# Data Federation By Using a Governance of Data Framework Artifact as the Tool –

Case Clinical Breast Cancer Treatment Data

Tomi DAHLBERG[a,b,1], Tiina NOKKALA[a], Jukka HEIKKILÄ[a], Marikka HEIKKILÄ[a]
[a] *Turku School of Economics, University of Turku, Turku, Finland*
[b] *Aalto University Business School, Helsinki, Finland*

**Abstract.** Widely spread breast cancer takes patients to an early grave. Early detection and ability to predict the effectiveness of treatments are among the means to fight this malignant disease. Data federation from dozens of data sources is needed for data analytics. The granularity, internality, structure and all other characteristics differ in federated data. We discuss alternative approaches to data federation and their theoretical basis, especially the ontology and governance of data. We developed an artifact in our on-going research. The artifact is used to support the federation of cancer data at a university hospital. We detected that our federative approach and the artifact improved the interoperability of data in the case. We suggest that our approach is capable to that also in other contexts.

**Keywords.** Data federation, Ontology of data, Master data management (MDM), Meta data, Governance of data, Medical data governance and management

## 1. Introduction

Ninety per cent of patients survived breast cancer five years after their confirmed positive diagnosis in the country of the present study (Finland) during the years 2005 and 2012 [9]. Still, breast cancer is one of the most common cancers taking women and also some men to an early grave, since it is one of the most common cancers. The number of diagnosed breast cancers has grown constantly since the 1950s, from 900 annual confirmed diagnoses to the current close to 5000 annual confirmed diagnoses [9]. The case investigated in our research is a data management project at a university hospital, which attempts to improve the survival rate of so called widely spread breast cancers. This malignant form of breast cancer accounts for most of the deaths. The project has two data management related objectives. The first is to improve the early detection of widely spread breast cancers. The second is to enhance the precision of predictions about the effectiveness of various cancer treatments. In other words, the other objective is to identify treatments that are in isolation or combined reliable predictors for (higher) survival probability. The project is expected to improve data analytics capabilities of data/information specialists, doctors and other stakeholders working at the university hospital. Advances in data analytics capabilities are deemed to require better integration - or as we call it - federation of data from multiple data storages with significant data inconsistencies. Our study aims to respond to that need.

[1] Corresponding Author, Information Systems Science at the Department of Management and Entrepreneurship, Turku School of Economics, Rehtorinpellonkatu 3, 20500, Turku, Finland

Federation of cancer data meets several challenges. A breast cancer patient could be treated in several units of healthcare institutions employing a wide range of healthcare specialists. Breast cancer could be detected in a periodic breast cancer sieving study, at a doctor's appointment in a public, private or occupational healthcare center, or during the treatment of some other disease(s). The detection and treatments of breast cancer generate a myriad of different data elements such as magnetic resonance and X-rays images, laboratory test and pathology analysis results, cytostatic, medication and/or surgical treatment medical reports as well as a lot metrics, referrals, prescriptions, analyses, diagnoses, reports and other data. Relevant data could spread out to several years and could even include genetic data about close relatives and/or data about the patient's life style and social environment. Clerical personnel, nurses and doctors participating to the various cancer detection and treatment tasks create, use, modify and store data about treatment events into dozens of information systems (IS), IS modules and related data storages The consequence is that the IS technical, information handling and socio-contextual characteristics of federated data differ. The entity and attribute definitions, formats, hierarchies and granularities of data storages are different. Data could be structured, unstructured or multi-structured. Data could be represented in numeric, alphanumeric, audio or video format(s). Data creation, use, storing and purging procedures as well as data volumes and velocity vary. The sources of data range from ISs to sensors and from internal to external data storages (e.g. code registers). Furthermore, data is used for different purposes in various use contexts at a time and over time and may thus have several valid contextual meanings.

The motivation of our on-going study is to support the data/information specialists of the university hospital to achieve the data management objectives of their project. To accomplish that, we design(ed) together an artifact to federate cancer data by applying the governance of data framework proposed by Dahlberg and Nokkala [5]. This article depicts the design and the use of the artifact as well as discusses the theoretical basis of the framework behind the artifact, called the federative approach to data management and governance, later the federative approach. The federative approach and comparable artifacts have been used earlier to federate data in other contexts, such as to federate dairy product master data, car reseller product and customer master data, construction company financial data and waste management service contracts data. Thus the purpose of our research is to investigate the applicability of the federative approach in a very demanding data management context and to collect empirical evaluative data about the usefulness of the framework/approach in various data federation contexts.

