the schema/ data distinction, we may represent the multiple database situation as shown in figure 3.

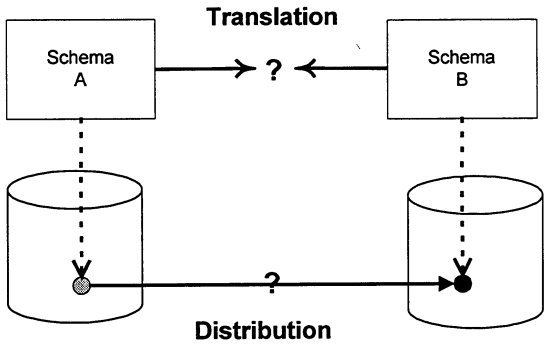

Figure 3: Multiple database situation

As we can see from figure 3, two problems arise in a multiple database situation: a translation problem and a distribution problem. The translation problem arises because the same fact is differently structured at different locations. Therefore, schema translation is necessary to map the structure of the source schema to the structure of the receivers' schema. This results in a mapping schema between the source and the receivers' schemas that is used every time a fact in the source database is updated.

The distribution problem arises because each fact update is translated and transported over an imperfect network to a limited set of users. During translation mapping errors may occur, which results in loss of data quality. During transportation, the data may get delayed, damaged, or delivered to the wrong recipient, resulting in inconsistencies among different locations.

3.2. Different levels of agreement

The translation problem is the result of differences between user contexts. Therefore, when an update is sent to the receiver, the translation program at the receiver needs to know many things: What is the message about? Who is it from? What does the information in the update mean? Which data fields do I have to update

and how? If the receiver knows the sender and has made agreements about what kind of updates can be expected and therefore knows how to react on them, the translation effort will be relatively simple. Therefore, depending on the degree of mutual understanding between sender and receiver this translation effort will be either simple or difficult.

Stamper offers a model that helps to understand the different degrees of mutual understanding between different contexts. (Stamper & Huang 1994, Stamper 1995). This framework describes six levels of communication between two subjects (persons or organisations) that operate in a different context. When messages (for instance, EDI messages) are exchanged, agreements on the first five levels are necessary to guarantee successful communication (on the sixth level). A slightly adapted version of the Semantic Framework for EDI is shown in figure 4.

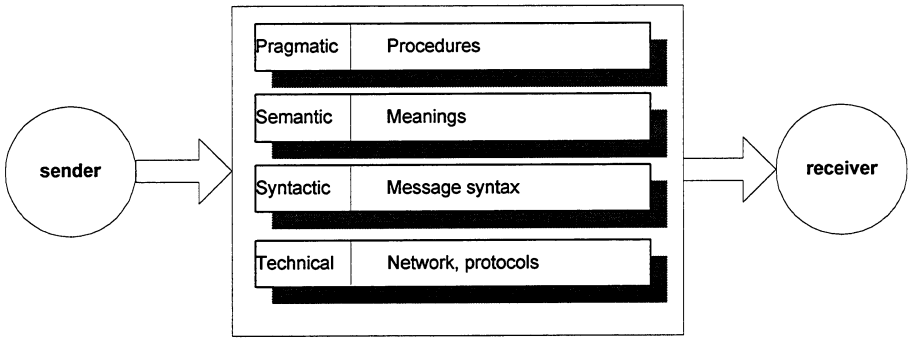

Figure 4. Semiotics Model for EDI (adapted from Stamper & Huang 1994)

On the first level agreements on the technical connection are established: these agreements specify the type of network connection and the network protocols that are to be used. On the second level, agreements on the syntax are established. Syntax relates to the structure of a message: which data-elements can be used, and how and in which order are they displayed (for example: name, address, domicile). The syntactic level can be compared with a grammar dictionary that specifies which words are available in a language and how sentences are constructed. The semantic level deals with agreements about the *meaning* of data. Here, the

relation between data elements with other elements is specified and constraints are defined. This level can be compared with data dictionaries in Database Management Systems, where the conceptual data model is constructed and agreements about for instance the range of product numbers or the definition of turnover are established. Finally, on the pragmatic level agreements on the *intention* of a message are established. For instance, when company A sends an order, company B must understand that company A wants one of their products. Furthermore, they must understand that company A wants them to react through returning an order confirmation. Thus, on this level, procedures are established. For understanding and describing the different data integration and data distribution approaches, we will use Stamper's framework to understand how, and to which level, translation within a certain approach is established.

