The MDM Golden Record is Dead, Rest in Peace – Welcome Interpreted Interoperable Attributes

Tomi Dahlberg
Ph.D., Professor of Information Systems (act)
Åbo Akademi University, Institute for Advanced Management Systems Research
Senior Research Fellow, Turku School of Economics at University of Turku
Executive in Residence, Aalto University Business School

Abstract

So-called golden record philosophy has dominated master data management (MDM) development activities during the recent years. Yet, this philosophy has been unable to deliver the promised solution, that is, organization-wide federation of customer, vendor, employee, product and other master data and through master data the federation of transactional data. Although significant improvements have been made, MDM solutions crafted have remained fragmented instead of being organization-wide. This article proposes that the key to organization-wide (master) data federation lies in interoperable attributes and their metadata. Two assumptions inherent in the golden record philosophy need to be revised to achieve this. Firstly, the ontological assumption needs to be that data is contextually defined instead of there being a single version of truth as the golden record philosophy assumes. Thus data may have several true meanings depending on its use context and the time of the data usage. Secondly, data management is increasingly done in open systems environments. Organizations have lost partly or fully control over their information architectures as they use a myriad of information systems and information sources developed and provided to them by vendors and other external parties. Finally, this article suggests that the interpreted interoperable attributes philosophy can be used to federate data beyond master data and transaction data internal to an organization, most notably big data.

Keywords: Master Data Management, Ontology of Data, Data Management, Metadata

1. Introduction - Background of this article

The purpose of our research team is to write articles about how to federate data from a myriad of information systems, and their data storages, in open systems environments. Such data federations include data architecture modeling within enterprise architecture (EA), federations of data to execute business transactions, eCommerce and managerial reporting as well as Big Data analyses/analytics. In open systems environments, data federations are characterized by differences in the formats, structure, granularity and other characteristics of data. Federations of internal and volume-wise rapidly increasing external data as well as federations of structured and unstructured data from multiple data sources, such as business transactions, digital sensors and social media are some of the new requirements. In order to achieve our research objectives we feel that we need to answer the following two questions: why has the golden record philosophy not been able solve master data management (MDM) problems in organizations, and why the golden record concept may lead to erroneous conclusions if data is federated on the basis of this philosophy. The philosophy – that is, the data ontological approach - on how to federate customer, product and other master data also defines how one thinks that data should be federated in general. Data federation can happen only if there is a connecting data element such as a customer or a product. Thus, it is necessary to address "the death" of the MDM golden record philosophy to understand how data can be federated in general.

This particular article is aimed for practitioners to get their feedback on the thinking of our research team. To help the reading of this article I have omitted references in the way they are used in scientific articles. I have also tried to write this article in natural language whenever possible. Those interested about references may look at DAMA International's DMBOK Handbook, the writings of Yair Wand and Rob Weber, Juhani Iivari or Kalle Lyytinen. This article also continues work done in two previous articles entitled "A Framework for The Corporate Governance of Data – Theoretical Background and Empirical Evidence" published in July 2015 and "Governance of Data, Data Governance and Data Management in Electronic Commerce – A Generic Governance Framework and Two Cases" conditionally accepted to an another journal and likely to be published in 2016. The writers of the mentioned two articles are Tiina Nokkala and I. I am Tiina's Ph.D. research supervisor. Our research and writer team includes also professor Jukka (Jups) Heikkilä and Ph.D. Marikka Heikkilä. I have previously written an article entitled "Framework and Research Agenda for Master Data Management in Distributed Environments" in 2011 together with Jukka and Marikka Heikkilä. Our ontological stance to master data and its management was stated already in that article, and we have had ample

opportunities to discuss our stance with both practitioners and scholars over the years. Any feedback on this article is highly appreciated.

Finally, this article is one of our final inputs to and result in the 4-year data to intelligence (D2I) shock program funded by TEKES (the Finnish Funding Agency for Innovation). In addition to Jups, Marikka and Tiina I want to thank the dozens of persons I have had the opportunity to discuss the topic of this article over the years. The journey has only started.

2. Why does the golden record philosophy not make sense?

There are two main reasons why the golden record philosophy doesn't work as expected.

