MASTER DATA MANAGEMENT: PRODUCTS AND RESEARCH

(Practice-oriented Paper)
Master Data Management

Jochen Kokemüller
Fraunhofer IAO, Germany
Jochen.Kokemueller@iao.fraunhofer.de

Anette Weisbecker
Fraunhofer IAO, Germany
Anette.Weisbecker@iao.fraunhofer.de

Abstract: Master Data Management is the discipline of creating and maintaining high value, high quality master data. In this contribution we give a definition of this data category underlining its importance to an overall high level of cooperate data quality. The current situation of commercial master data management solutions is presented. It is based on the results on six systems of two surveys we conducted. Here we discuss the systems capabilities for information integration, data modeling and information quality. After the current situation we provide an outlook on future developments in the area of master data management and discuss the relevance of Peer-To-Peer technologies. On this behalf we go into some detail discussing the specialized architecture VIANA.

Key Words: Master Data Management, MDM, Data Quality, Information Quality, Information Integration, Survey

INTRODUCTION

Enterprises face a growing amount of data. At the same time users require that access to this data returns fast and good results. The problem space is therefore two-dimensional. One dimension is the information availability, especially spanning multiple systems this is actually a problem of information integration. Challenges include mapping and transport of information from one system to the other crossing political and technological boundaries. The second dimension is the fitness of the information for use, the information quality.

The data category that serves as a fundament to business transactions is master data. As this data category experiences a high value the whole discipline of master data management is built around it. In this contribution we have a closer look on how to define master data. Therefore, we categorize enterprise data into three categories and discuss their respective properties. Here we define master data in its relationship to the other categories. We show that high quality data in an enterprise can only be achieved, if the underlying master data is of high quality. We therefore understand master data management as a key discipline and the fundament to achieve and maintain high quality enterprise data. Not only for this reason Gartner expects a significant grow of this market segment from $1 billion in 2007 to $2.8 billion in 2012, despite the current economic gloom [7].

Several vendors offer Master Data Management (MDM) systems. In this contribution we compare six systems and show the state-of-the-art of commercial MDM systems. We discuss which aspects work as a common denominator and therefore are defining aspects to the industry.

The remainder of this contribution is organized as follows. First we give a categorization of enterprise data, before we describe possible architectures for master data management solutions. We then analyze the current situation of the MDM industry by providing results of two surveys. We then continue with an analysis of possible future developments by presenting research activities in this area. Here we discuss the Peer-To-Peer information integration architecture VIANA before we conclude.

Categorization of enterprise data

Prior to discussing the categorization we introduce the concept of the active phase in the life time of a data entity. In this active phase the object is involved in an active process and is accessed and changed frequently. Afterwards the object enters the passive phase where it still needs to be accessible, but with a substantially lower access rate. In some cases a data entity begins with a passive phase, enters then the active phase before it returns to the passive phase. Both phases together build the entities entire life span.

Cooperate data can be classified into three categories: transactional data, inventory data and master data (Figure 1). These categories provide valuable insight into several characteristics of enterprise data. Table 1 compares the key characteristics of these categories.

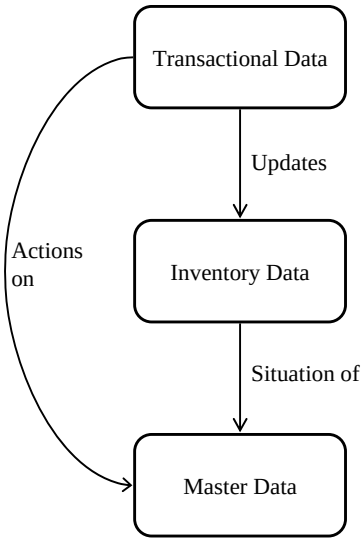

Figure 1: Relationship between data categories

Master data is fundamental to an enterprise. A company invests substantial effort in its acquisition, i.e. acquisition of customers, employees and so on. Master Data may be roughly defined as the fundamental data for transactions that is slowly changing. The data classes most commonly associated with it are products and customers. Nevertheless, accounts, business partners, employees, suppliers and others are also often cited master data classes. In comparison with other data categories they provide only a small share of the data volume present in an enterprise. A company with 1.000 employees is already a big company, yet 1.000 entries in an employee table are little data. The life time of those objects is usually pretty long. In Germany product master data has to be accessible for at least 30 years after production for product liability. The active phase is usually shorter than the overall life time still it is often a long time period. Decades are common for all classes. Changes to master data do not happen very frequently. Yet, as they live very long, the amount of changes done over the entire lifetime accumulates to a noteworthy amount. Master data is often even used across enterprise boundaries, e.g. in product catalog exchange.

Inventory data represents the status of master data. It can be seen as stock amounts of products or leave status of employees. Usually, there is more inventory data than master data because a product may be present in several stocks or an employee went on several leaves. The life time of this data depends by definition on the active phase of master data. It starts with the beginning of the master data's active phase and ends with its lifetime. Likewise, the active life time is equal to the active life time of the master data. The separation of this data class is mainly because it changes frequently, as stock amounts change due to every transaction. Yet, it's the situation of a master data object that changes, not that actual object itself. As a result, inventory data accumulates substantial historic data over its lifetime. Finally, this data category is often used in several IS of one enterprise at the same time.

