HOW TO ADDRESS MASTER DATA COMPLEXITY IN INFORMATION SYSTEMS DEVELOPMENT – A FEDERATIVE APPROACH

Research paper

Dahlberg, Tomi, Turku School of Economics at the University of Turku, Turku, Finland, tomi.dahlberg@utu.fi

Lagstedt, Altti, Haaga-Helia University of Applied Sciences, Helsinki, and Turku School of Economics at the University of Turku, Turku, Finland, altti.lagstedt@haaga-helia.fi

Nokkala, Tiina, Turku School of Economics at the University of Turku, Turku, Finland, tiina.nokkala@utu.fi

Abstract

We investigated the failure of an IS project within a global industrial company. The challenges in the complexity of product master data were one of the failure reasons. We detected a research gap in how to handle master data complexity in IS development, especially in data storage integrations. The traditional approach is to define so-called golden records, "single versions of truth", for each record, and then harmonize and cleanse data so that only or mainly golden record values will be used. We offer federative approach as an alternative to the golden record approach. According to this approach data interoperability is achieved by identifying shared attributes, by federating data on the basis of shared attributes' metadata, and by developing IS functionalities to process the metadata and their cross-references. We compare the ontological stances of the approaches theoretically and with figures. We present the results of a case, where the federative approach was probed. Our study contributes to research by showing how to link data management to IS development to address the complexity of master data in data interoperability projects, by comparing the golden record and the federative approaches, and by showing how the federative approach can be used in real-life contexts.

Keywords: Information systems development, master data management, data integration, data federation, golden record approach to data integration, federative approach to data integration, case study.

1 Introduction

Due to digital data explosion (Hilbert and Lopez, 2011), organizations have more data available than ever to be consumed for a myriad business purposes. The dominance of storable digital data over storable analog data is, however, a recent phenomenon. IDC (2011) as well as Hilbert and Lopez (2011) estimated that the amount of digital data created by the mankind surpassed the amount of analog data during the years 2002–2003, rose to 94 % in 2007 and is currently close to 100 % of storable data created by the mankind. The annual growth rate of digital data was 58 % during the years 1986-2007 (Hilbert and Lopez, 2011) and the growth is estimated to continue at that rate at the minimum. The 58 % annual growth rate was calculated from the number of server computers and their data processing capacity. The figure does not include data created with digital cameras, PCs and smart mobile phones, or with sensors and other Internet of Things (IoT) devices. Consequently, the annual growth rate of storable digital data is probably significantly higher than 58 %. Any way, by combining a conservative 60 % annual growth rate and the 100 % share of digital data from all data created, one is able to calculate that almost 61 % of all storable data created by the mankind has, at any time since 2010, been created during the last two years. During 2018, the mankind creates every third month as much storable digital data as we created from 10,000 BC until the end of 2013 in any format. In 2028, the same happens every 20th hour, in 2038 every 11th minute and in 2048 every 6th second. From the estimates of Hilbert and Lopez (Hilbert and Lopez, 2011), it is also possible to calculate that in 2018 the mankind will create 49 zettabytes storable digital data and 4,555,005 zettabytes in 2043 (60 % annual growth).

Data explosion is not just a volume issue although the number of information systems (IS) has exploded similar to digital data. Until this millennium, the majority of digital data was structured internal data processed with the internal ISs of an organization. Even enterprise resources planning (ERP) and other ISs that were purchased from IS vendors were managed and governed in the same way as internal ISs. Data were stored into the databases of those ISs and cumulated into data storages, such as reporting data vaults. Stored data consisted of transactional data, reports, documents and contents, to which master and reference data were linked (Cleven and Wortmann, 2010; DAMA, 2009; DAMA, 2017). The most important data management issue was that organizations knew and controlled the data models and the designs of the data storages they used. We call this the closed systems period.

In data management, organizations face now the challenges of what we call the open systems period. They have lost partially or totally the ability to know and control the logical and physical data models and the designs of the data storages they use. Self-developed ISs and closed software packages have been replaced with more open ISs developed for multiple organizations and/or with software and integration platforms. Independent IS service vendors are responsible for the data modelling of these ISs and platforms. User organizations acquire and deploy them as IS services - increasingly from clouds. In addition to processing and recording business transactions, organizations create digital data with sensors and other devices in their manufacturing, logistics, and other processes. Communication and message data, social media data, audio and video data, and analytics data are the data assets of these new data sources. Data available to an organization have enlarged to include unstructured and multi-structured data as well as additional dimensions of data, such as spatial and temporal. Data are also increasingly external to an organization or shared between organizations, such as data transmitted between ecosystem partners, for example between a buyer and a seller. Cumulatively, for (large) organizations, there are data about the same persons and organizations, facilities and locations, products and services, accounts and other concepts in dozens, hundreds, or even in thousands of data storages. Data in them differ in format, structure, granularity, and in other characteristics, including their meaning.

The role of master data is centric for data integration and/or federation, and for the resulting data interoperability. With master data we understand non-transactional data that is shared between ISs, for example, customer data and product data (Loshin, 2010). The generic research problem of the present article is: how should the complexity of master data be addressed in IS development, especially in data integration and/or federation projects. We see a data management research gap here, which the transformation from closed systems period to the open systems period has created.

The other source of motivation for this article comes from a recent case study where we investigated the failures of IS development (ISD) projects (Dahlberg and Lagstedt, 2018). In that study, one of failures happened in a publicly listed corporation with operations spread to over 200 locations in over 70 countries with close to 20 000 employees. The company wanted to develop and roll out a new product data management (PDM) IS to all of its business and geographical units. The company's product and service portfolio had been streamlined after several mergers and acquisitions, whereas ISs harmonization and integration had been postponed. From the company headquarter executives' perspective employees in all units had similar standing orders, manufactured similar products and offered similar customer services. Business processes appeared mature and product data unified to them. The ISD project was deemed a legacy ISs replacement project that would deliver a "one company PDM solution" by harmonizing product data with no need for new functionalities. The executives of the company considered the ISD project business critical and gave their strong support. The project team members were experienced and had good understanding of the methods used in the investigated ISD project.