Firstly, information systems and their data storages are increasingly characterized by open systems environments in comparison to closed systems environments a few decades ago. A closed information systems environment means that an organization develops itself or is responsible for the development of the information systems, including their information architectures, that the organization needs to enable the execution of its (business) activities. Thus, the organization knows exactly the scope of each information system, including the meaning of data entities, attributes and their dependencies in and between ISs. In such IS-environments the "information architecture" of information systems as a whole can be designed so that there are no unnecessary overlaps. Data interfaces are consequentially pre-defined, that is, the "interfaces" of each information system defines the data that the information system is able to receive and submit and how data interchange is initiated and/or called for. In other words, data is interchanged and federated by applying the data model of each IS. Initially data federation was done with transfer files, typically via separate batch processes. Application programming interface (API) concept emerged later to help data integrations between ISs. It is noteworthy that organizations did not have holistic enterprise information architectures in the sense we understand them today.

In most organizations this kind of environment is long gone.

Dozens of more or less overlapping information systems developed by independent software vendors contain same customers, vendors, employees, materials, products/services, functional locations, financial/managerial accounts and other master data entities and attributes. Consequently, in addition to master data, overlapping is true also for other types of data such as business transactions, reports, documents and contents. This impacts negatively organizational activities from daily operations to managerial reporting. Overlaps and fragmentation leads to the question, what source(s)

of data should we use and especially trust. Furthermore, even if the APIs of various ISs would be known their limitations make data federations challenging. For example, data federations between asynchronous and synchronous or, simply, differently timed automated processes / applications, can be very demanding. Another limitation is that APIs do not contain metadata about the meaning of data in ISs or changes in the meaning of the data within ISs. Still the need to federate data, usually from multiple different perspectives, is strong in order to produce managerial reporting and other holistic pictures of customer, vendor, employee and other artifacts and about organizational activities related to these artifacts. The growth of external data is also characteristic to open system environments. The volumes of external data about the customers, products and the activities of an organization grow at an astonishing pace. Compare this to closed systems environments, where data was largely structured data from organizations' internal ISs. Master data management (MDM) concept was, later on, introduced as one of the means to solve customer, product and other data federation needs.

Secondly, the ontological assumption that data has the same meaning – which is the basis of the "the one version of truth" thinking in – is flawed. This assumption of the golden record philosophy appears in two forms.

One assumption is that data entities and attributes have only one meaning. This leads to the conclusion that if there are several ISs with overlapping data, it is possible to create a single version of truth for data entity and attribute values, which are common to or shared by the ISs. This could be referred to as the golden record, that is, a record that has the true values of shared data attributes, for example for a customer. The logical consequence is that all ISs should use these true values. The other logical consequence is to conduct a "match and merge" process is to remove other than true values.

The second assumption states that if ISs have differences in their data models, it is possible to agree and develop a unified data model for shared data. The idea to remove inconsistencies by harmonizing data is inherent in the golden record philosophy. My claim is that the meaning of data is contextual and subject to changes over time, as the examples given later in the article will show.

Organizations, trying to harmonize master data on the basis of the golden record philosophy, have been detected to fail in their efforts to establish organization-wide data interoperability and federation. Success has been limited to silos of organizational activities, such as bill of materials (BOM) in purchasing and procurement or customer data intelligence (CDI) in marketing and sales. These improvements are obviously important achievements when compared to the typical mess of data at the starting point of MDM development. Yet, the golden record philosophy and data harmonization based MDM has fallen short on its promise to solve data federation needs on an organization level.

From an enterprise information architecture and an organization-wide data management perspective, the main reason for the failure is that the interpretation of data (= the meaning of data) is genuinely similar only within the organizational / functional silos, such as purchasing materials to manufacturing for a limited period of time. Consequently, applications / information systems that support and enable the activities/processes of those silos, could share a data model or have compatible data models that interpret data entities and attributes similarly. The same is not true on the organizational level, that is, organization-wide data models with similar meaning of data entities and attributes do not exist. Instead of just one interpretation there are several different interpretations on at first similar-appearing data, such as material (e.g. logistics and capital value of BoM materials versus logistics and capital value of spare parts differ) or customer data (e.g. legal unit versus sold-to versus bill-to versus ship-to data differ). These various data interpretations, that is, different data contexts, are valid, true and important for the operations execution of the organization and could be supported and enabled by different ISs / applications. What would happen if the organization, as an example, would define one true logistics and capital value for materials as the golden record philosophy suggests? Business critical data would be lost. Moreover the meaning of data changes over time. For example, (BOM) materials purchased to manufacture products may change over time due to technological development or supplier changes. What would happen to materials data in BOM inventory, spare parts inventory or in installed base register if one true value would be forced to all ISs? Again business critical data would be lost.

