TRANSPARENT DATA SUPPLY FOR OPEN INFORMATION PRODUCTION PROCESSES

Complete Research

Laine, Sami, Aalto University, Espoo, Finland, sami.k.laine@aalto.fi

Lee, Carol, Northeastern University, Boston, Massachusetts, USA, car.lee@neu.edu

Nieminen, Marko, Aalto University, Espoo, Finland, marko.nieminen@aalto.fi

Abstract

Open Data and Open APIs have been recognized as valuable approaches for society and business. The validity of data driven decision making can be questioned due to inaccurate data and lack of sufficient provenance knowledge. We explored healthcare data entry situations in administrative patient encounter processes. A wide variety of empirical data was analysed by using frameworks from three disciplines: Human Computer-Interaction, Data and Information Quality, and Software engineering. The analyses revealed ambiguity in timestamps that cannot be recognized from a single perspective. More importantly, they cannot be recognized from the limited perspectives of Open Data or Open APIs that focus on the data layer and data recorded to databases. Unless identified, contextual variations are made visible with additional provenance metadata, they will endanger the validity of data and data driven conclusions. In the future, Open Data and Open APIs should be developed towards Open Information by opening current black-boxes with additional provenance metadata. We also developed general requirements for Transparent Data Supply that would solve several current data quality problems. Information production processes capable of fulfilling these requirements could help secondary users to assess the fitness of information assets for alternative purposes.

Keywords: Case Study, Data Quality, Open API, Open Data, Data Lineage, Data Provenance, Information Production Process, Total Data Quality Management, Transparent Data Supply.

1 Introduction

It has been argued that analytical competitors will use data-driven decision making to outperform their rivals (Davenport and Harris 2007). Senior management has found information systems to be more important in a more competitive and dynamic environment (Booth and Philip 2005). Information is becoming more known as a strategic asset and a source of competitive advantage.

New trends, such as Open Data (Dietrich et al. 2012) and Open APIs (Holley et al. 2014), aim to improve access to these strategic assets. However, information can have subtle characteristics that are not completely captured by explicit and abstract documentation about data sets and technical interfaces. These implicit characteristics might become recognizable only after information is used for a particular and/or unexpected purpose. As described below, these hidden details should be made visible for all participants in further development or consumption of information assets. With this information, informed participants can better understand and reflect on issues that are not currently explicitly documented at original data sources or derived information products.

In practice, the entire information production process should be traceable. The problem of traceability is referred to as ‘provenance’ or ‘lineage’, especially in information and computer science terms. Although information about provenance is very central to decision making, it has been found that end users are least satisfied with lineage metadata in four main metadata categories: definitional, data qual-

ity, navigational, and lineage (Foshay et al. 2007). There is clearly a need to improve the support for traceability and to provide more metadata about contextual factors affecting the quality of information.

Current Open Data and Open API ideologies, mostly associated with data products or technical software, have a very limited part in more elaborate socio-technical information production processes. The limited focus becomes a significant problem because Open Data or Open APIs can be used to block the visibility to previous phases across information production processes.

In this article, we aim to explain why Open Data approaches should pay more attention to the complete Information Production Process as defined by Total Data Quality Management (Wang 1998; Wang et al. 1998). We also suggest requirements to improve the quality and transparency of Information Production Processes.

1.1 Research Background

Originally, the research project was aimed to seek ways to improve quality of data and decision making in secondary uses, such as hospital administration, service development, policy making and clinical research. The project was focused on exploring data quality issues in Finnish hospital productivity benchmarking as the healthcare practitioners currently did not use published data marts (i.e. Open Data) about benchmarking results due to data quality problems (Linna and Häkkinen 2007).

We used qualitative research methods to explore, analyse and explain consequences of healthcare data quality problems. One of our empirical findings was the ambiguity of timestamps in administrative healthcare data. They were illustrative examples of data quality problems pervading current Open Data and API approaches. Therefore, this article focuses on administrative timestamp ambiguities.

1.2 Research Questions

In this article, our first aim is to illustrate current hidden data quality problems that can be embedded in technically sound and seemingly identical data sets:

RQ 1: *How exactly are timestamps created in administrative patient encounter processes?*

Then, we further analyse these findings and identify five data quality problem themes that should be resolved to improve the quality of data and data-driven decision making:

RQ 2: *What kind of timestamp accuracy problems exist in administrative patient encounter processes?*

Finally, we discuss what our previous findings means for development and procurement of information systems:

RQ 3: *What should be done to solve the similar data accuracy problems in information assets?*

2 Openness in Information Management

From the perspective of Open Data, the key features of Openness are availability, access, reuse, redistribution, and universal participation, which result in free access to data sets or technical APIs (Dietrich et al. 2012). These data products or technical interface services can be used to disseminate almost any information assets, such as administrative, cultural, scientific, financial, or environmental information. They are also opened for many reasons, such as increasing transparency or adding social and commercial value.

2.1 Data Quality Management

Data and information quality is an established international multidisciplinary research domain. The evolution of data quality research during the last decades has been reviewed in relation to its topics and methods (Madnick et al. 2009).