Requirements collection and specification was an enormous task at the beginning of the new PDM ISD project. Multiple teams from the diverse business units and geographical locations of the company were engaged to do that. The assumptions that legacy ISs could be replaced without functionality and business process enhancements were challenged almost immediately, and were among the key failure reasons of the project. This and other ISD method related failure reasons as well as the data collection and analysis methods used in the case study are described in details in (Dahlberg and Lagstedt, 2018). In this article, we focus on the following issue: How should the complexity of master data, especially data integration / federation, be addressed in ISD projects to ensure data interoperability?

The insufficiency of the so-called golden record, "single/best version of truth", approach (DAMA 2009, DAMA 2017) and the single data domain, "product data", focus were the detected master data management related failure reasons of the failed PDM ISD project. Instead of mature processes and unified data, the company's business and geographical units had dissimilar processes that created the diversity of the data models in their legacy ISs. Only the products and services offered to customers were commensurate. Despite of these data modelling challenges, the PDM IS was specified and its development was carried out. The golden record approach soon led to serious problems. The "unified global master data" was a new concept to the users of product data in the company. They were familiar with their "local master data" models. The technical properties and semantic meanings of seemingly similar product data entities and attributes differed remarkably in local master data models. Later, during the implementation phase, these process and data inconsistencies created invincible data migration problems between the legacy ISs and the new PDM IS. Business and geographical units were unwilling to use the new PDM IS when they discovered that almost all employees would need to change fundamentally their way of working. That was a surprise to the IS developers, the project management of the PDM IS and to the company headquarter executives. The PDM IS was never taken into use.

In this article, we contemplate whether the federative approach to data management and governance (Dahlberg and Nokkala, 2015; Dahlberg et al., 2016) offers a viable alternative to the golden record approach in ISD (integration) projects. We investigate this question especially in situations where master data from multiple ISs and organizations are integrated / federated in order to make the data of data storages interoperable. We define data federation as the federation of two and usually more data storages and/or data sources, when data interoperability of those storages and/or sources was not considered at the time of data modelling, IS development and/or the daily operations and use of the data.

The purpose of this article is to investigate the above described research gap in addressing the complexity of master data in IS development projects by comparing the golden record and the federative approaches. The other purpose is to show how the federative approach has been and could be used in ISD projects, especially in data integration and interoperability projects. From the generic research problem and the purposes of our study we formulated the following research questions for this study:

RQ1: How does the golden record approach address master data complexity in ISD projects?

RQ2: How does the federative approach to data management and governance address master data complexity in ISD projects?

RQ3: What kind of tools does the federative approach to data management and governance offer to ISD to federate the (master) data of multiple data storages / sources to achieve data interoperability?

In next chapter, we compare the ontological stances of the golden record and the federative approaches and illustrate their comparison with two figures. Chapter three describes the methods we used to collect and analyze data in a university hospital case. We then present the tools we used to federate data in the case, and to support the data integration/federation of ISD projects in general at the university hospital. We end the article with a discussion and conclusions chapter.

2 Theoretical Background and Ontological Stances of the Golden Record and the Federative Approaches

Efficient data processing, consolidation, analytics, federation and integration are probably the most important properties of digital data that are infeasible with analog data. Data processing, consolidation and analytics are used to cope with large amounts of data and to extract meaningful information from raw data, for example, into managerial reports. Data analytics supports decision-making, and has led to the emergence and growth of algorithm-supported and algorithmic decision-making.

Data analytics and algorithmic decision-making can be built "easily" on data (extracted) from single data storage with a known data model and data structure. Yet, it is usually possible to obtain richer and deeper insights by integrating or federating data from multiple data storages, that is, by combining multi-source, multi-format and multi-dimensional data. The main challenge is, how to integrate or federate data so that this does not compromise data quality and/or lead to erroneous algorithmic decisions caused by missing, erroneous, incomplete and in other ways low quality data, or caused by the mixing of semantically different meanings (Newell and Marabelli, 2015). For example, differences in the granularity and semantics of product data between a buyer and a seller may lead to significant amounts of non-productive manual work. The research project described in the last Chapter of this article addresses these issues. Its aim is to automate and integrate the transfer of product and other supply-chain logistics information between industrial ecosystem partners by applying agreed UBL standard based messages, open source reference API programs and blockchain code governed by smart contracts.

The significance of data interoperability in master data and other shared data, the topic of this article, has increased as one of the consequences of digital data explosion. Data integration or federation between multiple data storages with inconsistent (master) data properties often requires IS development. But how should the complexity of (master) data be addressed in IS development? The theoretical description of two alternative approaches shows that differences in their data ontological stances lead to differences in, how the complexity of master data is addressed in ISD work (Wand and Weber, 1990; Wand and Weber, 1993: Wand and Wang, 1996; Wand and Weber, 2002; DAMA, 2009).

2.1 The Golden Record Approach and IS Development

The master data management (MDM) concept was introduced some 15 years ago as the means to consolidate, cleanse and standardize fragmented product, customer and other master data (DAMA, 2009). The first efforts simply brought data (storages) together. These efforts failed to produce much progress. The golden record approach then emerged as the solution to the problem, as of what to do with inconsistent and fragmented master data. The golden record approach has dominated MDM research and practice during the recent years (see, e.g. DAMA, 2009; DAMA, 2017; Dreibelbis et al., 2008; Berson and Dubov, 2007). The data management body of the knowledge method (DMBOK) (DAMA, 2009; DAMA 2017), which is representative to this mainstream data management practice, advocates the golden record approach. "*A golden record is a single, well-defined version of all the data entities in an organizational ecosystem. In this context, a golden record is sometimes called the 'single version of the truth,' where 'truth' is understood to mean the reference to which data users can turn when they want to ensure that they have the correct version of a piece of information. The golden record encompasses all the data in every system of record (SOR) within a particular organization*" (Whatis, 2013). The DMBOK handbook (DAMA, 2009, p. 173) states in a similar tone that "*master data management*

requires identifying and/or developing a 'golden' record of a truth for each product, place, person, or organization. In some cases a 'system of record' provides the definitive data about an instance … Once the most accurate, current, relevant values are established, master data is made available for consistent, shared use across both transactional application systems and data ware-house/business intelligence environments".