One reason for the emergence of the golden record philosophy lies in textbook entity relationship diagramming/modeling (ERD). The limitation of the textbook ERD is that metadata describing the meaning of the data attributes is considered irrelevant and is thus not documented. When information systems were developed in closed systems environments, as explained above, this was not necessary because the meaning of data was known in each information system. A separate information system was developed for each organizational activity including its separate unique data model. Note that the classical data diagramming/modeling techniques were developed and textbooks were written during the era of closed systems environments. Still today ERD models do not specify how the data governance is implemented for data entities and attributes. The golden record philosophy with the assumption of one reality results in a flawed governance model. Due to the one reality assumption the contextual data needs of most data user groups are not taken into consideration. If data entities and attributes, which appear similar but represent different contexts, are mechanically federated and harmonized from a myriad of information systems (with different data models), the result can be

almost anything and most likely business critical data is lost as the examples above indicated. This occurs not only for master data, but also for other types of data, such as transactional data. The claim of this article is that (in open systems environments), it is necessary to understand the initial, the maintained and other use contexts / meanings of data to be able to federate data meaningfully.

In some organizations the golden record philosophy and data harmonization based MDM has led to a so-called Master Information System philosophy. This means that those organizations have agreed that there is a specific master information system for each organizational activity, such as sales or materials purchasing. For example, a CRM system could act as the master information system for customer data. This approach, the Master Information System philosophy, is able to produce the same improvements and has the same limitations as the golden record philosophy in general.

The claim that the meaning of data is contextual and changes over time is probably easiest to show with examples. The body temperature measurements of a human being constitute a good neutral example. The layman's rule of thumb is that a person is healthy when the body temperature is 37 degrees in Celsius. Yet, after sleep, stressful activity, medical surgery, and in many other situations – contexts of data - a clearly lower or higher temperature is fully normal. Moreover, the measurement device, the measurement method and their calibrations impact the result of the measurement. The same is true for other contextual characteristics of a body temperature measurement, such as is the person lying, sitting or standing or from what body part is the temperature measured. How should body temperature data be federated and harmonized? Creating a database / data vault / data lake, where all human body temperature measurements are stored, is useless unless we know the reasoning of the measurements, that is, the purpose and how the data was created and why was the value of a measurement changed if that has happened. The golden record -philosophy would solve data federation by agreeing that there is only one correct way to measure body temperature, which is then used to produce true values. Sounds good, but is alien to reality. Those who create and maintain data have rightful – vested - interests for conducting the body temperature measurements in their way in their context / reality. The proposition of the golden record approach to measure body temperature with one agreed correct way is useful for each particular context should they have several ways to do that. (This is the improvement potential of the golden record approach discussed above.) Secondly, the practices to measure body temperature evolve and improve over time. For example, devices used in households to measure body temperature have developed from mercury-based to alcohol-based thermometers and to digital devices. These devices produce different results. Should we ignore and/or delete results that have been achieved with less reliable devices as the golden record philosophy would

suggest? Obviously not, instead we should include metadata about he device used in each measurement. In summary, we would loose valuable contextual information by agreeing that there is only one correct way to measure human temperature and might not be able to use historical data because past measurements are done in different ways. The solution to data federation needs lies in data interpretation done with the help of contextual metadata not in forcing data to be similar.

Replace now human body temperature data with material data, customer data or other master data constructs. Take material data as an example. Should we use the material information such as prices and logistics data from the time when the material was purchased (i.e. BOM data), or after work and other materials were added to BOM data in manufacturing, or when the product was sold to a customer with modifications done at the delivery, or when add-ons, replacements and repair maintenance were added to the product by after-sales, spare-parts and other service operations?