Transactional data represents business transaction as orders, quotes, invoices or applications for leaves. They represent an action as a sale on a master data object or an event. On completion, the action updates the inventory data, e.g. adjusts the stock amount or the employee's status. For regulatory compliance transactional data has to be archived (e.g. invoices in Germany for 10 years, quotes for 6 years). This is the life time of transactional data. The active phase is significantly shorter. It corresponds to the creation including updates to it in approval or similar processes. This is usually done in a matter of days. As changes are only done in the relatively short active phase, the overall amount of changes is low.

	Real world	**Example**	**Amount**	**Life time**	**Length of active phase**	**Change frequency in active phase**	**Overall amount of changes**
Transactional Data	Action or Event	Order	Huge	Medium	Short	Medium	Low
Inventory Data	Situation	Stock amount	Medium	Equal to active of master data object	Equal to life time	High	Many
Master Data	Object	Customer	Small	Very long	Long	Low	Medium

Table 1: Characteristics of data categories

Master data builds the fundament of most enterprise data. It is easy to see, that the information quality of transactional and inventory data depends directly on the quality of the master data. In the context of Business Intelligence (BI) initiatives it is common to raise the information quality in an Extract-Transform-Load (ETL) process because good business decisions need to be based on good data ("garbage in, garbage out"). Yet, the quality elevated information is often not fed back into operative systems. Especially in large and grown enterprises where one master data class is held in several systems it is often impossible to achieve a single view on one object. As a consequence, a customer might be known in the service and sales departments but one department does not know the history or the existence of the customer to the other department. Similarly, in one system a customer might be present several times and receive multiple copies of one mailing resulting in unnecessary costs and loss in customer satisfaction. [5] [16] give a concise discussion and classification of data quality problems.
The discipline that tackles the creation of an overall information quality for master data is master data management (MDM). Its aim is the "fitness for use" [21] of the data. This is achieved by methods known to the data quality community like reference reconcilitation and data enrichment. Additionally, the target is to always provide access to the most recent information by information integration [14], [9].

Architectures

Three architectural variants are commonly seen in commercial MDM systems. We start with the centralized architecture (Figure 2). In this variant all master data of one or more classes is stored in one system. This system provides then processes for data alteration and monitors data changes. As the integrated system may require that the MDM system provides it with objects using the IDs of the integrated system, the MDM system usually implements a mapping table mapping MDM IDs to foreign IDs. Additionally, the MDM system may keep track of the object's version deployed to the integrated systems. Whether a system keeps track of several versions or just one is vendor dependent, yet always exactly one dataset of an object is marked as the best version. This version is known under several names, the most common are: "Single Version of Truth" and "Gold Copy". All new information is integrated into this version and it is ensured, that this version is always the best in terms of data quality. Starting from it, the views for the integrated systems on the information are calculated and deployed.
The second variant often encountered is a leading system. Here, the data is not integrated in a separate database of an MDM system, but in one of the integrated systems: the leading system. Usually, the system with the highest expressiveness for the particular data class is chosen. The transformations of the MDM Workflow (Data scrubbing, field adjustments, etc.) are then carried out either by the leading system itself or a MDM system wrapping the leading system.
A lightweight way to integrate master data is by using a directory. In the directory references to certain master data objects are stored. The objects themselves remain distributed over several systems. This approach is the only one that offers a purely virtual integration. As a consequence it has the least influence over the data. It may integrate schematic diverging data, cannot enforce data quality algorithms and cannot generate a gold copy.
The last possibility for master data management is a Peer-to-Peer (P2P) based approach. Here the IS is wrapped by peers. They work in a networked structure where all participants are equal with respect to what they are able to do. This Peer-to-Peer collaboration [20] reflects the organizational structure of autonomous enterprises that directly and equitable share information and are responsible for the integration to their neighbors. P2P integration is not to be confounded with simple pair-wise spaghetti integration because it provides functional components for semantical information integration and acts as a framework for efficient implementation of complex information integration scenarios. Peers provide flexibility towards the integrated systems and allow to this side heterogeneous behavior. Additionally, they present themselves organized and homogeneous towards neighboring peers.