With data federation we understand activities that facilitate the simultaneous use of data from storages, which have different IS technical, informational and socio-contextual data characteristics. The underlying idea of the federative approach is to make data storages interoperable through data storage cross-mappings by first identifying shared attributes and by then describing the IS technical, informational and socio-contextual meta data (i.e. cross-mappings) of those attributes. The federative approach builds on a contextual stance to the ontology of data extended to open systems environments and on paying more attention to the governance of data. Contextual stance to data ontology means that data is considered to represent truthfully the social use context(s) of data [20, 21]. Data may have several meanings representing each use context to the extent that some of the meanings could be contradictory. The contextual stance differs from the canonical stance to data ontology [6], which proposes that it is possible to agree "one single version of truth" for data values and

then use those values in all contexts. The open systems environment concept describes the transformation from the use of a few internal data storages with structured data to the increased use of multiple data storages with also external and multi-structured data from varied data sources. This transformation is one of the reasons for the need to govern data better. In summary, the objective of our research is to respond to the following two research questions:

**Research Question 1 (RQ1)**: In general, how does the federative approach to the management and governance of data support data federation in open systems environments, and, in particular, to detect widely spread cancer cases earlier and to predict better the effectiveness of cancer treatments.

**Research Question 2 (RQ2)**: What are the impacts of the designed artifact on the federation of breast cancer data?

The rest of the paper is organized as follows. First, as the theoretical background, we discuss data ontology and the governance of data by comparing contextual and canonical stances to data ontology. We then explicate methodological issues. In section four, we present our design artifact, its use to federate breast cancer data, and other results of the study. We end the article with a discussion of its scientific and practical contributions and conclusions for researchers and practitioners.

## 2. Theoretical Background

### *2.1. From closed to open systems environments with data federation consequences*

Until this millennium the majority of digital data was structured internal data created and processed with the internal ISs of an organization. Internality and structuredness of data characterized also data stored into the data storages of ISs purchased from software vendors or outsourced ISs. Data consisted of transactional data, reports, documents, and contents to which master, reference, and meta data were linked [2, 6]. We call these closed systems environments.

In a closed systems environment, an organization took the responsibility over its own ISs and thereby knew the meaning of data stored [6]. Although data models were IS specific, that was not a problem. Differences in data models, for example in patient data, were known and justified by the differences in the use contexts of specific IS. Since differences were known, partly since their number of ISs was limited, it was possible to link/federate different sets of data, should that be needed, or to consolidate and cumulate data into data vaults [6]. The "information architecture" of ISs could be designed so that there were no unnecessary data and/or data model overlaps. Another benefit of the closed systems environments was that also the data interfaces were closed. That is, there were clear specifications on what data each IS was able to receive, use, produce and submit. Such definitions were also determined – or bound - early. Thus it was possible to interchange and integrate/federate data by applying the data models of each IS. Initially data integration was done with transfer files, typically via separate batch processes. Application programming interface (API) tools and open APIs emerged later to help data integration.

This is no longer the situation. Organizations have replaced self-developed IS with software packages and services purchased from independent IS service vendors. The numbers of IS used in organizations and the volumes of digital data have exploded during the last 15 years [11]. In addition to business transactions organizations create

digital data with sensors and other digital devices. For example, hospitals' operating rooms are equipped with numerous devices that record digital data about operations and the operations environment. Data available to an organization has enlarged to include unstructured and multi-structured data, such as communication and message data, audio, video and analytics data. In healthcare organizations, text mining (e.g. from medical reports) and image processing are widely used. The significance of spatial and temporal dimensions of data have also become more important (e.g. patient condition surveillance time series after a surgical operation) as opposed to traditional clerical IS, which often provide data about the current status of a transaction only. Data is also increasingly external to an organization, an organizational unit, and/or is shared between organizations. Consider, for example, patient data transmitted between hospitals and other healthcare organizations. We call these open systems environments.

When IS and data storage solutions are acquired as packages or as (cloud) services in open systems environments, the data models and interfaces of data storages are given and incompatible between acquired packages/services. By having done this user organizations have transferred the responsibility of data modeling to IS vendors [4]. The data model of a popular commercial software package represents the model of a generic user organization (over the functionalities of the software). This explains, why there could be incompatibilities between different instances of the same software package in a single organization [3]. Furthermore, a user organization may not even have access to the data model of a software package. A software vendor makes decisions about changes into the generic data model of the software. Add to this the ever-increasing deployment of IT with larger numbers of ISs in use. There could be data about the same persons (e.g. patients), facilities and locations (e.g. healthcare facilities), things (e.g. cytostatic materials), concepts (e.g. disease diagnoses) and other data elements in dozens, hundreds or even in thousands of data storages [4] with (partly) unknown interconnections. As a summary, data definitions are bound late and are in a constant uncontrollable flux. Similarly, the IS technical, informational and socio-contextual characteristics of data differ between data storages.