The ontological assumption of the golden record approach is that it is possible to define and agree on one version of truth so that all data entities and data attributes mean the same in different data usage contexts. The golden record approach could therefore also be called as the canonical approach to (master) data. The practical imperative is to establish canonical data models where context and time specific values of the golden record attributes are considered purpose-specific anomalies and replaced with the golden (canonically true) data values, as in the PDM IS case. The DMBOK method (DAMA, 2009, p. 171) explains this in the following way: "*Purpose-specific requirements lead organizations to create purpose-specific applications each with similar but inconsistent data values in differing formats. These inconsistencies have a dramatically negative impact on overall data quality*."

In the golden record approach, master data could be processed and managed with a separately developed MDM IS or by developing additional MDM functionalities into so called master information systems. For example, an organization might agree that its PDM (product data management) IS the master IS for product data. Consequently product data should be created and maintained in the PDM IS. Product data should then be transferred (=populated) to all other ISs from the PDM IS. Such a decision usually requires that extra MDM functionalities be added to the PDM IS. Terms match, merge, and cleanse and transform (=replace) describe how the golden record approach addresses the ambiguity of master data in IS development projects. During our research, we collaborated with a few master data IS service vendors, most notably with Ineo ltd., to understand how product and other master data are migrated, integrated and made interoperable. According to our research and practice based understanding, the steps to address the ambiguity of master data in IS development are those shown in figure 1:

1. Match: List the attributes of (all) data storages within a master data domain, for example, the product data domain. Detect from the lists, for example, database scripts, similar attributes (=match).

2. Merge: Determine the attributes of the golden record by identifying attributes with most matches. Develop IS functionalities to merge those attributes to the golden record / the master system IS.

3. Clone and transform: Cleanse merged data by determining the values of golden records and clone them to (all) integrated data storages. The golden record true values should be used to replace purpose-specific (anomalous) values of (all) integrated ISs. After that, new master data records and record value modifications should be created to the golden record / the master system IS and transferred (populated) with possibly needed transformations to the other integrated ISs.

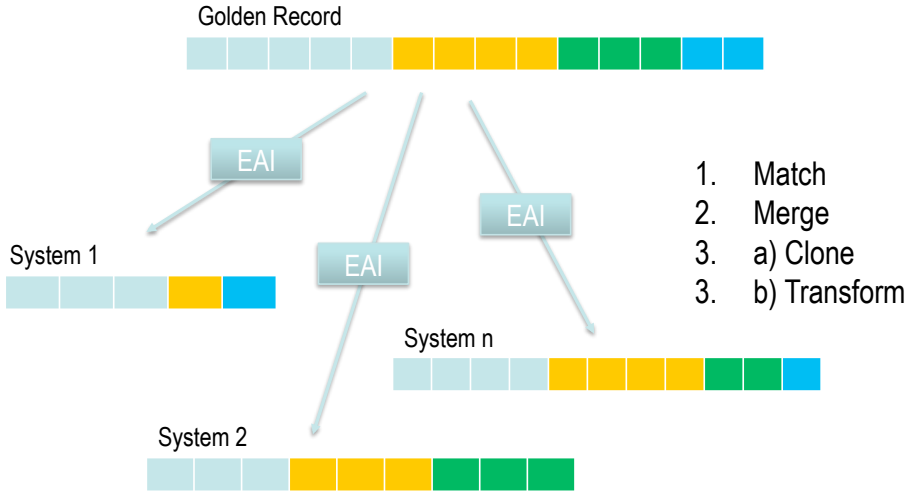

Figure 1. The golden record approach to master data (together with Timo Seppänen, Ineo Ltd). EAI means enterprise application interface technologies, tools and services

2.2 The Federative Approach to Data Management and IS Development

Although the golden record approach has made significant improvements possible, MDM solutions have remained fragmented instead of being organization-wide (Dahlberg 2010; Dahlberg et al., 2011). As a whole, the golden record approach has been unable to deliver the promised solution, that is, organization-wide interoperability and/or integration of customer, vendor, employee, product, and other data facilitated by well-managed master data. In many organizations, this approach has motivated the execution of large single-domain MDM development projects, especially to harmonize product or customer data. The golden record approach based canonical ontology motivated proposition has been that multi-domain projects are impossible to execute because they are too complex and large during the match and merge phases. The benefits of even successful single-domain MDM development projects have, however, remained lower than estimated. One likely reason is that business transactions, reports and contents most often include multi-domain data (Cleven and Wortmann, 2010). For example, consider a situation where a sales-person (employee) sells a product to a customer with an agreed payment term (P/L account). The employee, product, customer and P/L account represent four different (master) data domains. The value of data harmonization remains low, should only the quality of product master data be high achieved with harmonization. Data migration challenges have also been difficult to solve. The PDM case described in Chapter 1 and issues discussed in Chapter 2.3 are typical.

The limitations of the golden record approach detected in MDM projects and contextual stance to philosophical ontology have been the motives to develop the federative approach to data management and governance, and to introduce it as an alternative to the golden record approach (Dahlberg, 2010; Dahlberg et al., 2011; Dahlberg and Nokkala, 2015; Dahlberg et al., 2016). In some earlier studies, researchers argued against the view of one universal or a unified composite data ontology already in the 1980s. They proposed federated data solutions and architectures (see Heimbigner & McLeod, 1985; Sheth and Larson, 1990). The ontological stance of our federative approach, however, builds largely on the work of Wand and Weber (Wand and Weber, 1990; Wand and Weber, 1993; Wand and Wang, 1996; Wand and Weber, 2002). The federative approach could be seen as an attempt to bring the philosophical ontological stance and the ideas of Wand and Weber from the closed systems to the open systems period. Ontologically, the major difference to the golden record approach is that context and time specific values are not regarded as anomalies but are rather deemed as valid and true representations of different data usage contexts. Great care is recommended in replacing and removing context and time specific values, as that could lead to the loss of business critical data in relevant contexts.