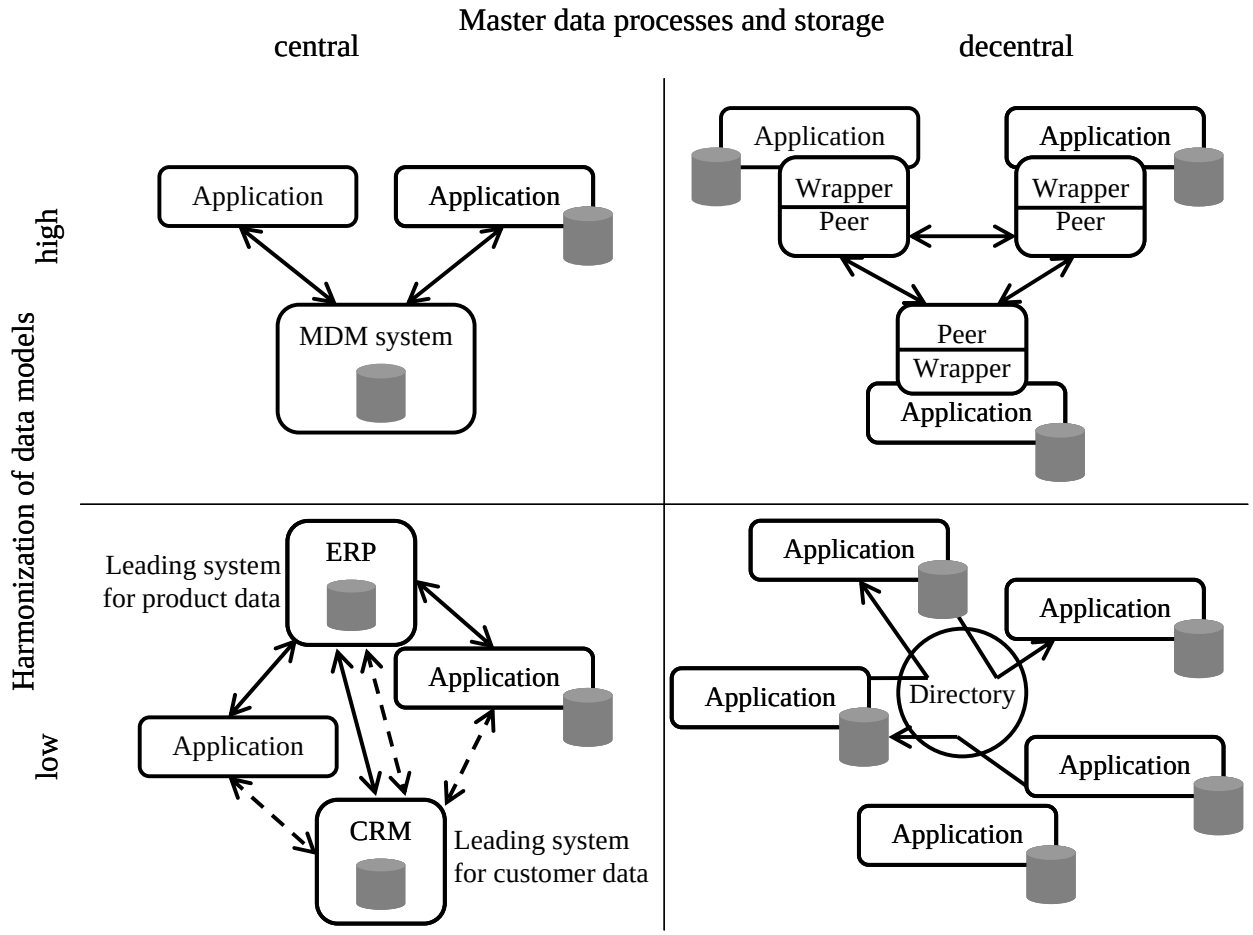

Figure 2: Architectures for master data management [13]

In Table 2 we compare the different design approaches with respect to the autonomy they allow towards the integrated systems, how MDM relevant process for information integration and information quality may be implemented, the effort necessary for the deployment of a system following the architecture and characterize which integration paradigm the architecture follows.

	Allowed autonomy for integrated system			**Processes**	**Effort**	**Paradigm**
	Design	**Interface**	**Access**			
Central	Low	Medium	Medium	Good	Medium	Materializing
Leading	Medium	Medium	Medium	Medium	Medium	Materializing
Directory	Medium	Low	High	Low	Medium	Virtual
P2P	High	High	Medium	Good	Low	Materializing

Table 2: Characteristics of master data management architectures

THE CURRENT STATE OF MASTER DATA MANAGEMENT

With the above discussed theoretical background on master data and the possible instantiations on systems managing this data we will now continue with a detailed look on commercial systems. Our main objective is to present insight into the current development of commercial systems with respect to their architectures especially for information integration, their data models, the security and methods for data quality tasks. We sent two questionnaires to the six active vendors for Master Data Management (MDM) solutions in Germany. While some vendors like Oracle possess several solutions, every vendor only completed the questionnaires for one product

(Table 3). This was no limitation we gave the vendors from our side. Except for Tibco we had an extensive telephone interview (approx. 1h each) afterwards with every vendor to add to the quantitative findings from the surveys additional qualitative insight. In the following we are going to present our results on the current state of the art of MDM solutions.

Vendor	**Complete product name**	**Product version**	**Shortened name**
IBM	IBM InfoSphere MDM Server for PIM	8.5	IBM InfoSphere MDM
Oracle	Oracle Customer HUB-Version	8.1.1	Oracle Customer Hub
SAP	SAP NetWeaver Master Data Management	7.1	SAP NetWeaver MDM
STIBO Systems GmbH	STEP 5	5.02	Stibo Step 5
Sun Microsystems	Java CAPS (Composite Application Plattform Suite)	6	Sun CAPS
TIBCO Software	TIBCO Collaborative Information Manager (CIM)	7.2.0	TIBCO CIM

Table 3: Systems in survey

In the following we distinguish between the MDM system and the integrated Information Systems (IS) whose master data get integrated. To the question what the primary architecture is all vendors but Sun responded the centralized architecture.