The transfer to the above-depicted open systems environments led us to conclude, that data federation should be done on attribute level, not on entity and/or data model levels as advocated e.g. by [6, 7]). In order to federate two or more data sets it is necessary to have at least one connecting, that is shared, data attribute. In closed systems environments, the data attribute lists of IS-specific data models are known to the user organization and bear the informational and the socio-contextual meta data of attributes, even if such meta data is usually un-documented. As the IS technical meta data known to an organization, it is sufficient to know shared attributes' format, length and whether those attributes are mandatory and primary or secondary search keys. After that the values of shared attributes could be matched, cleansed and merged. This is done to create so-called golden records [1, 6, 15]. In open systems environments data federation requires more IS technical, informational and social meta data. For example, how does an organization federate data, if a shared attribute is mandatory search key in one data storage but not in another data storage? As an example, patient identification could be the mandatory primary key in a patient register IS, but not in a laboratory IS. A laboratory IS may register events on the basis of mandatory sample identifications instead of patient identifications. In open systems environments, informational and socio-contextual meta data are needed to understand, how shared attributes have been created and what are the meanings of the attributes in their use contexts, since the user organization does not have such information. In summary, the

conclusions that lead to development of the federative approach are as follows: First, identify attributes that make it possible to share - and to federate - two or more data storages. Second, define necessary IS technical, informational and socio-contextual meta data characteristics for each shared attribute and cross-map them. In our case, patient-id, cancer diagnosis id, TNM-classification (tumor nodes metastasis) and date were identified as the shared attributes for cancer data in roughly a dozen data storages.

### *2.2. Governance of data and data federation*

We built our artifact on the theoretical basis provided by the governance of data framework proposed by Dahlberg and Nokkala [5]. Their framework suggests from a governance perspective how to solve the challenges created by the digital data explosion [11] and the above-described transformation from closed to open systems environments. Dahlberg and Nokkala [5] notified that most organizations experience difficulties in determining what data they have and use, what data they could and should have and use, and who is and should be responsible for the various phases of data lifecycle and for ensuring the quality of data [e.g. 3, 16]. Due to these difficulties organizations find it hard to determine what data they should manage and govern as well as how to tie together the management and governance of data [3, 16]. Furthermore, the imprecisions of data governance influence negatively data federation, as people wanting to federate data do not know what data is available and accessible to them, and what is the quality and reliability of the data they have access to. The framework of Dahlberg and Nokkala [5] seeks the solution to the governance of data challenges from corporate and IT governance principles, most notably from the principles expressed in the ISO/IEC 38500 governacne of IT standards family [13]. The purpose of their framework is to help organizations to establish clear governance and management accountabilities for data used in the various activities of an organization and its stakeholders.

At a more concrete level, the framework of Dahlberg and Nokkala [5] proposes that an organization needs to establish clear governance and management accountabilities for and in the following issues:

- Data types, such as transaction, report, content, document, master, reference and meta data [2, 6]
- Data sources mapped to data types, such as IS business transactions, sensor, message, audio and video data sources
- Temporality and spatiality of data mapped to data types, such as historical, current, future; single spot and time-series data as well as location data
- Internal-external origin of data mapped to data types
- Structure of data mapped to data types such as structured, unstructured and multi-structured data
- Life cycle phase(s) of data mapped to data types, such as creation, usage and modification (CRUD) of data as well as storing and purging of data
- The meaning of data mapped to data types and the life-cycle phase of data. This is the ontology of data issue discussed in Section 2.3.

In the design of the cancer data artifact the issues listed above were used to support the definition of some informational and socio-contextual meta data properties of shared cancer data attributes. For example, we considered, in what kind of data types could the patient identification attribute appear, such as transactions, reports and master

data? What are the sources of a patient identification attribute, such as patient IS, radiology image IS? Is a patient identification internal or external attribute, that is, registered and created internally or retrieved from the Population Register Centre? Who is responsible for the content of a patient identification attribute, for example, persons with role X in patient registry, persons with role Y in the radiology unit? In the design of the cancer data artifact we used experiences from earlier comparable designs, such as the federation of dairy product data. Those designs were developed in commercial projects for real-life solutions. These desings are currently in everyday use.

### *2.3. Ontology of data and data federation*

One of the data governance propositions by Dahlberg and Nokkala [5] is that an organization needs to define its stance to the ontology of data. Why? Most of us are familiar with situations where various data storages provide fragmented, overlapping and even controversial data on the same topic. For professionals, such as doctors treating cancer patients and data/information analysts helping them, the question is, which data storage(s) should they trust and use - for example, in efforts to detect malignant breast cancer cases. In order to use data effectively and to obtain data compatibility, coherence and interoperability, users must be aware of data accessible to them from various data storages. If data is federated, they also need to understand the meanings of pre- and post-processed data. This is where the ontology of data enters the scene. In general ontologies describe the meaning of data in their use context(s): For what purpose is data created, used and stored as well as what does data mean in each of its use context during the life cycle of data. According to Wand and Weber [19, 20] the underlying ontological premise is that the data (of an IS) represent reality, and events that change reality. As the consequence of this ontologies define data quality requirements (for an IS [19,20]), which are then needed to track real-world situations and their changes, and to report situations and changes rightfully [19]. The key principle in the design (of a 'good' information system) is to strive for completeness. Completeness is achieved, when ontological constructs - "things", their properties and values - are mapped in the design constructs, no more, no less. Ontologies also explain what kind of deficiencies data and data representations may have [18].