From the philosophical perspective, an IS is ontologically an abstract representation of one or more real-life data usage contexts, that is, data is contextually defined. According to Wand and Weber (1993) the key principle in the design of a "good" IS is to strive for ontological and contextual completeness. Completeness exists when ontological constructs (things, their properties, and values) are mapped to the contextual design constructs. Three types of representational problems could hamper completeness. Firstly, construct overload could be present, that is, one design construct could map into two or more ontological constructs. For example, a buyer may acquire several products or a product may be acquired. Secondly, construct redundancy, that is, two or more design constructs may represent a single ontological construct. For example, an order may split into several design constructs such as customer and vendor transactions, data flows carried out with order forms and product catalogues, or to order data registered into several data storages. These design constructs are redundant for the ontological construct order but all of them could also be contextually valid and intentional. Thirdly, construct excess, that is, a design construct does not map into any contextually meaningful ontological construct. For example, non-functional properties of a product, such as the technical data attributes of a product item could be irrelevant to a buyer but need to be included or attached into the product item data since that data could be vital to the planning or manufacturing units of the buyer organization.

In principle, these problems can be avoided should the domain realm ontology map completely to the design realm. In practice, this is impossible since the data of an IS represents a specific context and the data of multiple ISs several contexts (Wand and Wang 1996). This results largely from the division of work. Consequently, with ISs an organization sees the real world through and from the lenses of dif-

ferent contexts (1...n). Due to this reason organizations typically developed separate ISs and data models for each IS usage need already during the closed systems period. For example, procurement, manufacturing, logistics, sales, and accounting (contexts) had their own ISs and data models. It was possible to transfer data between different ISs (and data models) and to make data interoperable because the differences in data models were known, and their logical and physical designs were controlled. Data transformations and mappings provided the means for data transfers and integrations, if needed.

During the open systems period, ISs are bought from IS vendors, and there could be several ISs and applications for any context, such as procurement. Data models and their designs are no longer under an organization's control and could change constantly. Moreover, cameras, mobile devices, sensors and IoT devices, as well as social media and other similar applications create data with different contextual and semantic meanings. Data deficiencies could cause serious problems to data interoperability, unless the ontological and contextual reasons for data deficiencies are understood. Differences in data structures, formats, and other data characteristics may even cause differences in how a user perceives the real world through direct observations and its various representations within ISs. Thus, during the open systems period, ability to understand the ontological contextual meaning of data increases in importance and is a major means to avoid representational and observational problems. The proposition of the federative approach is that data interoperability and/or integration is achieved through shared attributes and their metadata. According to our understanding, the steps to address the complexity of master data in IS development projects are those shown in figure 2:

1. Identify: Identify shared attributes between data storages that are federated to make data interoperable. Describe IS technical, data processing (=who create and maintain data) and contextual meaning (=for what purpose is data created and used) metadata properties of each shared attribute.
2. Define and develop cross-references between the shared attributes through their metadata.
3. Connect (=federate) data storages by defining business rules for metadata cross-referencing.

The federative approach proposes that MDM "projects" usually need to address multiple data domains simultaneously, although there could still be a particular single domain focus, e.g., product data. In other words, an MDM "project" is advised to concentrate on a specific (limited) use context, and to make the data of all domains interoperable within that context step by step. The logic is to ensure that concrete and validated data quality improvements are achieved constantly by first federating two data storages, by then adding a third, and so on. Master data governance, management and data quality improvements are carried out with continuous improvements rather than in a single project. That is, master data management and the federation of data are seen as a way of organizational life.

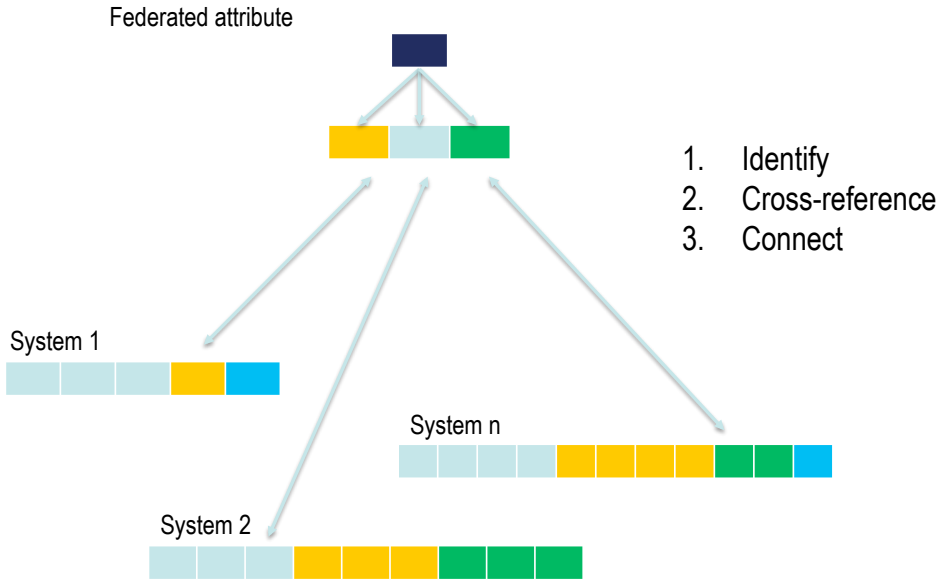

Figure 2. The federative approach to master data (together with Timo Seppänen, Ineo Ltd). Cross-reference and cross-mapping are synonyms in this approach are both terms are used in the text.