A very characterizing dimension on the target market of a system is given by the supported master data classes (Table 4). While some vendors claim to support all possible classes others claim to support only some selected classes. To our knowledge no system prohibits the creation of new master data classes or the alteration of existing classes, yet the support a system brings for a specific class extends the mere data model. It additionally includes processes, interfaces and even specialized algorithms for data quality concerns. Another aspect is the support of standard data schemas especially for data exchange, e.g. BMEcat for catalog exchange. One system (Stibo) already brings mappings from the shipped data model to BMEcat and other data exchange formats. Another system (Oracle) brings a strong customer focus. Oracle has several MDM products in its portfolio. The system we got data for is specialized for the customer data domain. The system from IBM implements the IBM Finance Industry Model which obviously focuses on the financial domain. Finally, SAP focuses in its data model on the integration of SAP ERP systems.

IBM InfoSphere MDM	**Oracle Customer Hub**	**SAP NetWeaver MDM**	**Stibo Step 5**	**Sun CAPS**	**TIBCO CIM**	
✓		✓		✓	✓	**All**
			✓			**Products**
	✓		✓	✓		**Customers**
	✓		✓	✓		**Suppliers**
	✓			✓		**Business Partners**
	✓					**Subsidiaries**
	✓					**Policies**
	✓					**Employees**
	✓			✓		**Patients**

Table 4: Supported Master Data Classes

All MDM systems integrate data continuously. Data changes are transported from the IS directly to the MDM system and are transported afterwards to other IS. This procedure can also be executed by all systems periodically. E.g. all changes are transported within the hour. Only one system (SAP) does not support a purely virtual integrated scenario. All other systems can integrate IS directly by functioning to them as their central data store. Yet, all systems focus on materializing integration by replicating data between all involved players.

Materializing integration is often encountered in data warehouses where the data is manifested in specialized schemas for business intelligence. Because of the familiarity of the integration paradigms customers require at times, that they do the integration task only once and achieve solutions to both challenges. Only three systems (SAP, Sun, Tibco) are prepared for this requirement and provide access for analytical purposes. The main focus of

all systems is the integration of master data for purely operational functions. It can be expected that in the future this gap between MDM and DWH will be diminishing.

Data modeling

A very important aspect of MDM systems is which data models it supports. This question has direct impact on the quality of the data model. Depending on the view on the data some people prefer hierarchically or object orientated modeled data while most IS prefer relational modeled data. Additionally, unstructured data is best viewed as such and not confined too rigorously into structured data models. In general, it is questionable whether there is an optimal data model for all purposes. Interestingly, MDM systems follow here very different approaches (Table 5). While one vendor (SAP) aims at supporting all data models all others specialize on certain models. Nevertheless, the relational data model is supported by all but one (IBM).

IBM InfoSphere MDM	Oracle Customer Hub	SAP NetWeaver MDM	Stibo Step 5	Sun CAPS	TIBCO CIM	
	✓	✓	✓	✓	✓	**Relational**
✓		✓	✓		✓	**Object oriented**
✓		✓		✓	✓	**Hierarchical**
		✓				**Flat**
		✓				**Unstructured**

Table 5: Supported Data Models

Next to the possibilities for internal data modeling, arises the question how data schemas may be mapped onto each other. Four Systems (SAP, Stibo, Sun and Tibco) support XSLT as a specialized XML based language for transformations and only one system (Tibco) the newer standard XQuery. While those languages are very expressive for data mappings they are often considered as not being very efficient at run time. To that end all systems but one (Oracle) support Java as a language for data mappings. Other high level programming languages like Visual Basic (Oracle) and Perl (Stibo) experience a very sparse support. Others are not supported at all (.Net, C, C++, Lisp, TclTK).

The support for XML standards continues when asked for languages for data modeling. XML Schema Definition (xsd) is supported by all systems and the older Document Type Definition (DTD) by all but one (SAP). As MDM Systems can be used for central data integration even beyond an enterprise' boundaries we get to a very interesting question: what standards are supported in data exchange? Here we see discrimination between systems and standards (Table 6). On one hand there are two systems (IBM, Oracle) that do not support standards at all. One supports two standards (SAP), one four standards (Stibo) and two even six standards (Sun, Tibco). On the other hand, the supported standards are very diverse. Only one standard (BMEcat) is supported by three systems and four standards are supported by two systems. Table 6 does not include the additional answers provided by the vendors.

IBM InfoSphere MDM	Oracle Customer Hub	SAP NetWeaver MDM	Stibo Step 5	Sun CAPS	TIBCO CIM	
				✓	✓	**UN/EDIFACT**
				✓	✓	**ebXML**
						openTrans
		✓	✓		✓	**BMEcat**
			✓			**xCBL**
			✓		✓	**cXML**
						eCX
						CIF
						ONIX

Table 6: Supported standards in data exchange. Without additional answers from vendors.