We consider the work of Wand and Weber's [19, 20,21] seminal for determining the quality and possible deficiencies in the data (of an IS) [4]. On the other hand, we also consider that their propositions need to be extended into open systems environments. The federative approach attempts to do that. Both in closed and open systems environments, the multidimensionality of real world cannot be properly represented, because all dimensions change constantly. It is especially difficult to distinguish emerging genuinely different real world states. As a consequence different real world states could be mapped to same IS/data storage states (fields). For example, a medical IS/data storage might allow the entry of data only into a fixed amount of data fields. After increased knowledge about malignant cancers there could be the need to enter data into a larger set of data fields – which the medical IS/data storage does not support. In open systems environments, it even more likely that one real world state is mapped into several IS/data storages. For example, data about the same cancer diagnosis could be entered into patient, pathology, surgical, and other IS/data storages (without cross mappings). After that changed states are entered only to those IS/data storages that are directly related to specific state changes. As the result of this the state

description of IS/data storages will differ. One evident conclusion is that is necessary to know the meaning of data (shared attributes) in the data storages that are federated.

The significance of ontologies for distributed data grows in open systems environments, because the number of real-world contexts and possible data deficiencies also grow. This happens when an organization transforms from a well defined closed systems environment with a few IS to an open systems environments with lots of IS and data sources. Ontologies do not solve all data deficiency problems. Yet, the help to control the following three representational problems that could hamper ontological completeness [20], and which appear to happen constantly between data storages in open systems environments:

- Construct overload. This means that one design construct maps into two or more ontological constructs. For example, the construct "cancer" could represent a thing and a property of a thing in federated data storages.
- Construct redundancy. This means that two or more design constructs represent a single ontological construct. For example, the construct, 'cancer diagnosis' may split into several design constructs in federated data storages, such as the confirmation of a diagnosis, test results of a diagnosis, or a doctor's diagnosis. Thus the (domain specific) ontological construct and the design constructs are not the same.
- Construct excess. This means that a design construct does not map into any ontological construct. Non-functional unused definitions and properties of IS and data storages are typical examples. The data storages of ISs purchased from software vendors, such as patient IS or pathology IS, may have unused data fields designed by the software vendors. A software vendor designs the functionalities and the data model of an information system to meet the requirements of potential customers a generic customer.

Wand and Wang [18] describe situations where data becomes deficient from users' viewpoint due to the above-described deficiencies in the mapping of real-world system ontologies to information system constructs. With Figure 1 we extend their approach to multi-contextual open systems environments, where real world could be seen through almost unlimited number of data usage contexts. Possibilities for data deficiencies are multiplied from the situation that Wand and Wang described. Figure 1 illustrates that real world is seen through the lenses of different contexts (1...n). As explained earlier organizations typically have one or more separate IS for each context, such as surgical, laboratory, radiology and pathology contexts. In open systems environments, getting the ontological and the IS/data storage meanings of data right increase in importance, and is the key to avoid representational and observational problems.

Data federation has been especially topical in master data management (MDM) literature [15, pp. 179-180]. Master data are non-transactional data shared by several IS [e.g. 1], which from data federation perspective provide natural links to various data storages. The MDM concept was introduced some 15 years as a way to consolidate fragmented customer (e.g. patient), product (e.g. cancer) and other master data (e.g. healthcare locations) [6]. The first efforts just brought data storages together, and were unable to produce much progress. So-called golden record approach emerged then as the solution to the problem of inconsistent and fragmented data storages brought together. During the recent years the golden record approach has dominated MDM thinking [see e.g. 1, 6, 7]. The data management body of knowledge (DMBOK) method [6] describes "*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*" [22]. DMBOK [6, p. 173] also states: *"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 warehouse / business intelligence environments."*