2.3 Examples of Golden Record Versus Federative Approach

Since the discussion in Chapters 2.1 and 2.2 is theoretical, we feel that the differences between the two approaches become clearer by looking at a few examples. A missing attribute, especially a missing mandatory attribute, is a typical data migration de-dupling challenge. For example, a new master data IS / master IS (e.g. PDM IS) may have product identification attribute as its mandatory search key attribute. This attribute could be missing from one or more ISs, from which data is migrated. If data migration is the only necessary data activity of the IS development project, then this challenge could be dealt programmatically, for example, with a data completion and conversion program. In data completion and conversion, business rules define the value of the missing attribute and its values on the basis of some other attributes and their values present in the migrated data. At the same time, the completed attribute is converted into the format of the receiving (golden record) IS. Alternatively, migrated data could be completed manually by adding the missing attribute value to each record after migration.

This challenge is more complicated should the usage of the data storages, from which data is migrated, continue after the migration. For example, the purpose could be to continue the use of local ISs and their local data attributes intact with the exception of the shared attributes, i.e., global master data. The solution offered by the golden record approach is to replace the shared attributes and their values in the local ISs with the golden record attributes and their golden values. The data models of local ISs need to be modified, if necessary. This solution is not only expensive but could be impossible unless the organization controls the logical and physical design of all relevant data models. That is unlikely during the open systems period. The federative approach suggests that completion and conversion rules are the solution in themselves, since they contain those shared attributes with their metadata descriptions, which are needed to make data interoperable or integrated. The use of such conversion rules could be made reciprocal to facilitate data transfers from local ISs to an MDM IS and vice versa, e.g., to allow the entries and modifications of product master data by those who best know the data.

Duplicates are another typical (master) data challenge. For example, a company could have registered 350 000 product items when the actual number of genuine product items is 100 000. Duplicates are a data quality error should there be only one data storage. Both approaches advise data harmonization and cleansing for this challenge. The federative approach advises extreme caution in the harmonization and cleansing of duplicates to avoid accidental removal of business critical data, for example an accidental merging of two open transactions. The federative approach considers duplicates in one data storage as a (user) data entry error. The proposal is to remove systematic and systemic reasons that cause double entries instead of jus cleansing data without removing the causes of defects.

This challenge is again more complicated should duplicates be the result of data transfers from several data storages. Duplicates probably reflect contextual differences between the data models of local ISs, MDM IS and/or Master IS. The logic of the two approaches is similar as in the example of the missing data attribute. Data conversion, that is, describing aliases and synonyms with the help of data conversion rules has been used for decades. According to the golden record approach, data conversion is one of the possible means to replace local values with true golden values. The federative approach suggests that data conversion rules (e.g. conversion tables) are the solution as they contain shared attributes with their metadata, and that conversion rules and tables could again be made reciprocal.

Synonym and changing data values are probably the most challenging complexity of (master) data. For example, a product item may have several permanent or semi-permanent prices: a procurement bill of material (BOM) price, a component price after manufacturing, and a spare part price used in after sales services. Product catalogues and/or price lists are among the means to communicate (semi-) permanent prices. Still, any of the prices may change over time at different intervals. Accounting research and practice have developed several models and methods to deal with these pricing issues, but here we are only interested about the (master) data management challenges. A new product item could replace an existing product item due to technological advancements, supplier changes or some other reason. Furthermore, each supplier could have a different coding scheme for the same product items.

The normative solution of the golden record approach, that is, the replacement of local values with the most-true golden record value cannot be applied in these situations. The alternatives are to treat differ-

ent prices (values) as local values or as additional attributes of golden records, and to promote (product) data value standardization. The downside is that the core principles of the golden record approach (DAMA 2009; DAMA 2017) are diluted and/or that the content of the golden record becomes either very generic with very few attributes or very broad with multiple attributes, many of which are empty. Moreover, these solutions are still unable to solve challenges of master (data) value changes over time and the changes of values. The issues depicted above have been the other motivation to develop the federative approach to data management and governance. The proposed solution to all issues is to define shared attributes, to describe their metadata, and to use the metadata to cross-reference shared attributes in order to make data interoperable or integrated. The remaining challenge is how to audit-trail attribute and metadata descriptions, and their changes over time. Some new technologies, most notably blockchain, also known as distributed ledger technologies, offer new means to solve this challenge.

3 Methodology - Probing of the Federative Approach

A MDM best practice benchmarking study (Dahlberg, 2010), discussions with MDM IS vendors and Gartner Inc., and the literature cited above were used above to describe the golden record and federative approaches. In addition to that, we wanted to probe the federative approach in an as complicated data management case as we were able to find. We supported the informaticists of a Finnish university hospital in making breast cancer data interoperable. In Finland, over 90 % of patients diagnosed with breast cancer between 2005 and 2012 were still alive five years after the diagnosis (Finnish Cancer Registry, 2017). Yet, metastasized (=widely spread) breast cancers kill patients rapidly, that is, the remaining 10 %. A large number of highly skilled medical, surgical, nursing, rehabilitation and other professionals participate into cancer treatments, and the data from a large variety of ISs is used. The university hospital's breast cancer specialists have access to enormous amounts of internal hospital data and to external data from other (healthcare) organizations. Still, the detection of malignant breast cancer cases is currently largely manual and relies on the expertise of the best specialists. The reason is that most data properties in relevant data storages differ. Data coding varies although the HL7 data standard is widely used in Finland and ranges from structured patient data and written unstructured medical reports, radiology pictures and ultrasound videos. Data sources vary from manual IS entries to real-time time-series created by IoT devices. The objectives of the case study were to find malignant breast cancers earlier than before and to evaluate the effectiveness of various cancer treatments.