In the early 1980s, researchers at Massachusetts Institute of Technology (MIT) founded the Total Data Quality Management (TDQM) framework (Wang 1998; Wang et al. 1998). The TDQM defines continuous data quality improvements with the consecutive cycles of Define, Measure, Analyze, and Improve. These tasks are applied across the entire information production process (IPP) consisting of three phases: data supply, data manufacturing and data consumption.

According to TDQM, the same data set can be source data or information product. For example, the citizen address data can be listed as an Information Product for Finnish Population Register Centre. At the same time, it is a data source for companies and public organizations. Therefore, any Open Data product or Open API service should be considered simultaneously as a potential data source as well as an information product in a wider information environment consisting of numerous information production processes.

According to TDQM, information assets based on Open Data products and Open API services should be managed by well-defined and quality controlled information production processes. A key lesson of TDQM is that information manufacturers and suppliers need to expand their knowledge about how and why the consumers use information (Wang, 1998). Likewise, information consumers need to understand how information is produced and maintained to be able to assess the fit of Information Products for their particular purposes. The network of suppliers, manufacturers and consumers should work together and be managed through cross organizational information management rather than be limited by organizational boundaries.

2.2 Challenges for Openness in Information Management

The results of Finnish hospital productivity benchmarking are not used by practitioners due to data quality problems (Linna and Häkkinen 2007), such as erroneous data, biased algorithms and obscurity between phases across information production processes (Laine and Niemi 2013). Similarly, the conclusions of administrative database research have been questioned because these studies rarely validated their data (van Walraven et al. 2011). In two out of five studies, people with the diagnostic code did not actually have the disease. In addition, the quality and usefulness of published health service waiting times are difficult to interpret due to subtle variations in organizational incentives, patient characteristics, treatments, hospital processes and work practices (Stoop et al. 2005). In practice, open access to data and statistics does not seem to matter since secondary users do not trust them due to hidden inaccuracies and lack of sufficient provenance information.

2.2.1 The Challenge of Provenance

Currently, Open Data and API approaches often limit themselves to a small phenomenon occurring at the data manufacturing phase. A significant problem arises when Open Data data sets or application interfaces are used to block the visibility to any previous phase across wider information production processes. Similar to administrative database research, Open Data sets are often provided without explicit information about how exactly values in the data set were created at physical and social realities in the first place (van Walraven et al. 2011). Information manufacturers and consumers are unable to evaluate and verify the actual quality of data and its true fit for alternative purposes.

The lack of such provenance information is a common problem in secondary data usage, although it is considered to be a central issue for decision makers. For example, information manufacturers and consumers agree on the importance of definitional, data quality, navigational and provenance metadata (Foshay et al. 2007). However, they also noticed that manufacturers and consumers do not agree on the priority and content of metadata. Consumers interpreting the data are the least satisfied with provenance metadata. Also, the limited data provenance documentation is often provided by IT professionals, whom are focused on traceability across data manufacturing phases. Provenance metadata describes source data properties, data transformations, aggregations or repeat official business rules. These issues do not cover many of the significant details occurring at the data supply phase. For ex-

ample, they do not describe how exactly reality was observed to determine the correct data value, how each data value was actually handled before being stored to databases and what kind of measurement errors and biases might be embedded to data sets. There is clearly a need to provide more metadata about contextual factors affecting the quality of information.

2.2.2 The Challenge of Semantic Heterogeneity

Service Oriented Architecture (SOA) principles, such as encapsulation and abstraction, promote reusability and flexibility by hiding internal details of technical implementation and contextual variations of situated semantics (Bieberstein 2005). Applications can insert, update and delete data instances in standardized data sets in various ways. These technical variations in applications and user interfaces can become hidden semantic heterogeneity and latent bias in seemingly homogenous data sets unless they are explicitly captured, stored and delivered for future uses.

In 2012, Gartner updated its definition to "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" (Beyer and Laney 2012). Therefore, Big Data refers to increasing technical and semantic heterogeneity (i.e. variety) in data sources.

In the future, Context Aware-Computing and adaptive user interfaces will become more common and versatile (Moran and Dourish 2001). Subtle adaptations in user interfaces can alter details that distort semantics in relation to other configurations or user interfaces. For example, variations in the order of tasks, types of input mechanisms, default values embedded to selection lists, or wording in labels can influence accuracy of information by altering error profiles or semantic meanings. Many current technological trends, such as above, increase data quality problems like semantic heterogeneity. These contextual variations should be made visible for secondary users to allow the evaluation of information products for their particular purposes.

2.2.3 The Challenge of Information Environments

Currently, management methodologies like TDQM face more problems regarding data standardization. Information is a commodity that can come from outside of its own organization and quality control. Previously, organizations could control their own governance and management within their environment, and stakeholders may not have been able to collaborate freely. Such controlled environments could be distinguished from Open Information Environments (OIEs): users have access to sources they may have no control over; new sources of data may emerge; applications of data might change radically over time; and new uses of data might emerge (Parsons and Wand 2014). In addition, as the development of data and applications are outsourced, information services are changed rapidly and hidden from all stakeholders' views. The semantic diversity and provenance challenges increases along with the openness in information environments.