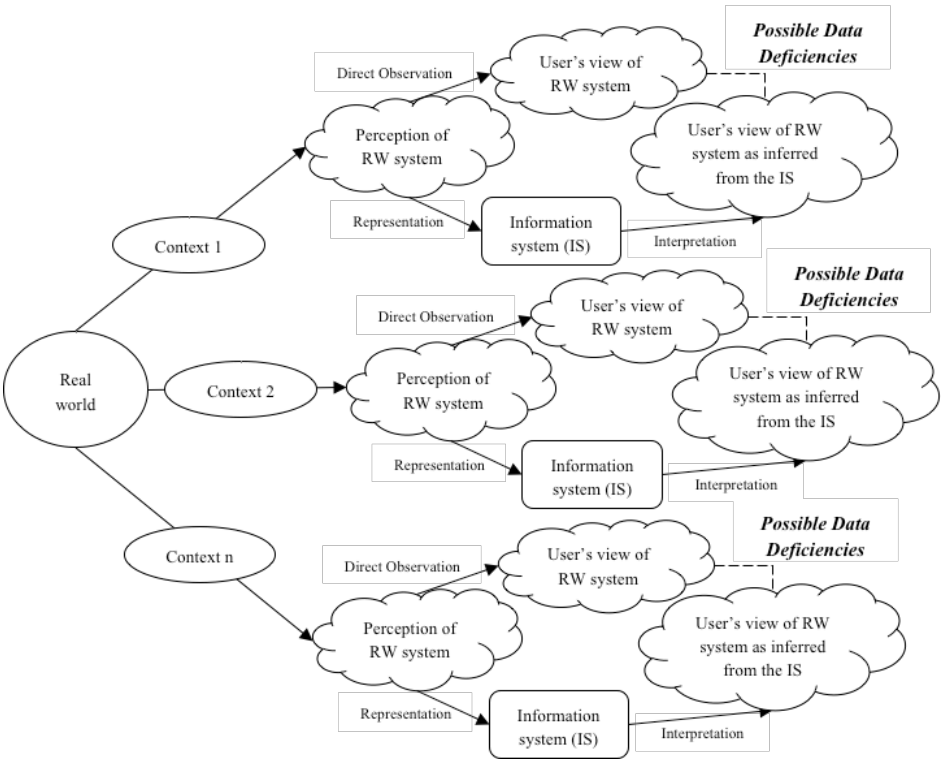


**Figure 1.** Possible data deficiencies from users' viewpoint in a multi-context environment (modified from Wand and Wang, 1996).

The golden record approach has been unable to deliver the promised solution, that is, organization-wide federation of customer, product, location and other master data, and through master data the federation of transactional and other types of data. Although improvements have been made, MDM solutions have remained fragmented instead of being organization-wide [4]. In our opinion the reason lies in the canonical ontological assumption, which is contrary to the contextual ontological assumption and other principles of Wand and Weber [20,21,22]. We discussed earlier the federative approach as a response to the challenges created by the transformation from closed to open systems environments. The federative approach was also developed as an attempt to overcome the problems of the golden record approach. The ontological assumption

of the federative approach is that data are contextually defined over the life cycle of data in the various use contexts of data within an organization's open systems environment. The proposition is that organization-wide (master) data federation is done through shared attributes and the meta data of the shared attributes. Shared attributes make data storages interoperable and meta data describes the IS technical, informational and socio-contextual meaning of shared attributes in each use context.

The federative approach acknowledges that the golden record approach could be used in a single homogenous use context. For example, that would be the case, if (breast) cancers would always be treated in the same way. Most purpose-specific data requirements are, however, seen to represent different use contexts of data. Data represents reality and the rightful vested interest of various use contexts, specific times of data usage, and the dynamic interplay between socially constructed data concepts [12]. In other words, purpose-specific data representations are not anomalies but correct representations of different real-world contexts and states. Therefore, the practical imperative is not to replace contextually different data, by matching, cleansing and merging, but to make related data interoperable by using shared attributes for that. This does not mean that an organization should have two different data values for the same record, for example the social-security identification or the name of a patient, unless there are contextual reasons to do so. Words identify/match, interpret and link by cross-mapping describe the federative approach.

Let's look at an example to compare the two approaches. The layman's rule of thumb for a normal human body temperature is 37 degrees Celsius or 98,6 degrees Fahrenheit. Yet, after sleep, stressful activity, medical surgery and other similar events clearly lower or higher body temperatures are normal. Contextual characteristics of measurements also impact results, such as is the person lying, sitting or standing. Moreover, the measurement device, the measurement method and their calibrations impact results. Rectal, vaginal, otic, oral and axillary measurements produce systematically different metrics [14]. Unless a doctor treating a (cancer) patient knows when, how, by whom and for what purposes measurements were taken and is able to determine the reliability of the data, (s)he may want to have additional measurements. Similarly, a new improved measurement method and its results do not mean that prior values should be replaced with the latest most accurate master data values as the golden record approach suggests.