The case study was conducted during the first half of 2016 (Dahlberg et al., 2016). Data collection was organized through workshops. A Finnish MDM IS vendor (Ineo) gave us the permission to use and develop further their data federation matrix tools. They had developed those tools in dozens of master data, mainly SAP, migration and implementation projects. We reorganized their tools, classified metadata properties into IS technical, informational and contextual categories and gave the improved tools back to the vendor. The vendor was not otherwise involved in the case. In this article, we present the matrix tools at a generalized level to protect the intellectual property rights of the MDM IS vendor and the hospital. Prior to the first workshop, we had three meetings with the informaticists. In these meetings, we planned together the execution of the workshops, and made them familiar with the federative approach and the tools. Then, in a workshop, we jointly interviewed a specialized group of breast cancer experts, such as pathologists or patient ISs support experts. Insights and feedback collected were used prior the next workshop to draft the next version of the data federation design with the tools. We followed the proposed steps of the federative approach shown in Figure 2 as follows:

- Step 1. Identify the most relevant ISs, ISs modules and data storages needed for data inter–operability. Identify specialists who know, how the data of an identified IS, IS module or data storage is created, processed and used to treat breast cancer patients and to detect malignant breast cancer cases. Invite the identified specialists to a workshop for a group interview.
- Step 2. Cross-reference identified shared attributes that make data interoperable between ISs, ISs modules and data storages. Start from two or a few ISs, ISs modules and data storages and increase their number as the work progresses and learning happens.

- Step 3. Connect the federated data storages by using the IS technical, information processing and contextual metadata defined for each shared attribute during Step 1.

We used these three steps iteratively. Thus it was possible to complement – both add and reduce – the number of ISs, ISs modules, data storages, shared attributes and metadata elements. For example, we identified initially three shared attributes and added the fourth attribute later. Similarly, encouraged by the tools we received from the MDM IS vendor we considered 30-40 metadata characteristics for each shared attribute in the beginning. We then noticed that (in this case) focus on a smaller number of the most important metadata characteristics was enough. It is worth to mention that several international standards exist for IS technical metadata, such as the ISO 19115, 19119 and 19139, whereas the information processing and socio-technical properties are not covered in ISO or other data standards.

In the collection of empirical data on the case, we followed the guidelines of Yin (Yin, 1994; Yin, 1999) and Eisenhardt (1989) for case studies and for the building of research constructs from case studies. As Eisenhardt (1989) and Yin (1999) mention, case studies can combine different data collection methods, including interviews, observations, and archival material. We wrote a case protocol (Yin, 1999) to guide empirical data collection. We used all the other data collection methods in the Yin's container (Yin, 1994) with the exception of direct observation (of cancer treatments).

4 Findings of the Case Study

During the case study, we designed and used step-by-step two data matrixes shown in Figures 3 and 4. They are illustrated in generalized formats. The informaticists found the matrixes easy to understand and use. The matrix tool of figure 3 was used during the first two iterative steps of the case. We first placed the data storages of information systems and, if necessary, ISs modules into the matrix as matrix columns. We then added the shared attributes into to matrix as matrix rows, and finally cross-mapped matrix columns and rows. Please, note, that the identification of shared attributes was preceded by the laborious task of compiling and checking the full attribute lists of data storages to be used in data federation. The informaticists conducted this pre-processing task, which was necessary to identify shareable attributes. Figure 3 shows the outcome of the iterations with generic IS and attribute names at the end of the case study. The data matrix tool was also used to determine that the shared attributes with attribute values really existed or that the attribute and attribute values could be deduced from some other attributes and their values in a data storage by using appropriate deduction rules.

	Patient IS	Laboratory IS	Surgical IS	Radiotherapy IS	Pathology IS	Information System N
Social security identification	X	X	X	X	X	X
(Cancer) diagnosis code	X	X	X	X	X	X
Tumor node metastasis (TNM) code	X	X	X	X	X	X
Date of events	X	X	X	X	X	X

Figure 3. Tool to identify shared attributes, X means that an attribute exists or can be deduced

The identification of shared attributes proved to be an easy task for the informaticists and also made sense to the interviewees during the workshops. The matrix tool helped them to compile shared attributes from all the federated ISs, ISs modules and data storages into a single table. The approach to start from a few data storages and then include additional data storages also proved useful. It was a surprise to us that only four attributes and (physical) database scripts (=attribute lists) were needed to make cancer data interoperable. The combined logic of the four shared attributes is that a person has been diagnosed to have breast cancer, which can be malignant only if all events happen within a short time period and have a certain tumor node metastasis (TNM) code. The TNM code does not exist in any

data storage prior the positive diagnosis of a malignant breast cancer. After that the TNM value is updated to relevant data storages. It is, however, possible to determine the probability of TNM code typical to malignant breast cancers with deduction rules, that is, to collect evidence with algorithms.

The matrix tool shown in Figure 4 was also used during the two first steps of the case. Tables 3 and 4 provide necessary data to connect and cross-map data storages technically. This matrix describes the three types of metadata characteristics for each shared attribute. Figure 4 illustrates the metadata characteristics of the social security ID attribute in a generalized format, such as field/attribute length, definition of initial attribute value entry process, or definition of the semantic meaning of the attribute. Please, note, that all of attribute values could have different meanings in the various data storages. The characteristic values of the matrix are used to support the cross-referencing between data storages.

Social security identification	Patient IS	Laboratory IS	Surgical IS	Radiotherapy IS	Pathology IS	Information system N
IS technical metadata						
Field length						
Other properties						
Information processing metadata						
Initial entry						
Other properties						
Socio-contextual metadata						
Definition of the meaning						
Other properties						

Figure 4. Tool to describe the metadata elements of shared attributes for cross-referencing

The content of the cells in the data matrix of Figure 4 provide answers to the following questions:

- What IS technical metadata characteristics does the shared attribute have (format, length, hierarchy, granularity, mandatory attribute, search key, location,…)?
- What data and information processing metadata characteristics does the shared attribute have (data type from processing perspective, source or origin, level of structure and other dimensions, persons and processes responsible for data entry, using and updating, purging…)?
- What semantic socio-contextually determined meanings of metadata characteristics does the shared attribute have (meaning in each use context, purpose of creation, changes of meaning during the attribute's life-cycle such as entry, using and updating, purging…)?
- Who is responsible for the management and governance of the shared attribute (each metadata element including information security, privacy and data quality)?