4. Three data integration approaches

In the database literature, the problem of data(base) integration between multiple databases is referred to as *multidatabase*, *heterogeneous* or *federated* (Heimbinger & LcLeod 1985,) database systems. Sheth & Larson present a taxonomy of these systems (Sheth & Larson 1990). In this paper, we are interested in federated database systems, since these systems are autonomous, while at the same time they participate in a federation to allow partial and controlled sharing of data.

Federated Database Systems (FDBS) can be categorized as loosely or tightly coupled. An FDBS is *loosely coupled* if it is the user's responsibility to create and maintain the federation and there is no control enforced by the federated system and its administrators. A federation is *tightly coupled* if the federation and its administrators have the responsibility for creating and maintaining the federation and actively control the access to component DBSs. We will now examine how data integration is established within the tightly and loosely coupled systems.

4.1. Tight coupling approach

The tight coupling approach basically consists of three steps. In the first step, the local database schemas are translated into component schemas expressed in the Common Data Model (CDM). The CDM describes the local schema in a single database language. In the second step, the component schemas are integrated into one or more federated schemas. The integration of the component schemas in one federated schema is normally established through view or schema integration (Batini et. al. 1986). This procedure compares the different component schemas, through identifying naming conflicts and structural conflicts. When schemas are compared and differences are detected, the difference must be resolved, after which the schemas can be integrated in one federated schema. In the third step, the transformations between local, component and federated schemas are constructed. This means that the mappings between the different schemas are generated together with an appropriate distribution or allocation schema. This schema contains information about the distribution of the data among different locations. Each time data is sent from one location to others, the mappings assure that the data is translated to the right context while the distribution schema assures that the data is sent to the right locations.

In terms of the abstract problem description, the *translation* problem in the tight coupling approach is solved through the definition of a single schema through view integration. This single, federated schema is used in the translation of messages between different locations, resulting in unambiguous message exchange. The *distribution* problem is solved through view updating (Batini et. al. 1992), or update synchronization (Ricardo 1990, pp. 511). View updates are used to synchronize multiple copies over different locations, thus providing fast access at multiple sites to the same remote data. The view updating task is relatively easy in the tight coupling approach, since the distribution schema describes where the copies are stored, and the mappings between the schemas translate the updates to the right contexts during the update process.

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

4.2. Loose coupling approach

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

In terms of the abstract problem description, the loose coupling approach leaves both the *translation* and *distribution* problem to the user. Firstly, the user must understand the semantics of the location where he retrieves his data to formulate a valid query. Secondly, the loose coupling approach requires that the user knows where the data that he wants to retrieve is located. Therefore, the loose coupling approach is specifically appropriate for situations where many different database systems are connected with each other where each location occasionally needs information from another location.

4.3. Context mediation approach

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

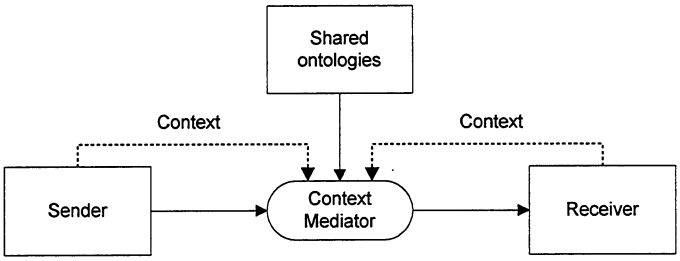

Figure 5: Architecture Context Mediation

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

To make these transformations, the context mediator is based on a *shared ontology* of the financial service community. This shared ontology (or domain model as Goh. al. describe it in the 1999 paper) basically describes how this community views the structure of financial information of different companies in

different countries. Using this shared structure, Goh. et. al. (1994,1999) are able to map the schemas of the local databases on each other and to translate different currencies and scalefactors.