Similarly, who and what is a customer when consumers are our customers? Is a customer the person/address that purchases our product/service, or the person/address who is the user of our product/service, or the person/address who pays the bill for our product/service, or the person/address used in the delivery of our product/service, or is it a household instead of just one person? One may create similar differences in the meaning for other master data constructs as well. Information technology does not limit the amount of data we could have on a customer, product or other (master) data entities.

With big data and Internet of things, the significance of data interoperability and federation increases even further. Two consequences appear inevitable. Not only information structures, rules, mappings and other similar characteristics of data are metadata but also business data becomes metadata since it can be used to federate data. Secondly, unstructured data sources do not match at all with the golden record philosophy, but require an interpretation-based approach to data.

Data Lake is one of the novel concepts, which reflects these consequences. For example, all wellbeing and medical data about persons could be stored into a data lake. The data in the data lake could include self-registered data about physical exercises, social welfare data, data about medical incidents, vaccinations, allergies, hospitalizations, medicine prescriptions and lots of data from other relevant data sources. If a medical doctor would then create a data federation to include all body temperature measures and their history for his/her patient from the lake without knowing the contextual metadata of the measures, could that data be used to diagnose the patient? If this would be based on the assumption that all measures are true, since there is only one way to measure the body

temperature, such an assumption might lead to a seriously erroneous diagnosis. Contextual metadata is needed to support the medical doctor in the use of the data for making the diagnosis.

Returning back to closed information systems, there is an additional flaw in the golden record thinking. Within information systems research, the writings of Wand and Weber became very influential at the end of 1980s. According to them it is possible and thus necessary to design an information system so that it is a true representation of its organizational (social) use context. Consequently, the data of a well-designed information system is a true representation of its users' social reality. The information system needs to fulfill three criteria with its data model: stationary (=data description of the organizational use context), stage shifts (=description of transactions that change stationary situation from current to an updated stationary situation), and reporting (=about the stationary situations and their stage shifts) requirements. There is also a strong consensus that these data representations of the reality are based on the vested interests of the IS users, which - surprise, surprise - are contextually different according to Wand and Weber. This is the meaning of social reality. Our research team's ontological stance builds on this work and extends it to open systems environments. In open systems environments there are significantly more non-similar contexts, with several different representations of reality based on non-similar vested interests. Secondly, in open systems environments responsibility for the development of information systems has shifted to software vendors, as explained earlier. The data models of the ISs used within an organization could be partly or totally unknown to the organization. Furthermore, since each user organization has applied the ISs to her contextual reality, the meanings of the data attributes differ between those organizations even when the data model is the same. During the last years we have witnessed the emergence of platform environments where the ISs of multiple organizations share information through the platforms. The conclusion is that data federation over multiple true representations of reality requires metadata about each reality / context and their mappings.

3. How to solve the data federation challenge in practice?

The question how to link/federate data between information systems is an old issue. For example, in early 1980s, when the first generation software packages were developed, they often shared "parameter data" – as master data was called at that time. For example, by sharing financial and managerial account keys, customer reporting keys and other parameter data, it was possible to transfer data between software packages and to automate accounting data entry in so-called pre-systems to general ledger. The same parameter data handling modules were added to each software package if

needed. For example, a sales software package included financial and managerial accounting keys. When a sales transaction was entered, relevant financial and managerial accounting references were registered as a part of the sales transaction. Selected data attributes of the sales transaction could then be transferred to accounts receivables and financial/managerial accounting software packages as accounting events. Technically, data transfers were executed through point-to-point integrations and/or as batch file transfers with control reporting.

Later, parameter data has also been called static data. SAP, probably, coined the word master data to describe parameter/static data, which is shared by several modules / applications / ISs, is non-transactional, managed separately and attached to transactions, and which can be used to federate (transactional) data. Even more important, SAP created the first true inheritance-chains that allowed business scenarios to be designed into master data models, which facilitated automation during the creation of business transactions. Unfortunately, these functionalities are rarely and poorly exploited.