3 Methods and Data

3.1 Case Organizations

The Finnish healthcare system is based on public services. Primary healthcare is provided by health centres in municipalities. Specialized healthcare is provided mainly in hospitals governed by hospital districts. Hospital districts are organized into five groups around university hospitals responsible for providing highly specialized services. In addition, small sectors of both private and occupational healthcare exist.

Our case organizations are two Finnish hospital districts and their respective university hospitals. They are similar-sized university hospitals providing similar specialized healthcare services for almost 500,000 residents and over 20 surrounding municipalities. Their demographical population coverage

and service portfolio are roughly the same and are also governed by the same national legislation and policies. For example, they use the same internationally standardized codes, such as the ICD10, for diagnosis coding and NCSP for surgical procedures. They also use the same primary electronic patient record system, including the identical proprietary Operational Data Source database that is used for extracting raw operational data for data warehousing and reporting purposes.

3.2 Research Methodology

Explanatory case studies, such as our research project, provide accurate descriptions of the facts, considerations of alternative explanations, and a conclusion based on credible explanations of the facts (Harder 2010). We aim to describe our empirical findings about timestamps, analyse the alternative meanings and explain the kinds of accuracy problems that exist in identified timestamps.

As a qualitative case study, we increased the credibility of our findings with five of the six potential types of triangulation: data source, data type, methodology, theory, and analysis (Mills et al. 2010). We acquired informal interpretations from several professions, official management rules from administrative documents, technical details from system documents and physical first-hand evidence from user interface walkthroughs. Complementary data sources, data types and research methods provided us in-depth empirical data about data entry situations across clinical pathways of neurology patients in different hospital units.

We also used knowledge from three disciplines to analyse contextual variations hidden in timestamps. Combining these three theoretical perspectives and their distinctive analytical units, we reveal and explain variations that are not evident from a single theoretical perspective. Drawing upon theories from human computer-interaction research, we analysed timestamps in relation to Context of Use (CoU): users, tasks, tools and environments as defined by ISO 9241-11 standard (ISO 1998). Alternative contextual meanings were analysed further according to three data supplier roles, defined by the TDQM framework (Shankaranarayanan 2000). Software Engineering provided the third framework for analysing contextual variations in timestamps. Software systems consist of three integration layers: data, application and user interfaces (Fowler 2003).

3.3 Data Collection

3.3.1 Preliminary interviews

Data entry collaboration was found to be very complicated between varying hospital units and professions. Individual interviewees often remarked that they do not know how exactly certain individual data instance is entered to the electronic patient record system since it is completed by other professions or units. Interviewing individuals one by one ended up being an inefficient research method. Therefore, exploratory focus groups were considered better research methods to explore data entry situations across clinical pathways. Physicians, nurses and secretaries were brought together to discuss subtle nuances and practical problems related to each data entry situation. In the preliminary phase, we also collected a wide variety of administrative and technical documents from hospital district intranets.

3.3.2 Preliminary analyses and planning focus groups

Preliminary empirical findings from interviews and various data management documents were used to develop ‘administrative process charts’. These charts were printed and set in the middle of the table at group meetings. In practice, they were swim lane flowcharts consisting of administrative process steps and their respective user interface screen names. These flowcharts were used to guide the flowing discussion and remind groups of the important topics that should be discussed in step-by-step detail.

3.3.3 Focus groups

A focus group is defined as a moderated discussion among six to twelve people who discuss a topic of interest under a guidance of moderator (Tremblay et al. 2010). The purpose of group discussions and user interface walkthroughs was to explore and document Context of Use across clinical pathways. We specifically explored data quality controls and its related workarounds. We also paid particular attention to semantic data definitions and their situated interpretations.

During focus group sessions, researchers systematically explored and documented each individual's data entry process phase-by-phase and interaction with each screen of the user interface guided by the developed flowchart templates. The majority of attention was paid to diagnosis coding, but also to three other significant data element categories: specialty codes, procedure codes, and timestamps.

The focus groups consisted of the primary users of the software systems: one physician, two nurses and two secretaries from different units. One of the nurses did not attend at one hospital and one of the secretaries did not attend the meeting at another hospital. Group sessions were recorded and transcribed resulting in two sets of approximately 30-page documents.

The primary author moderated the sessions and two other researchers took notes. A local information management staff member, responsible for electronic patient record user training, operated the software systems and captured screenshots during the walk-throughs.

3.3.4 Data Analysis

Considering the aim of the research, to explain hidden variations potentially embedded in Open Data, we emphasized creative views to qualitative analyses in contrast to a procedural view (Coffey and Atkinson 1996). Creative perspective emphasizes strategies that explore significant relations and interpretations within the data rather than code, sort and categorize the data. Therefore, we deliberately looked for alternative theories capable of revealing inaccuracies that are currently hidden in information products.

First, we aimed to identify contextual and semantic variations rather than look for errors or their root causes. We identified different timestamp types and their occurrences in data sources. For example, we identified the registration, appointment and discharge timestamps. We considered alternative explanations for individual timestamp values and documented their potential contextual meanings to a framework based on Context of Use: users, tasks, tools and environment (ISO 1998).