In terms of the abstract problem definition, the *translation* problem in the CM approach is solved through the definition of the shared ontology. If we look at this shared ontology as a unified schema description of the financial service community, we may conclude that the CM approach is based on a specific type of federated schema. This schema is more sophisticated than traditional schemas, because it separates the meanings of concepts such as *currency* or *country* from their context specific values (e.g. US$ or The Netherlands), hence producing a semantic richer schema. Although the CM approach also uses a shared schema, this schema is not developed from the component schemas through a very labor-intensive process, as is the case in the tight coupling approach. Rather, it is defined independently of the component schemas, based on the needs of a particular community. Thus, the intensive integration effort of the tight coupling approach is reduced.

With respect to the *distribution* problem, the CM approach leaves the collection of remote data to the user. Thus, the user must contact the right data sources himself to retrieve the required data. Hence, we may conclude that the CM approach lies between the tight and loose coupling approaches. It resembles the tight coupling approach in that it provides unambiguous message exchange. On the other hand, it resembles the loose coupling approach in that it is flexible in connecting to other locations to retrieve remote data typically for one time use.

4.4. Analysis three integration approaches

Table 2 gives an overview of the different data integration approaches in terms of the abstract problem description.

Table 2: Data integration approaches

Problem			Tight coupling	CM approach	Loose coupling
Translation problem	How?		view integration + mapping	central defined ontology + mapping	'flat' mapping
	Level?	Pragmatic			
		Semantic			
		Syntactic			
Distribution problem	How?		define replication schema	X	X

4.4.1 Translation problem

With respect to the translation problem, the tight coupling approach uses view integration to construct a unified schema of all local database schemas. Translation is accomplished through mapping the between the local schema and the central schema. In terms of Stamper's model, the tight coupling approach solves both the syntactic and the semantic problems of communication. However, the pragmatic aspects are not solved.

The loose coupling approach solves the translation problem through translating the data elements from queries from one context to the data elements of another context. We refer to this direct mapping between data elements, without using a semantic structure describing the meaning of the elements, as 'flat' mapping. Since flat mapping only translates between syntaxes, in terms of Stamper's model, the loose coupling approach does not solve the semantic and pragmatic problems of communication.

The CM approach solves the translation problem through the construction of a centrally defined ontology. This ontology describes the 'general' business structure in a specific application domain. Through mapping from the local schemas to this central schema, the translation problem is solved. In terms of Stamper's model the CM approach solves both the syntactic and semantic problems of communication, just as the tight coupling approach.

4.4.2 Distribution problem

The distribution problem is only addressed in the tight coupling approach. Through the definition of a replication schema, which describes which data is used by whom, the data is distributed to the different user databases. This approach provides a good solution whenever a limited number of databases is involved. When a large number of users is involved who are independent of each other, distribution depends on the discipline of each independent user. In large networks, this normally leads to poor distribution performance. The other two integration approaches do not solve the distribution problem.

5. Information Logistics

To solve the interorganizational data quality problem for article information, we need a method that supports the development of central agreements on user community level about the structure of article data and that aligns the data across many independent databases across an interorganizational network. Such a network can be very large. For instance, in the Dutch retail sector, about 4000 suppliers deliver their products to 40-100 retailers.

From the data integration approaches we discussed, the CM approach with the concept of defining a central ontology offers a good solution for the translation problem. However, the distribution problem is insufficiently solved in all three approaches. Therefore we propose a new method for data integration, specifically appropriate for large interorganizational business networks, which we named Information Logistics (IL). The IL method uses the ontology concept of the CM approach to solve the translation problem and concepts of Logistics to solve the distribution problem. We will discuss how the IL method solves the translation and distribution problem below.