The classification of metadata into the three categories and the inclusion of management and governance accountabilities were welcomed by the informaticists. Unclear or missing accountabilities, variations in data processes, and differences in the contextual meanings of data attributes were seen as typical data quality challenges that had previously prevented cancer data interoperability thinking at the university hospital. In some situations, it was unclear who was responsible for data processing activities. In other situations, the same data processing activity could be carried out in different ways depending on the person executing the activity. These situations usually resulted in data quality defects. The federative approach, and especially the matrix tools were able to reveal some of those defects.

In some situations, contextual semantic metadata descriptions were needed to describe the differences in interoperable data values. For example, the layman's rule of thumb for the normal human body temperature is 37 degrees Celsius (98.6 degrees Fahrenheit). Yet, after sleep, stressful activity, or (breast cancer) surgery a clearly lower or higher body temperature is normal. In addition, the measurement device, the method, and their calibrations as well as the contextual characteristics of meas-

urements, such as persons lying, sitting or standing influence the detected body temperature values. Rectal, vaginal, optic, oral, and axillary measurements are known to produce systematically different body temperature values (Kelly, 2006). In some units of the hospital, nurses reduce habitually the values body temperature measures with x.x degrees due to patient-care-related reasons, whereas the nurses of other units do not. The federative approach was able to capture (some of) these data ambiguities.

In summary, the federative approach distinguishes itself from the golden record approach and canonical data integration endeavours in general by suggesting that the compilation of data into single data storage (vault) is only one possible option to make data interoperable. Another option is to let the original data reside where they are and to make data interoperable by federating data on the basis of shared attributes' metadata and by cross-mapping the shared attributes. The two matrix tools shown in Figures 3 and 4 were developed to do that. From the IS development perspective, this suggests the development of an MDM and metadata repository IS is enough in such situations. The metadata repository IS contains data federation rules, the meanings of shared attributes, descriptions of data formats, other metadata characteristics definitions and the database scripts of data storages made interoperable. Metadata descriptions are created and updated only when new data interoperability needs emerge. Similarly, new metadata and cross-mapping descriptions can be added whenever needed, for example, to fulfil a new management reporting need. This also means that so-called big bang projects with single domain focus can be avoided. Instead of that, data interoperability is developed continuously at the pace of organizational learning. (MDM) projects could speed up such continuous development. The logic of the metadata-enabled cross-mapping is to make all attributes of federated data storages potentially interoperable in their original meanings. (Physical) database scripts are needed to support that.

5 Discussion and Conclusions

As the motive of the present article, we depicted digital data explosion and how ISs user organizations have lost logical and physical design control over the data models and ISs they use during the open systems period. Data is increasingly external and provided as an inseparable part of IS services along with unknown and rapidly changing data models. Despite of these developments, the ability to integrate or federate data is increasingly important for organizations as the amount of digital data continues to grow at a breath-taking speed. We also briefly described how the complexity of (product) master data contributed to an IS development project failure in a large global company. We discovered a research gap in how to address (master) data complexity in IS development and compared two ontologically different approaches to fill this research gap.

The ontological assumption of the golden record, the canonical, approach is that it is possible to find and agree single true values for the attributes of golden records. In other words, each physical record of the golden record data storage contains the truest values for the attributes of that physical record. The proposition is that the anomalous values of local ISs need to be replaced with the true golden values after the golden records have been created. The golden record approach addresses master data complexity by first matching the attributes of integrated data storages. The matched attributes are then merged to establish the golden record and the golden values of each attribute's physical records are determined and agreed. Finally, the truest golden values are cloned and transmitted to replace the attribute values of local ISs data storages. This is our answer to the first research question.

The contextual ontological assumption of the federative approach to data management and governance is that the data models of local ISs data reflect different true and valid real-world contexts. Replacements of local ISs data values with golden record data values may therefore delete business critical data, and should be avoided unless there are contextually determined reasons for that, i.e., there is (only) one context. The purpose of data federation is to make data interoperable. The federative approach addresses master data complexity by first identifying shared attributes between federated data storages. Cross-mappings between the shared attributes are then defined and implemented with the help of shared attributes' IS technical, information processing and semantic socio-contextual properties. The one- and especially bi-directional cross-mapping rules facilitate data transfers between data storages.

The attributes of the federated data storages become interoperable with the help of their database scripts or schemas with their original values. This is our response to the second research question.

We described two matrix tools that were designed and used in a case study. The objective of the case study was to make diverse breast cancer data interoperable through data federation in order to detect malignant breast cancers earlier and to evaluate cancer treatments' effectiveness. We developed the tools from earlier versions of similar tools used by a Finnish MDM IS vendor in other contexts, such as food and technology manufacturing as well as whole and retail commerce industries. The tools made the federative approach operationally useful in the real-world context of the case. The informaticists of the hospital found them easy to use. This is our response to the third research question.

Based on our experiences, the major intellectual difficulty of the federative approach lies in understanding the practical consequences of the approach's ontological stance. Most people agree intuitively with the statement that the meaning of data is contextual and should not be replaced (with a golden value) unless there are compelling reasons to do so. They also agree with the statement that the variations of data processes are most often the reasons of data quality defects. At the same time, the same people still attempt to enforce single values to the attribute values of ISs and propose that this is the only way to collect harmonized data into large databases. They also try to solve data quality defects with data cleansing projects without changing the reasons that are the causes of these defects.

The obvious question is, why? One possible answer is that ISs professionals and users are accustomed to the canonical data models of the ISs they develop and/or use. Another related possibility is that they have not been aware of alternatives, such as the federative approach. To develop any single IS, it is mandatory to define a canonical data model for the real-world context of that IS since any data attribute can have only one meaning in a specific context. In chapter 2, we discussed possible design defects in the mapping of real world and IS constructs (Wand and Weber, 1993). DMBOK and its established data management methods are useful for this. However, that does not mean that all real world contexts and their data models are similar, or that it would be possible to integrate them into one canonical data model without losing contextual meanings. In our opinion, the most important conclusion of our study is that data complexity of (master) data needs to be addressed differently in the IS development of a single ISs and in the integrating or federating of multiple ISs. The two approaches appear complementary to us. Determining suitable use contexts for each approach is an amenable topic of future studies.