Second, we sought for alternative frameworks that could be used to better illustrate the nature of the phenomena that we are interested in to explain. We identified two additional theoretical frameworks for further analyses about contextual meanings. In data quality research, TDQM has defined data supply to consist of data creation, data collection and data recording phases (Shankaranarayanan 2000).These data supply roles were chosen to highlight hidden semantic and pragmatic variations between original data creations and data sets entered into databases. Software system layers (i.e. data source, business logic and user interfaces) were chosen as analytical framework to highlight inaccuracies resulting from semantic mismatches in and between technical layers (Fowler 2003).

Third, we identified significant problems in holistic themes that were found to have negative effects on data quality. Individual data quality problems are often related to many problematic themes at the same time. Therefore, all previously identified timestamp occurrences were analyzed in relation to each problem's theme. These analyses were used to describe important factors related to each theme.

Finally, we derived generalized requirements for Transparent Data Supply. That is, because our research aims to provide practical contributions for practitioners.

4 Results

4.1 Administrative patient encounters

HL7 V3 Patient Encounter standard defines patient encounter as 'an interaction between a patient and one or more healthcare practitioners for the purpose of providing patient services or assessing the health status of the patient (HL7 2014). In this article, we focus on the two most common types out of seven unique types: inpatient encounter and outpatient encounter. The difference between the two is simple - an inpatient visitor is admitted to a hospital and gets assigned to a bed, while an outpatient visitor is just registered to an appointment for a short-time.

A simplified example of a healthcare process consists of the following stages: 1) patient arrives at the hospital, 2) patient signs into the hospital upon arrival, 3) patient receives appropriate treatments, 4) patient is discharged from the hospital, and 5) patient leaves the hospital. We limit our focus on these 5 stages, even though there are other administrative stages before and after patient encounters, such as the original hospital might have received a referral, entered the patient into a queue, scheduled an appointment and sent information to the patient.

At each stage, information in the electronic patient record systems is updated and HIS generates timestamps based on user's activities. Phases 1 (arrival) and 5 (departure) are actually unknown in reality. Hospitals are only aware of the administrative patient encounter process consisting of phases 2-4. Considering above, it is impossible to measure how fast patients receive services and treatments. Instead, the speed in which patients receive services can only be measured after they have been formally registered as patients waiting for services.

In this paper, we focus our analyses to a particular phase: patient informs hospital about arrival and the related timestamps. In this way, we use our in-depth data and analyses to show the contextual complexity of seemingly simple timestamp data elements.

4.2 Timestamps at administrative patient encounters

The administrative patient encounter process consists of arrival, registration, encounter, discharge and departure phases. For example, a patient arrives at the hospital for a scheduled appointment. In both hospital districts, the user interface screens are almost identical. The same electronic patient record system is used with only some slight differences in selection lists and basic configuration choices. Many of their selection lists are similar because they are governed by the same national legislation and use the same nationally standardized coding schemes. In the patient arrival phase, there are two interesting timestamps: registration and appointment timestamps.

4.2.1 Registration timestamps

The research question is simple:

- *How exactly is a registration timestamp value, such as "8:53", created in hospital processes?*

In the first hospital case, a patient can be registered by secretaries at reception desks or via automatic self-registration with barcode cards. In the other hospital case, patients are registered only by secretaries at reception desks. Therefore, timestamps can represent the moment of physical arrival to the hospital or the event of available administrative reception. In reality, these are not always exactly the same moment.

With detailed inspection of the timestamps, much more complexity is revealed. At hospitals, patients can be transferred from one unit to another. In these cases, the registration timestamp for a current encounter is actually derived from the discharge timestamp of the previous encounter, where one minute is added to that discharge time. Since hospitals districts are made up of networks of physical buildings

in different cities, administrative timestamps seems to show that patients had traveled from one unit to another in a minute.

The situation is semantically even more complex, since those previous discharge timestamps were entered manually into HIS from the discharge summary screen. In these situations, secretaries enters time manually and ambiguously writes using different meanings, such as patient will leave a unit later at this time, patient will be picked-up by someone at this time, or patient will be leaving just now. The original discharge time is ambiguous and is automatically supplied to a registration timestamp of the next encounter.

In addition, we discovered that a data entry policy alters timestamps even further. If a patient comes to the hospital in the middle of the night, the registration timestamp will be manually altered backwards to a previous date and time rather than supply the current date and time.

Technically, registration timestamp data can be created in many different ways. Sometimes timestamps are created automatically by inserting an insurance card to a machine. Some timestamps are created by pressing a single confirmation button or 'n=now'-key at the user interface screen. Also, timestamps can sometimes be altered manually in adjustable timestamp fields. For example, secretaries may adjust timestamps to represent the expected time of a patient pick-up.

USER	TASK	TOOL	ENVIRONMENT	MEANING
Patient	Self-registration	Barcode card	Current unit	"Arrival at location"
Secretary (current user)	Registration	EPR & key press	Current unit	"Available service at reception"
Secretary (current user)	Registration	EPR & manual adjustment	Current unit	"Midnight at previous day"
Secretary (at previous unit)	Discharge	EPR & manual adjustment	Previous unit	"will leave at this time"
Secretary (at previous unit)	Discharge	EPR & manual adjustment	Previous unit	"will be picked up at this time"
Secretary (at previous unit)	Discharge	EPR & key press	Previous unit	"is leaving unit now"

Table 1. We noticed that registration timestamps are ambiguous as seen in the "meaning" column. Timestamp values might have been created in completely different situations for many reasons and created by varying techniques.