In the present article, we point out some other earlier work we have used to develop the federative approach such as [10, 17] but do not discuss them in detail. We end Section 2 by answering the question, what is the significance of domain specific ontologies, for example medical vocabularies and code databases? In our opinion, they are highly useful in condensing and documenting knowledge. Still, a domain specific ontological construct and its (classified) value(s), such as cancer diagnosis codes, have different meanings depending on their use contexts. To federate data from dissimilar data storages it is necessary to describe three sets of meta data properties, e.g., for the cancer diagnosis code. IS technical meta data describes the field format, length and other technical properties of the cancer diagnosis code in each federated data storage. Informational meta data depicts how and who create and process the cancer diagnosis code in each federated data storage during the life cycle of cancer diagnosis code data. Socio-contextual meta data is used to define the meaning(s) of cancer diagnosis code in federated data storages

## 3. Methodology and the Case

With our artifact we support the federation of breast cancer data from data storages available to a university hospital in order to detect malignant breast cancer cases. That constitutes the case of our research. The university hospital has access to enormous amounts of relevant data both internal and external due to its role in the healthcare system in the country of the present research. A university hospital provides special healthcare services to the citizens of healthcare districts cooperating with the university hospital and sometimes also to other patients. Numerous professionals and software vendors have participated and participate to the development and operating of IS and data storages used by the hospital in general and by its (breast) cancer specialists in particular. Yet, the detection of malignant (breast) cancer cases is currently largely manual and based on the expertise of the professionals. The reason is that almost data characteristics in relevant data storages differ for reasons described earlier in this article.

Our ongoing research started in January 2016 and is executed in co-operation with the data/information specialists of the hospital. The development of the artifact and data collection is organized through workshops. Prior a workshop, the latest version of the artifact is prepared for its presentation at the workshop. The researchers and the data/information specialists crafted the first version of the artifact. Then in a workshop, researchers and data/information specialists interview a specific group of (breast cancer) specialists at a time, such as pathology specialists and IS-support personnel having responsibility for pathology ISs. We modify the design artifact after the workshop by using insight collected. The ability of the artifact to support data federation is evaluated lightly interim after each workshop and will be evaluated more thoroughly after the last workshop. If the data/information specialists and the medical CIO of the university hospital will consider the artifact and the federative approach useful, we will give the artifact and our approach to the university hospital to be used as generic tools in the federation of (medical) data.

We applied the governance of data framework proposed by Dahlberg and Nokkala [5] and the federative approach to the management and governance of data to craft (each version) of the artifact. The steps in the crafting of the artifact have been and are as follows:

- Step 1. Identify the most relevant ISs/modules and data storages for data federation. Identify groups of specific specialist that need to be interviewed about how data in those ISs/modules is understood and used.
- Step 2. Identify shared attributes that are needed to make data interoperable between the identified ISs/modules and data storages.
- Step 3. Describe IS technical, informational and socio-technical meta data for each shared attribute. .

The steps are iterative. Thus it is possible to complement – both add and reduce – ISs/modules and data storages, shared attributes and their meta data characteristics. For example we identified first three shared attributes and added the fourth later. Similarly, we had initially 30 to 40 candidates for the meta data characteristics of each shared attribute but were able to reduce their number later.

In the collection of empirical data about the case we follow the guidelines of Yin [23, 24] and Eisenhardt [8] for case studies and for the building of research constructs from case studies. As Eisenhardt [8] and Yin [24] mention, case studies can combine different data collection methods, such as interviews, observation and archival material. At the writing of this article it is not possible to determine, which of the six data

collection methods in the Yin's container [23] we will be able to use prior the project will be complete. We have a written case protocol [24] according to which we will use all other data collection methods with the exception of direct observing. Although we use the word artifact and follow a "standard" design science methodology to the extent possible, we regard our research more a single case research than a design science research. The reason is that the federative approach and comparable artifacts to the artifact developed in this case have been used earlier to federate master and social media site data in large commercial projects. As the consequence of this, the artifact designed in the present is not truly novel wheras the research context is new.

## 4. The Artifact as a tool to federate data and other results

The artifact - our tool to federate data – consists of two matrixes shown as Table 1 and Table 2. The data/information specialists of the university hospital found the artifact easy to understand and use. Although they are extremely busy and loaded with work, they have always found time for us. According to our prior experience the intellectual difficulty lies in understanding the ontological stance of the federative approach and the artifact. Most people agree intuitively with the statement that data is contextually defined but are accustomed to use the canonical data models of information systems. Some of them fail to make the intellectual step needed concerning the ontology of data. That is needed to federate data in open systems environments from incompatible data storages. That is why discussed the theoretical background of our research in detail as compared to methodology and results. The artifact makes no sense unless the ontological stance of the federated approach is understood and recognized.