A more harmonic view is provided when the vendors are asked for the supported data models in data exchange. Here we see once again a broad support for XML. All systems support valid XML. Two systems (Oracle, Stibo) require the XML Document to be validated against a schema definition. Those systems do not support not-validated merely well-formed XML documents. Four systems (IBM, SAP, Stibo and Tibco) support Comma Separated Value (csv) files for data exchange and all but one (Stibo) support plain text files.

Security

IBM InfoSphere MDM	Oracle Customer Hub	SAP NetWeaver MDM	Stibo Step 5	Sun CAPS	TIBCO CIM		
✓	✓	✓	✓		✓	**Class**	
✓	✓	✓			✓	**Field**	**Schema level**
✓	✓	✓	✓		✓	**Relation**	
	✓	✓	✓		✓	**Object**	
			✓		✓	**Attribute**	**Instance level**
	✓		✓		✓	**Relation**	

Table 7: Adjustable permissions on schema and instance level

An interesting area is security in the context of a MDM solution. Certainly, the question is allowed whether one wants to enforce permissions using a central MDM system or leaves this task to the connected IS and therefore to the systems the users interact with. Consequently, the systems follow different philosophies. Every system includes a local user directory. This may be useful for local administration but is unfeasible for large enterprise wide deployments. To this end, all systems can connect to an LDAP server. Other techniques are only supported very sparsely. Another aspect that all systems have in common is that authorization is based exclusively on user roles and not on users.

We continued our questionnaire with an investigation on which level permissions may be defined (Table 7). To this end we looked at schema level whether the access to certain classes, fields or relations may be controlled. On this level for example the access to the field birth date may be denied, the affected roles may then not see any birth date at all. We continued with the same objects on instance level. We asked whether permissions may be enforced on certain objects, attributes or relations, e.g. if for one role the details on the spare parts (relation) of one product may be omitted. For the sake of brevity we omitted the results for the different operations (Create, Read, Update and Delete) on those objects. An interesting finding though is that one system (Sun) cannot enforce permissions at all, while others are quite expressive in this domain.

Data quality

All systems contain mechanisms for data cleansing tasks in form of semantical transformations. Most important, all systems contain mechanisms for reference reconciliation. While some systems may work together with full-blown data quality (DQ) Suites, they provide only simple algorithms out-of-the-box for this important task. Comparisons of objects are mostly done on the identity of two fields. Weights may be assigned to these fields and if a certain threshold is passed objects are considered to be duplicates. Only three systems (Stibo, Sun, Tibco) can evaluate references between objects for reference reconciliation (e.g. [4]) and only one system (Sun) brings a trainable algorithm, e.g. [2].

While for all systems a workflow is configurable to define how to continue with possible matches, two systems do not support automatic correction (IBM, SAP) or cannot provide a work plan for manual interaction (IBM, Stibo). One system (Sun) requires that all entities must pass the DQ Algorithms successfully before it may be added to the system's database. Unlike all other systems it does not even contain a temporary storage, where dirty data may be stored. This may be a severe hindrance if the DQ process is to be postponed to not intervene more than absolutely necessary with operational activities [3]. In total three systems (IBM, Stibo, Sun) cannot ignore a duplicate match if detected.

All systems check entities on data quality immediately for every newly created entity. To most systems this is optional. Some systems (IBM, Oracle, Tibco) may also check data quality in intervals or through a manual trigger (IBM, Oracle, SAP, Tibco).

CURRENT RESEARCH IN THE AREA OF MASTER DATA MANAGEMENT

In the last few years in the integration of enterprises a development towards higher autonomy of the integrated enterprises can be observed. Technically this is manifested in upcoming usage of Peer-To-Peer (P2P) Technologies for enterprise integration. For example the Global Data Synchronization Network (GDSN) was developed by the standardizing organizations GS1, UCC and GCI in 2004 to achieve multilateral product data integration. In this