4.2.2 Appointment timestamps

The research question is again simple:

- *How exactly is an appointment timestamp value, such as "8:53", created in hospital processes?*

Inpatient encounters do not have any relevant timestamp for meeting physicians or nurses. Inpatients are just registered (i.e. admitted) to wards. Of course, electronic patient record system does have timestamps built in, such as last modification timestamp, but they are not used by the primary users. Since inpatient appointment timestamps do not exist for primary users, it is impossible to know when exactly the patient met the physicians or nurses for the first time.

Outpatient encounters do have timestamps for appointments, but an appointment timestamp is actually picked up from the previous scheduling module and presented in the screen as a preset value. Since the hospital staff never changes the value of these appointment timestamps, the value presented in the registration screen is actually the planned event time rather than the actual event time. Also, these planned appointment times were originally entered from the scheduling user interface screen rather than the

appointments user interface screen. In this way, the appointment timestamp might have been entered by a different person that is registering the patient.

We also found that sometimes appointments do not have any timestamp data. The reason for this is that the documentation had been delayed, even though a patient received the appointment services a long time ago. In these cases, the timestamp field is left empty.

Another ambiguity is that multiple appointments can be held at the same time. A physician can have several planned virtual encounters, such as phone calls, at the same time. Physicians can change the order or length of these phone calls subjectively according to their situations. There is no way to distinguish and measure these events individually. In these cases, it appears that a physician is providing several appointments at the same time.

Technically, appointment timestamp values are created by selecting graphically available service slots from a scheduling module calendar screen. Therefore, appointments appear to be at regularly intervals, such as 8:00, 8:30, and 9:00 and so on. Patient appointments seem to match their plans very well because appointment timestamps are automatically picked up from the scheduling module. They are not updated to reflect the reality and match actual event timings.

ACTOR	TASK	TOOL	ENVIRONMENT	MEANING
Secretary (at previous time)	Scheduling	EPR & free calendar slot	Current unit	"planned appointment time"
Secretary (at previous time)	Scheduling	EPR & free calendar slot	Current unit	"multiple indistinguishable event plans at the same time"
Secretary (at previous time)	Discharge	EPR & no update	Current unit	"Delayed documentation with no timestamp value"

Table 1. Appointment timestamps are ambiguous, as seen in the "meaning" column. Appointment timings are not created at the time of the appointment. They are not entered from the appointment registration user interface screens, but rather from the scheduling screens.

4.3 Contextual variations along administrative data supply phases

TDQM has defined data supply to consist of data creation, data collection and data recording phases (Shankaranarayanan 2000). Therefore, we analyzed timestamp meanings, such as arrival at location, in relation to data supply roles. Our analytical mappings illustrate how data is semantically transformed across creation, collection and recording.

The problem, at the data creation phase, is that semantic variations vanishes and becomes indistinguishable from each other after the data values has been recorded to the electronic patient record systems. The user interface or database of the electronic patient record system can represent the data to look identical to 'availability of reception', although the data value actually could be 'arrival at location' or 'will be discharged at this time' (table 2).

MEANING	Registration timestamps that look like "availability of reception" but are actually "arrival at location".		
SUPPLY PHASE	**CREATE**	**COLLECT**	**RECORD**
USER	Patient	EPR	Secretary
TASK	Self-registration	Data integration	Registration
TOOL	Barcode card	Registration Device	EPR
ENVIRONMENT	Current unit	Current unit	Current unit
NOTE	**MISSING!**	**MISSING!**	**"Open Data"**

Table 1. The problem is that Open Data can present data as it is entered to the system (record column). That is how data is stored in database tables or views and later presented for users on user interface screens. In reality, original data values are created at the hospital as presented in the "create" column. The contextual information is lost on the way when the original timestamp is collected by devices and stored at registration.

The problem lies in the accuracy of the data in the database to represent the reality. The actual differences and variations are not recognized because they vanish across the data creation and data collection phases. The lessons learned from Total Data Quality Management are: There is a need to identify the original data creator rather than rely on information entered at the time of the entered data. The original situation (create at table 2) should not be hidden by creating a black-box with open data or open API (record at table 2).

4.4 Technical obscurity across technical software layers

Software systems are based on three layers: the data source layer, the business logic layer, and the user interface layer (Fowler 2003). We analyzed the timestamp meanings in relation to software integration layers. Our analytical mappings illustrate how data can be semantically transformed across the data, application and user interface layers.

The problem at the original data creation user interface layer is that semantics vanish and become indistinguishable from each other after the data values has been recorded to the electronic patient record systems. The current user interface or database of the electronic patient record system can represent the data to look identical to 'availability of reception' although the data value actually could be 'will leave at this time' or 'arrival to location' (table 3).