As the use and number of information systems grew rapidly and shifted from in-house development to software packages and software vendor developed ISs, the limitations of point-to-point integrations in data federation grew. This is largely true also for so-called wall-to-wall concepts such as SAP ERP. Reasons for this are that a wall-to-wall solution could initially have been developed by several software vendors which have then been purchased by the wall-to-wall vendor, or that the modules of the wall-to-wall solution are implemented at different times in the user organization with module, that is, context specific objectives. The consequence is that the interpretation of data between modules and times is different. Finally, organizations seldom rely on one wall-to-wall solution but purchase ISs from several vendors.

Some 10-15 years ago, several large organizations noticed that they had common business data about customers, products in several functionally overlapping modules / applications / IS. I have witnessed a situation where a global company operating in 100+ countries had over 150 sales IS and customer registers to contact the same global and local customers. As mentioned earlier this was one of the reasons for the emergence of the master data and MDM concepts.

The first generation of MDM philosophy, 10-15 years ago, suggested that it would be a good idea to bring customer, product and other master data together and by doing so to be able to remove duplicates, correct data entry errors and harmonize data. This approach soon faced the difficulty that data brought together was too inconsistent and fragmented. Data harmonization succeeded only for registers and activities, which were close enough.

The challenges of the first generation MDM philosophy led to the second generation of MDM philosophy, characterized by the golden record philosophy. According to this philosophy, the entities and attributes of (main) data registers are listed and compared for a data domain, such as the customer data domain. The purpose is to detect which entities and attributes are shared between registers. In practice this is done with so called affinity matrixes. In an affinity matrix each column describes the entities and (key) attributes of one register and the rows describe individual entities and (key) attributes. By crafting such a matrix it becomes possible to detect, which entities and (key) attributes are shared between registers. For example, for customer data it is possible to craft an affinity matrix to compare the customer data of sales, crm, accounts receivable and other modules / applications / ISs, each of which have separate customer registers.

It has been claimed that by detecting shared entities and attributes, it is possible to determine so-called “global master data”. In the golden record philosophy, global master data means master data shared by all or most ISs. Global master data entities and attributes could be used to define golden records, that is, their structure and logical content. After that, by focusing on these entities and attributes it is possible to agree the single version of truth for each customer, that is, to create the physical golden records with their values. Each customer register should then use the content of the golden record. This is done so that new customers and changes to customer data are registered first to a master register/system to create the golden record, which then populates entries to other registers. Should a new customer be detected first in another register, this data is transferred to the master register / information system to be opened for use or is used in a pre-golden record status. The benefits and limitations of this approach in data federations have been described above. This approach has also been referred to as the canonical model, where only shared attributes are seen as the relevant parts of the golden record.

In addition to flaws in the data ontology and closed system environment assumptions, the golden record philosophy has also been criticized for its inherent single-domain approach. A lot of work is needed to list the entities and attributes of any domain (into an affinity matrix). In the golden record philosophy this is done to “match and merge” data, that is, to harmonize data to reach the single version of truth, (after which other versions are removed by forcing them into the single true value). For the golden record to make sense it is necessary to include at least the most important registers for a domain into the match and merge process. Since the crafting of an affinity matrix demands a lot of work, that leads to the conclusion and recommendation that it is necessary to focus on a single data domain at a time, as multi-domain approaches appear too arduous.

Fixing data quality problems in customer data, as an example, may help an organization. Yet, interactions with customers involve usually several master data domains. For example, when the organization sells a product to a customer, a payment term is agreed. The organization will also create accounting entries to book the sales and may wish to register the employee, who conducted the sales transaction. This simple example has four master data domains: customer, product, accounting and employee data.

The interpreted interoperable attribute philosophy to master data recommends that all master data of a selected activity or process should be addressed instead of a single domain. This MDM philosophy is inherently multi-domain. It is noteworthy that the results of a development activity or project improve the quality of all data domains within the scope of the development activity / project immediately upon completing the development activity / project. By limiting the scope of the development activity / project it is possible to run smaller development activities than with the golden record philosophy, should that be needed. According to the interpreted interoperable attribute philosophy it does not matter if customer data, as an example, have different interpretations outside the scope of the development activity / project. Within the development activity / project the focus is, and needs to be, in describing the links and meanings of interoperable attributes. Actually, the existence of other interpretations/meanings is normal and expected for data outside the scope of the development activity / project.