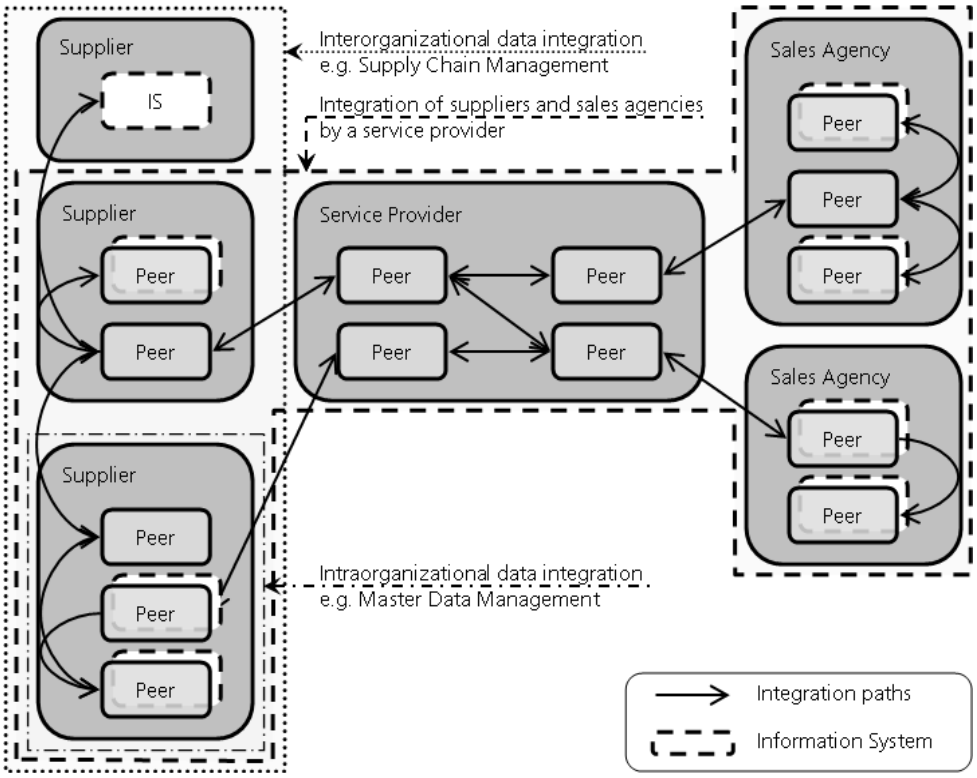

Figure 3: Peer Architecture with integration paths and supported scenarios.

concept the product catalog is offered in a central hub, where prospectors may find brief information and require full details on certain products. Those master data objects are then requested and transmitted pair wise [18].

For the optimization of supply chains RFID technology is continuously raising its importance. Big customers like Wal Mart require from their suppliers, that they furnish there shipments with RFID Tags. Here we observe a cost-benefit-asymmetry. The costs are with the mostly small or medium sized (SME) suppliers who because of their size cannot take advantage of the RFID Technology. The benefit is received nearly exclusively by the customer without costs. In a research prototype [19] P2P technology is employed so that SMEs can lower the necessary investments to achieve benefits from RFID technology in their own logistic chains. To this end a P2P storage network spanning the supply chain is developed that contains the necessary data for process optimization.

VIANA

For the integration of very small enterprises, especially for very small independent sales agencies [11] the project M3V (www.m3v-projekt.de) which is funded by the Federal German Ministry of Economy and Technology develops the P2P integration architecture VIANA [12]. We are now going to discuss this architecture in some detail.

The architecture was designed under the assumption that it will be accepted best if all participating partners achieve a benefit from it. Therefore, it supports all three scenarios described below (Figure 3) that can create an overall Win-Win situation.

Integration using a central service provider: This is realized in the integration of suppliers and sales agencies by a central hosted service provider. In this inter-organizational integration a data hub is present. It integrates data from all parties into its integrated database. Analyzing this scenario we see, that the major benefit of the platform is received by the sales agencies while the major investments – especially for the integration of multiple information systems – is to be invested by the suppliers. Therefore, we introduce the next two scenarios.

Integration without a central service provider: This scenario is usually found in Supply Chain Management (SCM) where product catalogs are published along the value-chain to principals who in turn create transactional and store inventory data. Especially for highly integrated manufacturing processes as just-in-time (JIT) or just-in-sequence (JIS) [8] processes, high data quality is of vital importance. In this scenario usually no single participant of the primary value chain is able or interested in providing a central service for high data quality.

Intra-organizational integration: In this scenario an enterprise integrates its internal IS to achieve an overall increase in data quality and information homogeneity. It is important that the benefit may hereby be achieved without the integration of any external IS. Effectively, this scenario is intra-organizational master data management.

The Architecture of VIANA – a PDMS that integrates Information Systems (IS) – concentrates on the materialized integration of master and operational data. It focuses on write operations and mechanisms to enhance data quality. The key concepts are shown in Figure 3. The communication and integration is established by peers. A peer wraps

an information source and transports changes in the data along configured paths. We emphasize that those paths are not used for querying data as in virtual integration but for the propagation of write operations. Read operations are executed exclusively locally.

Types of instantiation

Peers are used to accomplish the integration between enterprise information systems. Certainly, they do not replace those systems. Their only function is the integration with neighboring systems equally represented by peers. From a peer it is demanded that it provides the necessary functionality to cooperate with other peers. Like in the Wrapper/Mediator Architecture [17], the functionality a peer provides towards the integrated information source depends on the capabilities of the source. It is required that the peer, independent of the implementation details, always initiates an operation in VIANA when a local atomic transaction is executed. We understand that this is a challenge if the corresponding IS cannot trigger events on atomic write operations. In this case periodic checks for changed data could be implemented. We emphasize that a sufficient high frequency is needed to lower the chance of conflicting write operations which would demand user interaction. We now discuss several types. The numbers in parenthesis are references to Figure 4.