MEANING	Registration timestamps that look like "available service at reception" but are actually "will leave at this time".		
SOFTWARE LAYER	**User Interface (Original)**	**Application**	**Database (Current)**
USER	Secretary (at previous unit)	EPR	Secretary (current)
TASK	Discharge	Enforce business rule	Registration
TOOL	EPR & manual adjustment	Software code	EPR & timestamp
ENVIRONMENT	Previous unit	Data center	Current unit
NOTE	**MISSING!**	**MISSING!**	**"Open Data"**

Table 1. The problem is that, in the database registration timestamp, data can look like it was entered as "availability of reception", and yet the original value might have been created at a different user interface screen rather than the one inserting it to the database. That is, because a codified business rule in the application presets a value to the current user interface screen that reinserts it to database.

Problems arise if someone believes that the data in the database represents the reality and does not recognize the actual differences and variations in the original application and user interface layers. The lessons learned from software engineering theories are: there is a need to identify how the data instances have moved across the technical layers: data, application and UI. In addition, potential variations in the original user interfaces and applications logic should not be hidden by seemingly standardized data sets or interfaces.

5 Discussion

5.1 Timestamp accuracy problems

Multiple data entry techniques

The same timestamp data element (e.g. registration or appointment moment) might be created and updated in various ways at alternative phases of the administrative patient encounter processes. The value can be entered by using a physical barcode card, keyboard, or mouse. It could be automatically filled by computerized business rules. In addition, the keyboard can be used in multiple ways, such as pressing the confirm-button to save an event, pressing the 'n'-key to fill an empty timestamp slot, adjusting timestamp values via graphical icons, or manually typing in the complete moment of time. Alternative devices and data input techniques lead to different cognitive situations and error profiles.

Obscure data creation situations

Originally, timestamp data values might have been created somewhere else in another context. The original user, task, tool and environment might be completely different than the one recording data to the database. Therefore, TDQM methodology defined the creator, collector and recorder roles as a part of the theoretical framework. The actual actors entering the timestamp data might have been the patient itself or the previous secretary who reserved the appointment a long time ago. The original task that created the registration timestamp value might have been a discharge task in the previous patient encounter rather than the current registration task. The data entry tool might have been another device, such as self-registration machine rather than the electronic patient record system. Furthermore, the original registration timestamp data entry environment might have been organized completely differently, as the social and physical unit may be in another building than the one responsible for the registration. In practice, these contextual variations make timestamps ambiguous and inaccurate.

Ambiguous and inconsistent definitions

Timestamps are defined and documented ambiguously. Self-registrations with barcode cards or reception desk registrations cannot be distinguished clearly. They are presented in the same user interface fields and displayed with same labels. Discharge time can be used ambiguously to describe for example "now", "pick-up time" or "planned departure". Ambiguous labels or definitions are also used inconsistently. For example, a single date timestamp simply labelled as "time", might be displayed separately as date and time and labelled "registration date" and "registration time". In the previous case, one has to know what "time" it really is and where it has been derived. Semantic consistency breaks down completely when registration timestamps are actually derived from completely different events, such as discharge. These semantic changes are not documented in user interfaces or database tables, but learned in practice. Also, appointment timestamps represent actual planned times rather than appointment events. Thus, the meaning has changed across.

Human errors and motives

Organizational processes are vulnerable to human errors and human motives. Although we considered these possibilities, we did not directly observe them in our case. All physicians, secretaries and nurses do not always follow these policies and described official workarounds perfectly. Timestamps can be forgotten or their timings can be manipulated. For example, it has been discovered that manual data updating problems lead to millions of dollars in yearly losses (Katz-Haas and Lee 2005). In addition, subtle variations in contexts lead to different error profiles. For example, barcode card registration errors differ fundamentally from manually adjusted timestamp values.

Human behavior patterns

Finally, various behavior patterns also exist, which can influence automatic timestamps: rushing tasks, delaying tasks, skipping tasks, manipulating tasks and so on. These behavior patterns cannot always be considered only as problems. They might be sometimes necessary and beneficial to solve more important problems in varying situations. In our case, the secretaries sometimes rushed documentation

and updated timestamps to the expected pick-up time of patients or skipped the whole timestamp field leaving it empty. Such behavior patterns are quite common in organizational work environments and result often from complex contextual circumstances (Koppel et al. 2008; Saleem et al. 2009; Yang et al. 2012). However, secondary uses, such as analytic calculations about task timings or process efficiency, become severely distorted.

5.2 Requirements for Transparent Data Supply

Our empirical findings and analyses have significant implications for the development and procurement of Open Data products. In this chapter, we derive requirements for Transparent Data Supply.

Quality Controls

At data entry situations, errors and environmental interferences should be minimized by effective quality control. That includes precise instructions, effective constraints and proper feedback mechanisms to guide the data entry towards the goal.

Precise Semantics

Data and metadata should be granular enough for distinguishing ambiguous meanings from one another, such as self-registrations and reception desk registrations. Semantic meanings should be precise across all contexts rather than generalized common concepts, which leaves room for ambiguities.

Documented Contexts

Data and metadata should store contextual variations, which are currently lost from data sets, based on the technical data layer and organizational data recording phase. In practice, contexts of data creation and collection should be captured and stored for later use in addition to the current data recording phase. Technically, the contextual properties of a user interface and application layers should be also stored in addition to the properties of the data layer.

Automatic Supply

Data and metadata should be collected automatically from the technical layer. That is, because manual data recording is subjective, inaccurate and laborious. The emphasis should be on automatic documentation of primary events rather than manual documentation for secondary purposes.