The matrix shown in Table 1 was crafted for and during the iterative second steps of the artifact design. Second steps focused on the identification of shared attributes. Table 1 is shown in a generic format to prevent the identification of the university hospital's information systems. We crafted the matrix by placing information systems/modules and data storages to the matrix as the columns of the matrix. We placed shared attributes to matrix as the rows of the matrix. Please, note that we have skipped the laborious task, where we looked at the attribute lists of federated data storages in order to identify shared attributes. The matrix of Table 1 shows the outcome of step 2 and could be used to check once more that the shared attributes really exist in all federated data storages.

**Table 1.** Data federation artifact – identification of shared attributes

| | Patient IS | Laboratory IS | Surgical IS | Radiotherapy IS | Pathology IS | Information system N |
|---|---|---|---|---|---|---|
| | | | | | | |
| Social security identification | | | | | | |
| (Cancer) diagnosis code | | | | | | |
| Tumor node metastasis (TNM) code | | | | | | |
| Date of event | | | | | | |

The identification of shared interoperable attributes proved an easy task for the data/information specialists of the university hospital and made sense to cancer specialist we have so far interviewed in our workshops. The matrix compiles shared attributes from all ISs/modules and data storages into one table. On the basis of our experiences we believe that the best way to craft the matrix is to add at time one IS/module and data storage.

The matrix shown in Table 2 was crafted for and during the iterative third steps of the artifact design. Third steps focused on the definition of meta data characteristics for each shared attribute. Also Table 2 is shown in a generic format to prevent the identification of the university hospital's information systems and sensitive data.

**Table 2.** Data federation artifact – definition of meta data characteristics

| **Patient Identification** | Patient IS | Laboratory IS | Surgical IS | Radiotherapy IS | Pathology IS | Information system N |
|---|---|---|---|---|---|---|
| | | | | | | |
| **IS technical meta data** | | | | | | |
| Field length | description | description | description | description | description | description |
| Other characteristics | | | | | | |
| **Informational meta data** | | | | | | |
| Initial entry | description | description | description | description | description | description |
| Other characteristics | | | | | | |
| **Socio-contextual meta data** | | | | | | |
| Definition | description | description | description | description | description | description |
| Other characteristics | | | | | | |

The content in the cells of the matrix shown as Table 2 were defined by answering to the following questions:

- What kind of IS technical properties does a shared attribute have (format, length, hierarchy, granularity, mandatory, search key, …)
- What kind of informational properties does a shared attribute have?
  - Is the data type of the shared attribute, such as transaction, report, document, content, master data, reference data or meta data?
  - What is the source of the shared attribute: business transaction system, sensor device, control device, spatial device, temporal device, social media device, or other?
  - Is the shared attribute structured, unstructured or multi-structured?
  - Is the origin of the shared attribute an internal or an external data source? If the source is external, how is the organization allowed to process and use the shared attribute and the related data storage?
  - Who enters and modifies the shared attribute during its life cycle
- What kind of socio-contextual properties does a shared attribute have?
  - What does the shared attribute mean in each use contexts during the life cycle of the attribute?
  - For what purposes is the shared attribute used created and what does it mean at the time of creation?

    - For what purposes is the shared attribute used and what does it mean when used?
    - Why is the shared attribute stored and what does it mean when stored?
    - What other life-cycle stages does the shared attribute have and what is the meaning of the attribute in each stage?
- Governance of data - Who is responsible for a shared attribute?
    - Who are responsible for each of the IS technical, informational and socio-contextual meta data characteristic of the shared attribute?
    - Who are responsible for the data quality of the shared attribute?
    - How are the availability and the access rights of data ensured for the shared attribute?

What distinguishes out our approach from many canonical data integration endeavors is that we do not attempt to collect all the data into single harmonized data storage (vault), which is then used for reporting. Instead of that we let original data reside where they are. Technically data federations are conduced with the help of meta data repositories. A repository cross maps federated data storages by using the meta data of shared attributes to do that. The meta data repository is s data storage for federation rules, meanings of attributes and their meta data, description of data formats, and definitions of cross mappings. Meta data descriptions are created, modified and used only when a data federation need is recognized. New federation rules can be added whenever needed, e.g. for a new reporting need. The idea is to avoid big bangs and to proceed at the pace of learning.

## 5. Discussion and conclusions

In the first section of this article we asked two research questions. In the theoretical background section we argued and showed that user organizations have lost control over the data models they use and partly also over their data. Data is increasingly external and provided as a service, with unknown data models, APIs and/or adapters. Yet, the ability to manage and federate data is becoming increasingly important for organizations to reap the benefits of digital data. To manage and federate data effectively it is necessary to know why the federated data sets have been created, for what purposes they are used, and why data is stored. We compared two ontological approaches to data federation. We also discussed them theoretically by reflecting the phenomenon against existing literature on ontologies and governance of data. We discovered that contextual meta data is needed in order to execute data federations in addition to IS technical and informational meta data. This is our answer to the first research question.