5.1. IL and the translation problem

The translation problem in the IL method is solved through the definition of an Information Product (IP). The concept of an IP is similar to the Information Product introduced by Wang (1998) in the TDQM approach. We will define an IP as:

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

The IP contains the central sectorwide agreements, which are based on a *user defined* ontology. This means that user groups on sector level determine the contents of the IP. We use the term product because the concept of a product emphasizes the relation with customers or *users,* which need a product for a specific *purpose*. The specification of the purpose of the IP is important because it limits the amount of information that will be contained within the IP. Furthermore, products have *owners* who manufacture the products and *distribute* them to the users. We will return to the distribution property of products in the next paragraph.

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

food ingredients, their expected effects and their way of preparation are related to each other. Together these IPs form the shared ontology for a specific user community as was described by Goh et. al. (1994).

An important characteristic of the IL method is that the IP is not only constructed but also *maintained* by the specific user community. For instance, for the food retail community in the Netherlands, both manufacturers and retailers have established a special working group that defines IPs under the auspices of the Dutch EAN organization. This organization is responsible for the EAN article numbering system (e.g. the bar codes on consumer products) and implements EDI message standards for all sectors in the Netherlands. Apart from defining IPs, this working group is also responsible for maintaining the IP, since changes may occur that require restructuring of the IP.

A second important characteristic of the IL method is that the users themselves are responsible for making the mappings between the component data models of each user to the central schema (the IP). In both the tight coupling approach and the CM approach the architects make this mapping. The main advantage is that this delegation of the mapping task reduces the complexity of this task. For instance, in the food retail sector in the Netherlands, about 1200 component supplier data models and 50 component retail data models have to be mapped to the central schema. For a single architect, this task is virtually impossible to perform but is relatively easy performed by the IT departments of the 1200 suppliers and 50 retailers. The reader is reminded that the central schema is not prescriptive in that it forces the participants to adapt their component schemas to the central schema structure. Rather, the IT departments of the participants may structure their component schemas in whatever way they want. They only have to make sure that externally communicated information is structured according to the central schema. In other words, each participant may speak its own language at home as long as they speak the centrally defined language when they work together.

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

5.2. IL and the distribution problem

IL specifically aims to solve the distribution problem in a large interorganizational network. The distribution problem arises because updates of product information in source databases need to be distributed to a limited number of receiver's databases, which are actually users of the information. Thus, the right information needs to be at the right time at the right place.

In IL, this problem is resolved through the definition of an IP distribution structure. This structure determines how information from data senders is send to data receivers, taking into account the variety of user and sender requirements. Users typically want to specify which information they want to receive, when, in which format, and having what quality. Senders typically want to control which users receive what information. The structure also takes into account how the central schema (the IP) is maintained and where conformance checking takes place (centrally or locally).

Taking a systems approach, we can define the goals, the inputs and the control variables that determine the design of the IP distribution structure (see Figure 6).

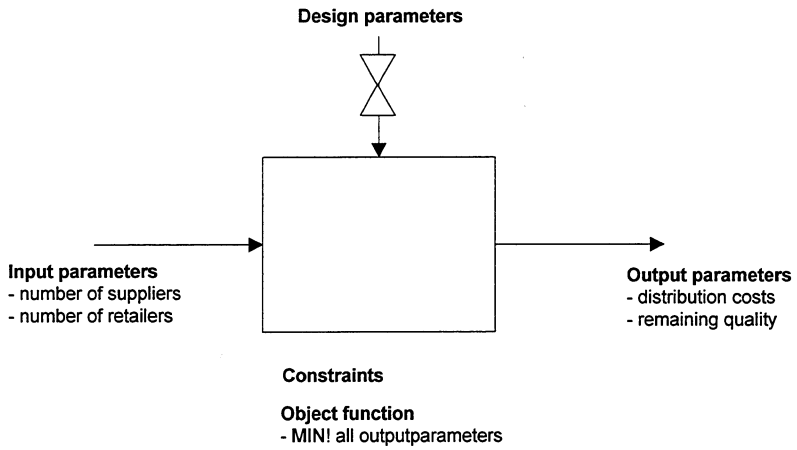

Figure 6: An information distribution structure

We will discuss the model shortly. The objective is to determine how the IP distribution network must be designed to minimize parameters such as the costs of distributing instances of the IP with a maximum of data quality in a network consisting of a limited number of suppliers and retailers under a number of constraints. Possible design parameters may be: (1) centralized (via a central data cross docking center) or bilateral data distribution) to save the costs of distribution, and/or (2) the availability of a central data quality conformance testing service to improve overall data quality. Currently, we are investigating which design parameters determine the structure of such a distribution network, and how these parameters should be set to determine an optimal structure for different types of distribution situations.