Traceable Contexts

Data and metadata should support traceability across documented contexts. Original data creation situations should be traceable from the recorded data links between all the three PP roles and three technical layers because semantics mismatches cannot be recognized from a single layer, but only in comparison to other layers or roles.

Openness

Data and metadata should be opened transparently for secondary users. Open Data or Open API should not be a black-box hiding the contextual details and variations in previous data supply roles or technical layers. Recorded data and technical data layer are not enough to understand individual data instances, but one should understand what has happened in the data creation and collection phases as well as in the application and user interface layers.

5.3 Implications for Openness

Currently, Open Data and API communities favor a simplistic and idealized view about benefits resulting from access to data (Janssen et al. 2012). At the same time, the validity of hospital productivity benchmarking (Linna and Häkkinen 2007), hospital waiting times (Stoop et al. 2005) and even peer-reviewed clinical research (van Walraven et al. 2011) have been questioned due to inaccuracies and contextual variations embedded in healthcare data sets and government statistics.

Traditionally, data and information quality management has emphasized quality controls and semantic standardization. They should also be emphasized in Open Data and Open APIs. However, data stand-

ards might not be precise enough or follow data recorders. Likewise, quality controls are just one of the many factors that can influence the actual quality of data. Our empirical analyses illustrated how even simple data elements might not be what they are in databases and data sets.

Currently, there is a wide agreement across disciplines that secondary users need to know how and why their data were created in the first place (Lee and Strong 2003; van Walraven and Austin 2011). This should also be acknowledged by Open Data and API communities that seem to neglect the importance of semantics and context in information products (Janssen et al. 2012). Open Data and Open APIs should always be delivered with provenance metadata about the actual data supply situations. Secondary users can then become aware of hidden inaccuracies and contextual variations that might lead to biased decisions and create obscure barriers against the use of published government data.

5.4 Generalizability

Our analyses were deliberately limited on administrative timestamp ambiguities, although our data covers other similar data elements that share similar problems. We believe that our current analyses about data creation situations could be replicated with other data elements. Preliminary inspections of other data indicate that our main findings are generalizable to other types of data elements as well as across industry sectors.

5.5 Future Research

There is a need to quantitatively evaluate the scope of identified inaccuracies embedded in Open Data products or Open Data interfaces. However, that requires additional provenance metadata that is often unavailable. Therefore, there is a need to research the quality and maturity of provenance support in Open Data products or Open Data interfaces. Most importantly, there is a need for Design Science Research that would build better provenance support for technical and managerial methods across Information Production Processes.

6 Conclusions

Currently, Open Data and Open APIs have been recognized as valuable approaches for society and business (Dietrich et al. 2012; Holley et al. 2014). However, our article draws attention to the current approaches of Open Data and Open API as a form of a black-box. We explored healthcare data entry practices in administrative patient encounter processes. Our data was analyzed using frameworks from three disciplines: Human Computer-Interaction, Data and Information Quality, and Software engineering. The analyses revealed ambiguity in timestamps that cannot be recognized from a single perspective. More importantly, they cannot be recognized from the limited perspectives of Open Data or Open APIs, that focuses on the data layer and data recorded in databases. Unless identified contextual variations are made visible with additional provenance metadata, they will endanger the validity of data and data driven conclusions.

In the future, Open Data should be developed towards Open Information by opening current black-boxes with additional provenance metadata. We explained why provenance should cover the original creation and collection phases as suggested by TDQM (Shankaranarayanan 2000; Wang et al. 1998; Wang 1998). In addition, extended provenance metadata should cover user interface and application characteristics. We also developed general requirements for Transparent Data Supply that would solve many current data quality problems. Information production processes capable of fulfilling these requirements could help secondary users to assess the fitness of Open Data for alternative purposes.

7 Acknowledgements

This research is funded by Tekes - the Finnish Funding Agency for Innovation.

8 References

Beyer, M. A., and D. Laney (2012). "The Importance of 'Big Data': A Definition", Gartner Inc.

Bieberstein, N. (2005). *Service-Oriented Architecture (SOA) Compass: Business Value, Planning, and Enterprise Roadmap*, IBM Press Pearson plc, Upper Saddle River, N.J.

Booth, M. E., and G. Philip. (2005). "Information Systems Management in Practice: An Empirical Study of UK Companies," *International Journal of Information Management* (25:4), pp. 287-302.

Coffey, A., and P. Atkinson. (1996). *Making Sense of Qualitative Data: Complementary Research Strategies*, Sage, Thousands Oaks, California.

Davenport, T. H., and J. G. Harris (2007). *Competing on Analytics: The New Science of Winning*, Harvard Business School Press, Boston, Massachusetts.

Dietrich, D., J. Gray, T. McNamara, A. Poikola, R. Pollock, J. Tait, and T. Zijlstra (2012). "Open Data Handbook Documentation", Open Knowledge Foundation, Cambridge, UK.

Foshay, N., A. Mukherjee, and A. Taylor (2007). "Does Data Warehouse End-User Metadata Add Value?" *Communications of the ACM* (50:11), pp. 70-77.

Fowler, M. (2003). *Patterns of Enterprise Application Architecture*, Addison-Wesley, Boston, Massachusetts.