The consequences for the daily master data use and management are interesting. The golden record philosophy leans strongly towards highly centralized master and metadata management and governance, at least for global data. The existence of a single metadata model (descriptions of data) is a logical consequence from the ontological assumption that data always has the same meaning At the entry of new master data, changes should be made first to the master register/system and then enforced to other registers/systems through data population. Or, if pre-entry processes/systems are allowed, a centralized global process needs to confirm the golden record, which means that the pre-entry process/system has to cover the needs of the golden record, usually the canonical model.

The interpreted interoperable attribute philosophy allows varied data management and governance arrangements since data is seen to have several meanings. The focus is on understanding the various meanings and links between data attributes and in describing them with metadata. Consequently all local interpretations/contexts are considered true and “global master data” is the sum of all local metadata interpretations.

In the interpreted interoperable MDM philosophy, the modules / applications / ISs are allowed to have their own master data registers, based on vested contextual needs. This is excellent news from the perspective of securing return on investments from money spent on ISs purchasing. An organization may continue to use its legacy ISs longer. Modules / applications / ISs are federated by an MDM service/application, which is a repository of interoperable attributes with their various metadata descriptions. In this star-type architecture, changes in the master data of any federated module / application / IS could be populated to other modules / applications / ISs by using known metadata and attribute mappings. Further, each module / application / IS could dynamically subscribe new and/or changed metadata entries. Thus data attributes and relevant metadata live side-by-side.

One CIO of an end-customer described such a MDM application as a re-fueling application, tanking up short-living, fit-for-purpose business applications with necessary master data, so that they are able to perform their tasks over a restricted time-period. The metaphor, of course, was to fueling gasoline into the tank of a car.

Another interesting and significant feature is that the MDM attribute repository could be used to create entirely new views to data and, thus, to federate data in a new way from the federated modules / applications / ISs data. At the same time, the establishment of information management at the attribute level means that previous record driven harmonization and dimensioning can be discarded. One is able to present any attribute-driven view from the federated data.

In summary, my proposition is that data federation done on the basis of interpreted interoperable (master) data attributes establishes the third generation of MDM philosophy. It could also be called "architectured enterprise information asset management" when the principles of this philosophy are extended to data federations outside of master data, that is, to big data. This philosophy still has its roots in MDM thinking, but attempts to solve the problems of the golden record philosophy. The starting point of the interpreted interoperable attribute philosophy is the ontological claim that data may have several contextually dependent meanings, that is, data may have several metadata descriptions as opposed to a single metadata model. (Master) data may even have several independent representations within one context at any point of time or over time. Data federation is done on the basis of contextual metadata, that is, through interpreting attributes that make data interoperability and federation possible. It could be possible to craft similar affinity matrixes as in the golden record philosophy by focusing on entities and key attributes. After that - as a new task - the alternative metadata descriptions of attributes need to be defined. This obviously cannot be done entirely manually. It is probably also possible to avoid the use of entity relationship diagrams – they may not

even be available – by analyzing the content of transactional databases to capture metadata. Actually, this appears to be the reality in most big data contexts.

4. Concluding remark

This article has attempted to explain the limitations of the MDM golden record philosophy and why that philosophy does not work on the organizational level. The two main reasons given are related. The assumption that it possible to define the content of master data attributes so that the content is the same always and everywhere is flawed. The current and increasing myriads of modules/applications/ISs about the same customers, products and other data attributes make the limitations of the golden record philosophy visible in practice. Organizations have moved from closed to open systems environments where they need to federate data between vendor developed ISs and between internal and external data sources.

The article has also attempted to depict how the interpreted interoperable MDM attribute philosophy is able to build new views on existing knowledge by extending the possibilities to federate data. By capturing metadata it is possible to remove the limitation of textbook data modeling of not including metadata descriptions. Actually, interpreted interoperable MDM attributes live their own lives, together with their metadata. These attributes are fit-for-fight for Big Data comparisons. This approach facilitates federation of data with different structure, format and granularity as long as there are attributes that can be used to federate data.