Compared to the other approaches we discussed, the IL method is specifically concerned with the implementation of a distribution structure to guarantee the synchronization of product information across databases in an interorganizational network.

6. Evaluation data integration approaches

The problem of data-integration in an interorganizational context is that an interorganizational situation requires both sharing of data to address subunit interdependencies and high flexibility to address local unique information needs.

In the literature we found three data integration approaches, that provide different solutions for the data translation and data distribution problems. To understand the applicability of these approaches in an interorganizational context, we will use the parameters of Goodhue's model to evaluate them. These parameters are respectively:

1. The ability to share data;
2. The flexibility to introduce local design changes;
3. The costs of integration.

The three parameters above only address the problem of data integration. Therefore we add to more parameters from the perspective of the distribution problem:

4. The flexibility to change the replication schema;
5. The remaining quality of the data.

We have used these parameters to evaluate the three approaches for data integration and distribution including the Information Logistics approach. The results are shown in table 3.

Table 3: Evaluation

Problem			Tight coupling	CM Approach	Loose coupling	Information Logistics
Trans-Lation problem	How		View integration + mapping	Central ontology + mapping	Flat mapping	User ontology + mapping
	Level	Pragmatic				
		Semantic				
		Syntactic				
Distribution problem	How		Define replication schema	X	X	Distribution structure
Integration costs			--	+/-	++	+
Design flexibility			--	+	++	+
Ability to share data			++	++	--	++
Replication flexibility			--	NA	NA	++
Quality			+	+	--	++

The upper part of table 3 defines each approach in terms of the abstact problem description. The lower part of table 3 shows the results of the evaluation.

With respect to the *costs of integration*, the tight coupling approach is clearly the most expensive, because integrating the schemas of more than a few hundred participants is virtually impossible. The costs of integration for the CM approach are also quite high, because the CM architect has to map all the local schemas to the central schema. In an interorganizational situation with a few hundred participants, the architect will need a lot of time to first understand each local schema and then translate it to the central schema. The integration costs of the IL approach are lower, because the mapping to the central schema is performed by each local participant, who have to learn only one schema, namely the central one. However, since the mapping is performed by more than one person, this may result in more errors, which means that data quality becomes more important. The integration costs for the loose coupling approach are the lowest, since no integration is needed whatsoever. In this approach, the human users solve all translation problems.

With respect to the *flexibility to introduce local design changes*, again the tight coupling approach scores very low, since every local design change has to be analyzed and approved at the central level. Because both

the CM aaproach and IL use a central ontology, the local design changes are decoupled from the central schema. Thus, at the local level many design changes can be implemented, without affecting the central schemas. The only consequence is that the mappings from the local schemas to the central schema must be re-established. The loose coupling approach has the highest score, since schema changes do not affect the mechanism of the translation process in the loose coupling approach (except when the syntax of the schemas is changed).

With respect to the *ability to share data,* the loose coupling approach has toe lowest score, because it does not support the semantic level of communication. Since the other approaches use a semantic schema, their ability to share data is high.

With respect to the *flexibility to change the replication schema,* only the tight coupling approach and the IL method have the ability to distribute data. The flexibility for introducing replication changes in the tight coupling approach is very low, since in normal replication schemas changes in the dynamics of the relationships between databases are administrated manually by a central administrator. Because an interorganizational network normally contains many participants, the number of changes will be very high. A central administrator cannot process all these changes within a reasonable time limit. The IL method scores very high, since changes in the dynamics of the relationships are administrated locally by the participant themselves.