The industrial corporation discussed in Chapter 1 participates into a 2,5 year research project that started in August 2017. Cumulatively three universities, over 40 industrial companies and the Business Finland are engaged and funding two sister projects. One of the main objectives of the projects is to develop automated and integrated supply chain data interoperability to manage and govern the product data life cycle within these industrial ecosystems. In addition to the federative approach, we build that research on standardized (=industrial ecosystem level agreed) process and data models (Korpela et al., 2016) taken from the OASIS/ISO UBL 2.2 standard. The 300 attributes of the data model establish the core of shared attributes to which data from the ISs of each participating company is federated. For the industrial corporation discussed in Chapter 1 these research projects offer the opportunity to pilot with the federative approach in a business ecosystem context. In these research projects, we as researchers are able to conduct additional case studies in international contexts and extend the use of the federative approach to the ecosystem level. With such additional studies we address the main limitations of the present study (single case, single country, limited data). Despite of these limitations, we suggest that the present study contributes to research by showing how to address the complexity of master data in IS development projects, by linking MDM and data management issues into IS development, and by demonstrating how the federative approach can be used in real-life IS development projects.

Our findings suggest also other propositions, which are suitable for future research. Ontological and other differences between the federation, integration, management and governance of single and multiple data storages in open systems environments offer ample opportunities for research. Our advice to researchers is to deploy these opportunities. Our advice to practitioners is to seek alternatives to the golden record approach in efforts to make data interoperable between multiple data storages, and to avoid large conceptually oriented single-domain (MDM) governance projects.

References

Berson A. Dubov L (2007) Master data management and customer data integration for a global enterprise. McGraw-Hill, New York.

Cleven A. Wortmann F (2010) Uncovering Four Strategies to Approach Master Data Management. System Sciences (HICSS), 43rd Hawaii International Conference, IEEE.

Dahlberg, T (2010) Master Data Management "Best Practices" Benchmarking Study 2010 – Publicly Available Report. Aalto University School of Business (Helsinki School of Economics Available via Internet from Researchgate.com, DOI: 10.13140/2.1.4201.9849

Dahlberg T. Heikkilä J. Heikkilä M Framework and Research Agenda for Master Data Management In Distributed Environments. The Proceedings of IRIS 2011 Conference, Volume: TUCS Lecture Notes No 15, pp. 82-90. ISBN 978-952-12-2648-9, available from http://tucs.fi/research/publication-view/?pub_id=IRIS2011[1]

Dahlberg T. Nokkala T (2015) A Framework for the Corporate Governance of Data – Theoretical Background and Empirical Evidence. Business, Management and Education, 13:1, pp. 25-45.

Dahlberg T. Nokkala T. Heikkilä J. Heikkilä M (2016) Data Federation by Using a Governance of Data Framework Artifact as the Tool - case clinical breast cancer treatment data. Information Modeling and Knowledge Bases XXVIII, in vol. 292 of Frontiers in Artificial Intelligence and Applications.

Dahlberg T. Lagstedt A (2018) There Is Still No "Fit for All" IS Development Method: Business Development Context and IS Development Characteristics Need to Match. Proceedings of the 51st Hawaii International Conference on System Sciences, pp. 4803-4812.

DAMA (2009) The DAMA Guide to the Data Management Body of Knowledge DAMA-DMBOK Guide. Technics Publications, LLC: Bradley Beach, NJ, USA.

DAMA (2017) The DAMA Guide to the Data Management Body of Knowledge DAMA-DMBOK Guide 2nd Version. In print.

Dreibelbis A. Hechler E. Milman I. Oberhofer M. van Run P. Wolfson D. (2008) Enterprise master data management an SOA approach to managing core information. IBM Press/Pearson plc: Upper Saddle River, NJ.

Eisenhardt K.M (1989) Building theories from case study research. Academy of Management Review, 14:4, pp. 532-550.

Finnish Cancer Registry (2017) available from http://www.cancer.fi /syoparekisteri/en/?x56215626=112197488

Heimbigner D. McLeod D (1985) A federated architecture for information management. ACM Transactions on Office Information Systems, 3:3, pp. 253-278.

Hilbert M, Lopez P (2011) The World's Technological Capacity to Store, Communicate, and Compute Information. Science 332:6025, pp 60-65.

IDC (2011) Overload: Global Information Created and Available Storage. http://www.idc.com.

Kelly G (2006) Body Temperature Variabilty: A Review of the History of Body Temperature and Its Variability due to Site Selection, Biological Rhythms, Fitness, and Aging. Alternative Medical Review, 11:4, pp. 278-293

Korpela K. Mikkonen K. Hallikas J. Pynnönen M. (2016). Digital Business Ecosystem Transformation: Toward Cloud Integration. In: Proc. Annu. Hawaii Int. Conf. Syst. Sci.

Loshin D (2010) Master Data Management. Morgan Kaufmann.

Newell S. Marabelli M (2015) Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of "datification". The Journal of Strategic Information Systems, 24:1, pp. 3-14.

Sheth A.P. Larson, J.A. (1990) Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys, 22:3, pp. 183-236.

Wand Y Weber R (2002) Research commentary: information systems and conceptual modeling—a research agenda. Information Systems Research, 13:4, pp. 363-376.

Wand Y. Wang R Y (1996) Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39:11, pp. 86-95.

Wand Y. Weber R. (1990) An ontological model of an information system. IEEE Transactions on Software Engineering, 16:1, pp. 1282-1292.

Wand Y. Weber R. (1993) On the ontological expressiveness of information systems analysis and design grammars. Information Systems Journal, vol. 3, no. 4, pp. 217-237.

Whatis (2013) WhatIS Glossary. Available online from Internet: http://whatis.techtarget.com/definition/ golden-record

Yin R.K (1994) Discovering the future of the case study method in evaluation research. Evaluation Practice, vol 15:3, pp. 283-290.

Yin R.K (1999). Case study research, design and method. 4th ed. Thousand Oaks, CA: SAGE Publications.