Wrapper to a database: (1a) This kind of interaction extends common databases with PDMS facilities. We show the integration of the database using publish/subscribe interaction. That is, because every write transaction in the database is to be published immediately to neighboring peers. A standard conformant way to achieve this in relational databases would be by using SQL/Trigger. More efficiently, it may be implemented using vendor dependent transaction logs. Interaction directly with databases is not limited to stand-alone databases but includes the databases of information systems (1b). While the latter may impose some difficulties in applying business logic to data it may at times be the only way to integrate legacy systems.

Wrapper to an information system: (2) Many modern information systems provide a way in which external applications can monitor their data and provide write access by some kind of interface. This type of interaction is similar to integrating databases directly, yet it eliminates the need to care about business logic. Thus, it is the preferred way of integrating IS.

Hybrid Wrapper for observation and access: If an information system does not provide the ability to observe its data for changes but merely provides the functionality to read and write data by its interfaces, then the observation task can be done directly on the database (3a) or periodically checking the system for updates (3b). The data access remains using techniques of the information system.

Interface for a remote information provider: (4) In this scenario a peer publishes an interface and is invoked by a remote information system. The information flow is unidirectional from the remote system to the peer. As a consequence, the peer only provides outgoing connections to other peers.

Interface to a remote information consumer: (5) Here, the information flow is unidirectional from the peer to the remote system. As a consequence, the peer only provides incoming connections from other peers.

We understand the P2P approach as a generalized architecture for materializing integration. A super-peer is supported by the architecture as a special instantiation thus a centralized P2P Architecture [1] can be resembled. Particularly the Hub-and-Spoke Architecture [6] forms a specialization of the here presented architecture. While this central hub has the notion of a global data schema it is important to note that the service provider only hosts the information necessary to supply its own services and that from the perspective of a peer no peer forms a central hub.

At the end, VIANA provides two views: (1) a heterogeneous view towards the integrated systems, allowing them to maintain the autonomy of this IS. This may be desirable for political, technical or financial reasons. Through information hiding [15] this heterogeneity is encapsulated by the infrastructure of VIANA which provides (2) a homogeneous view towards its neighboring peers. This allows a smooth integration of IS by the replication of relevant data through semantic links [10] between neighboring peers.