Data federation starts from understanding the contextual meta data. The avoidance of data deficiencies related to the human perception of real word states, the representations of the real world in data/ISs and their combined effects need to be also considered when data is federated. We designed an artifact that operationalizes these considerations into questions ranging from understanding the reason of data creation to agreeing the governance accountabilities for data. This is our answer to the second research question.

Our article has the typical limitations of largely conceptual papers. At the moment, after approximately five years of research [3,4,5] we have limited amount of empirical

publicly available data to support the federative approach discussed in this article. The same is true for artifact we designed. Yet, since the topic of the article is both theoretically and practically important, we feel that these limitations are understandable. We have also done our best to avoid systematic errors in our argumentation.

The most important scientific contribution of our research is the extension of the work by Wang, Wand and Weber [18, 19, 20, 21], and other works cited in the theoretical background section, to open systems environment. We also brought the governance of data framework by Dahlberg and Nokkala [5] to concrete level. Practically our research contributes by showing steps that are needed to design tools for data federation between data storages that appear incompatible.

To (data management) researchers our advice is to consider the ontological nature of digital data. This issue continues to grow in importance fueled by data explosion and big data to name but a few. For practitioners our advice is to understand what data they have to perform their diverse tasks. Ability to link tasks and multi-sided, yet reliable data is needed in the case context of our research and most contexts of digital societies.

## References


[1] A. Berson and L.Dubov, *Master Data Management and Customer Data Integration for a Global Enterprise.* McGraw-Hill, New York, 2007.

[2] A. Cleven and F. Wortmann, Uncovering four strategies to approach master data management, *System Sciences (HICSS), 2010 43rd Hawaii International Conference*, IEEE, 2010.

[3] T. Dahlberg, Master Data Management "Best Practices" Benchmarking Study 2010 – Publicly Available Report. *Aalto University School of Business (Helsinki School of Economics*, 2010. Available through Researchgate.com, DOI: 10.13140/2.1.4201.9849

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

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

[6] *DAMA, The DAMA Guide to the Data Management Body of Knowledge DAMA-DMBOK Guide*. Technics Publications, LLC: Bradley Beach, NJ, 2009.

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

[8] K.M. Eisenhardt, Building Theories from Case Study Research. *Academy of Management Review*, **14**:4 (1989), 532-550.

[9] Finnish Cancer Registry (http://www.cancer.fi/syoparekisteri/en/?x56215626=112197488, retrieved 9.1.2016

[10] D. Heimbigner and D. McLeod D. A Federated Architecture for Information Management. *ACM Transactions on Office Information Systems,* **3**:3(1985), 253-278.

[11] M. Hilbert and P. Lopez, The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, **332**, no. 6025 (2011), 60-65.

[12] J. Iivari, R. Hirschheim and H.K Klein, A Paradigmatic Analysis Contrasting Information Systems Development Approaches and Methodologies, *Information Systems Research*, **9**:2 (1998), 164-193.

[13] *ISO/IEC. ISO/IEC 38500:2008 Corporate Governance of Information Technology (International Standard).* International Organization for Standardization and the International Electrotechnical Commission, 2008. Source: http://www.iso.org

[14] G. Kelly, Body Temperature Variabilty (Part 1): 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 (2006). 278-293.

[15] D. Loshin, D, *Master Data Management*. Morgan Kaufmann (2010).

[16] B. Otto, Quality and Value of the Data Resource in Large Enterprises. *Information Systems Management,* **32** (2015), 234–251.

[17] A.P. Sheth and J.A. Larson, Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases, *ACM Computing Surveys*, **22**:3 (1990), 183-236.
[18] Y. Wand and R.Y. Wang, Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*, **39**:11 (1996), 86-95.
[19] Y. Wand and R. Weber, An Ontological Model of an Information System", *IEEE transactions on Software Engineering,* **16**:1 (1990) 1282-1292.
[20] Y. Wand and R. Weber, On the Ontological Expressiveness of Information Systems Analysis and Design Grammars, *Information Systems Journal,* **3**:4 (1993), 217-237.
[21] Y. Wand and R. Weber, Research Commentary: Information Systems and Conceptual Modeling—a Research Agenda, *Information Systems Research,* **13**:4 (2002), 363-376.
[22] *Whatis - WhatIS Glossary*, Available Online [online], [downloaded 14 December 2015]. Available from Internet: http://whatis.techtarget.com/definition/golden-record
[23] R.K. Yin, Discovering the Future of the Case Study Method in Evaluation Research. *Evaluation Practice*, **15**:3 (1994), 283-290.
[24] R.K Yin, *Case Study Research, Design and Method*. 4th edition. Thousand Oaks, CA: SAGE Publications, 1999.