Finally, With respect to the *quality of the data,* only the IL method scores very high, because explicit quality checking mechanisms are incorporated within the distribution structure to ensure the delivery of high quality data as requested. The tight coupling approach also has a high score, because normally the existence of one single schema means that many quality problems are prevented through the use of integrity constraints. We assume that having a single schema is equal with having a high degree of integrity. In

practice this is not always the case. The CM and loose coupling approaches do not incorporate any form of quality checking and therefore have a low score on this dimension.

7. Further research

Although we discussed the outline of the IL approach in this paper, the method is still under construction. In an explorative case study where we used this method we found three design parameters that together determined the structure of the distribution network. Currently, we are testing the applicability of the method in a similar case study. This study should provide insight in the applicability of the IL method.

8. References

BATINI C., LENZERINI M. and NAVATHE S.B. (1986) A Comparative Analysis of Methodologies for Database Schema Integration, *ACM Computing Surveys,* (18:4) pp. 323-364

BATINI C., CERI S. and NAVATHE S.B. (1992) *Conceptual Database Design: An Entity-Relationship Approach,* The Benjamin/Cummings Publishing Company, Redwood City, CA

DATE (1990) *An Introduction to Database Systems – Volume 1*, 5th edition, Addison-Wesley Publishing Company, Reading MA

ELMASRI R. and NAVATHE S.B. (1994) *Fundamentals of Database Systems,* Benjamin Cummings Publishing Company, Redwood City, CA

GOH C.H., MADNICK S.E. and SIEGEL M.D. (1994) Context Interchange: Overcoming the challenges of large-scale interoperable database systems in a dynamic environment. *in Proc. 3rd International Conference on Information and Knowledge Management (CIKM-94),* Gaithersburg, Md

GOH C.H., BRESSAN S., MADNICK S.E. and SIEGEL M.D. (1999) Context Interchange: New Features and Formalisms for the Intelligent Integration of Information *ACM Transactions on Information Systems*, Forthcoming July 1999

GOODHUE D.L., WYBO M.D. and KIRSCH L.J. (1992) The impact of data integration on the costs and benefits of information systems. *MIS Quarterly,* (16:3), pp. 293-311

HEIMBIGNER D. and MCLEOD D. (1985) A Federated Architecture for Information Management, *ACM Transactions on Office Information Systems* (3:3) pp. 253-278

LITWIN W., MARK L. and ROUSSOPOULOS N. (1990) Interoperability of Multiple Autonomous Databases, *ACM Computing Surveys* (22:3) PP. 267-293

MADNICK S.E. (1995) Integrating information from global systems: dealing with the on- and offramps of the information superhighway. *Journal of Organizational Computing,* (5:2) pp. 69-82

PELS H.J. (1988) *Integrated Information Bases, Modular Design of the Conceptual Schema.* Doctoral Thesis, H.E. Stenfert Kroese B.V., Leiden-Antwerpen, The Netherlands: In Dutch

RICARDO C. (1990) *Database Systems: Principles, Design, and Implementation,* MacMillan Publishing Company, New York NY

SHETH A.P. and LARSON J.A. (1990) Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases, *ACM Computing Surveys, (22:3)* pp. 183-236

STAMPER R. K. (1995) Signs, Information, Norms and Systems. *in: The Semiotics of the worksplace*, B. Holmqvist and P.B. Andersen (eds.)

STAMPER R.K. LIU K. and HUANG K. (1994) EDI Systems Design from Semiotic Perspective, *Working Paper L248*, TU Twente, The Netherlands

STRONG D.M., LEE Y.W. and WANG R.Y. (1997) Data Quality In Context, *Communications of the ACM*, (40,5), pp. 103-110

VERMEER B.H.P.J. (1996) Articles Are Not Alike! An Investigation of Problems due to Poor Synchronization of Article Master Data in the Grocery Sector, Eindhoven, The Netherlands: *TU Eindhoven Report EUT/BDK/77*, In Dutch.

WANG R.Y. (1998) A Product Perspective on Total Data Quality Management, *Communications of the ACM, (41:2),* pp. 58-65