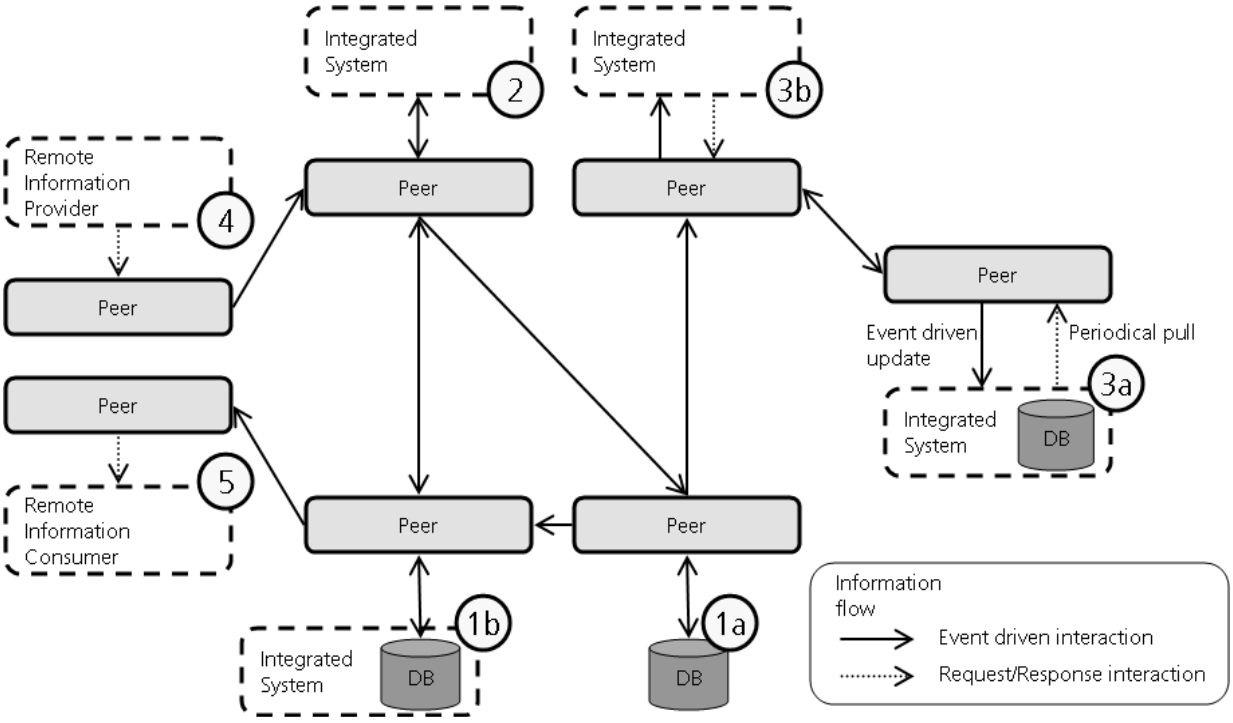

Figure 4: System overview depicting interaction patterns and information flow

CONCLUSION

In this paper we discussed the characteristics of organizational data. We showed how master data relates to other data categories present in enterprise, namely transactional and inventory data. From the description of the data we derived its value for enterprises which justifies a dedicated treatment of it. This is the domain of the discipline of master data management. We therefore continued first with a discussion of architectures for master data management, before we analyzed commercial solutions. Here we presented results of two surveys we executed with the system vendors. Discriminating features could be found in the field of data modeling and target industries. Additionally, the supported data exchange formats are very diverse which suggests that the vendors follow different strategies. While master data management aims at raising and maintaining high information quality, the systems often provide out-of-the-box only basic features for this task. Yet, most systems may work together with specialized solutions in this domain.

After the presentation of the current status of commercial master data management solutions we continued our discussion with upcoming technologies. We identified Peer-To-Peer (P2P) as such and described three solutions following this approach. We described the architecture of VIANA in more detail and showed how wrapped systems may replicate their data in a P2P context bridging the gap between heterogeneity and homogeneity in data models and autonomy.

REFERENCES

[1] Androutsellis-Theotokis, S. Spinellis, D. , A Survey of Peer-to-Peer Content Distribution Technologies, 2004

[2] Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P. Fienberg, S. "Adaptive name matching in information integration", Intelligent Systems, IEEE, 18, 5, Sep/Oct, 2003, pp. 16-23

[3] Cappiello, C. Comuzzi, M. "A utility-based model to define the optimal data quality level in IT service offerings", 17th european conference on Information Systems (ECIS), 2009

[4] Chen, Z., Kalashnikov, D. V. Mehrotra, S. "Exploiting relationships for object consolidation", IQIS '05: Proceedings of the 2nd international workshop on Information quality in information systems, New York, NY, USA, 2005, pp. 47-58

[5] Elmagarmid, A. K., Ipeirotis, P. G. Verykios, V. S. "Duplicate Record Detection: A Survey", IEEE Transactions on Knowledge and Data Engineering, 19, 1, Los Alamitos, CA, USA, 2007, pp. 1-16
[6] Erasala, N., Yen, D. C. Rajkumar, T. M. "Enterprise Application Integration in the electronic commerce world", Computer Standards und Interfaces, 25, 2, 2003, pp. 69 - 82
[7] Eschinger, C. , Report Highlight for Market Trends: Master Data Management Growing Despite Worldwide Economic Gloom, 2007-2012, Gartner, 2008
[8] Frazier, G. L., Spekman, R. E. O'Neal, C. R. "Just-in-time exchange relationships in industrial markets", The Journal of Marketing, 1988, pp. 52-67
[9] Haas, L. "Beauty and the Beast: The Theory and Practice of Information Integration", Database Theory – ICDT 2007, 2006, pp. 28-43
[10] Hose, K., Roth, A., Zeitz, A., Sattler, K. Naumann, F. "A research agenda for query processing in large-scale peer data management systems", Information Systems, 2008
[11] Kokemüller, J., Kett, H., Höß, O. Weisbecker, A. "A Mobile Support System for Collaborative Multi-Vendor Sales Processes", Proceedings of the Fourteenth Americas Conference on Information Systems, Toronto, ON, Canada, August 14th-17th, 2008
[12] Kokemüller, J., Kett, H., Höß, O. Weisbecker, A. "An Architecture for Peer-to-Peer Integration of Interorganizational Information Systems", 15th Americas Conference on Information Systems (AMCIS), San Francisco, USA, August 6 - 9, 2009
[13] Legner, C. Otto, B. , Stammdatenmanagement,.
[14] Leser, U. Naumann, F. , Informationsintegration, dpunkt.verlag, Heidelberg, 2007
[15] Parnas, D. L. "On the criteria to be used in decomposing systems into modules", 1972
[16] Rahm, E. Do, H. H. "Data Cleaning: Problems and Current Approaches", IEEE Data Engineering Bulletin, 23, 4, 2000, pp. 3-13
[17] Roth, M. T. Schwarz, P. "Don’t Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources", Proceedings of the 23rd VLDB Conference, Athens, Greece, 1997, pp. 266-275
[18] Schemm, J. W., Legner, C. Österle, H. , Global Data Synchronization — Lösungsansatz für das überbetriebliche Produktstammdatenmanagement zwischen Konsumgüterindustrie und Handel?, Physica-Verlag HD, 2008
[19] Schönemann, N., Fischbach, K. Schoder, D. "P2P Architecture for Ubiquitous Supply Chain Systems", 17th European Conference on Information Systems (ECIS), Verona, Italy, 2009
[20] Walter, P., Werth, D. Loos, P. "Peer-to-Peer-Based Model-Management for Cross-Organizational Business Processes", Los Alamitos, CA, USA, June, 2006, pp. 255-260
[21] Wang, R. Y. Strong, D. M. "Beyond accuracy: what data quality means to data consumers", Journal of Management Information Systems, 12, 4, 1996, pp. 5-33