Harder, H. (2010). "Explanatory Case Study", in *Encyclopedia of Case Study Research,* Albert J. Mills, Gabrielle Durepos, Elden Wiebe (eds.), SAGE Publications Inc, pp. 371-372.

HL7 (2014). "The HL7 V3 Encounter Standard," Health Level Seven International. Ann Arbor, Michigan, USA.

Holley, K., S. Antoun, A. Arsanjani, W. A. Brown, C. Cozzi, J. F. Costas, P. Goyal, S. Lyengar, H. Jamjoom, C. Jensen, J. Laredo, J. Maddison, R. Narain, A. Natarajan, J. Petriuc, K. Ramachandran, R. Ravishankar, R. Reinitz, S. Vaidya, and M. Vukovic (2014). "The Power of the API Economy: Stimulate Innovation, Increase Productivity, Develop New Channels and Reach New Markets," *Redguides for Business Leaders* (REDP-5096-00). IBM Corp.

ISO (1998). "ISO 9241-11:1998 Ergonomic Requirements Office Work Visual Display Terminals (VDTs) - Part 11: Guidance on Usability," International Organization for Standardization.

Janssen, M., Y. Charalabidis, and A. Zuiderwijk (2012). "Benefits, Adoption Barriers and Myths of Open Data and Open Government," *Information Systems Management* (29:4), pp. 258-268.

Katz-Haas, R., and Y. W. Lee (2005). *"Understanding Interdependencies between Information and Organizational Processes",* in *Information Quality,* Richard Y. Wang, Elizabeth M. Pierce, Stuart E. Madnick and Craig W. Fisher (eds.), M.E. Sharpe, Inc., Armonk, New York, USA, pp. 167-178.

Koppel, R., T. Wetterneck, J. L. Telles, and B. Karsh (2008). "Workarounds to Barcode Medication Administration Systems: Their Occurrences, Causes, and Threats to Patient Safety," *Journal of the American Medical Informatics Association* (15:4), pp. 408-423.

Laine, S., and E. Niemi (2013). "Transparency of Hospital Productivity Benchmarking in Two Finnish Hospital Districts," in *the Patient Classification Systems International (PCSI)*, Helsinki, Finland.

Lee, Y. W., and D. M. Strong (2003). "Knowing-Why about Data Processes and Data Quality," *Journal of Management Information Systems* (20:3), pp. 13-39.

Linna, M., and U. Häkkinen (2007). "Benchmarking Finnish Hospitals", in *Evaluating Hospital Policy and Performance: Contributions from Hospital Policy and Productivity Research,* Jos L. T. Blank and Vivian G. Valdmanis (eds.), pp. 179-190.

Madnick, S. E., R. Y. Wang, Y. W. Lee, and H. Zhu (2009). "Overview and Framework for Data and Information Quality Research", *Journal of Data and Information Quality* (1:1), pp. 1-22.

Mills, A. J., G. Durepos, and E. Wiebe (2010). *Encyclopedia of Case Study Research*, SAGE, Los Angeles, USA.

Moran, T. P., and P. Dourish (2001). "Introduction to this Special Issue on Context-Aware Computing", *Human-Computer Interaction* (16:2), pp. 87-95.

Parsons, J., and Y. Wand (2014). "A Foundation for Open Information Environments", in *the Proceedings of European Conference on Information Systems*, Tel Aviv, Israel.

Saleem, J. J., A. L. Russ, C. F. Justice, H. Hagg, P. R. Ebright, P. A. Woodbridge, and B. N. Doebbeling (2009). "Exploring the Persistence of Paper with the Electronic Health Record", *International Journal of Medical Informatics* (78:9), pp. 618-628.

Shankaranarayanan, G. (2000). "IP-MAP: Representing the Manufacture of an Information Product," in *the Proceedings of the International Conference on Information Quality, MIT, Boston, Massachusetts, USA*.

Stoop, A. P., K. Vrangbæk, and M. Berg (2005). "Theory and Practice of Waiting Time Data as a Performance Indicator in Health Care: A Case Study from the Netherlands", *Health Policy* (73:1), pp. 41-51.

Tremblay, M. C., A. Hevner, and D. Berndt (2010). "Focus Groups for Artifact Refinement and Evaluation in Design Research," *Communications of the Association for Information Systems* (26:1), pp. 599-618.

van Walraven, C., and P. Austin (2011). "Administrative Database Research has Unique Characteristics that can Risk Biased Results," *Journal of Clinical Epidemiology* (65:2), pp. 126-131.

van Walraven, C., C. Bennett, and A. J. Forster (2011). "Administrative Database Research Infrequently used Validated Diagnostic or Procedural Codes," *Journal of Clinical Epidemiology* (64:10), pp. 1054-1059.

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

Wang, R. Y., Y. W. Lee, L. L. Pipino, and D. M. Strong (1998). "Manage Your Information as a Product", *Sloan Management Review* (39:4), pp. 95-105.

Yang, Z., B. Ng, A. Kankanhalli, and J. W. L. Yip (2012). "Workarounds in the use of IS in Healthcare: A Case Study of an Electronic Medication Administration System," *International Journal of Human-Computer Studies* (70:1), pp. 43-65.