Telematics and Informatics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele

Semantic privacy-preserving framework for electronic health record linkage

Yang Lu *, Richard O. Sinnott

Computing and Information System, University of Melbourne, Victoria 3010, Australia

ARTICLE INFO

Article history:
Received 3 September 2016
Received in revised form 27 April 2017
Accepted 11 June 2017
Available online xxxx

ABSTRACT

The combination of digitized health information and web-based technologies offers many possibilities for data analysis and business intelligence. In the healthcare and biomedical research domain, applications depending on electronic health records (EHRs) identify privacy preservation as a major concern. Existing solutions cannot always satisfy the evolving research demands such as linking patient records across organizational boundaries due to the potential for patient re-identification. In this work, we show how semantic methods can be applied to support the formulation and enforcement of access control policy whilst ensuring that privacy leakage can be detected and prevented. The work is illustrated through a case study associated with the Australasian Diabetes Data Network (ADDN – www.addn.org.au), the national paediatric type-1 diabetes data registry, and the Australian Urban Research Infrastructure Network (AURIN – www.aurin.org.au) platform that supports Australia-wide access to urban and built environment data sets. We demonstrate that through extending the eXtensible Access Control Markup Language (XACML) with semantic capabilities, finer-grained access control encompassing data risk disclosure mechanisms can be supported. We discuss the contributions that can be made using this approach to socio-economic development and political management within business systems, and especially those situations where secure data access and data linkage is required.

1. Introduction

Data analysis and especially real-time data analysis is key to designing successful data management systems (Abello et al., 2015). Since the concept of *business intelligence* (BI) was proposed in 1990s, it has been widely-applied to describe data-centric applications with analytical functionalities. BI requires data inputs and produces data outputs (Wixom et al., 2014). Business users and services often perform analytical techniques based on data warehouse approaches, where data is collected and aggregated from different sources into a single large-scale repository. By implementing BI solutions in different domains, IT infrastructure costs can be reduced. Such an approach can also provide more accurate analytical results and save time for stakeholders and data users (Saltz, 2015). However it is often the case that this aggregation is infeasible or impossible due to security and privacy concerns. This is especially the case in the medical domain and use of data with other forms of social or environmental data for example.

Due to the seismic transformation from paper-based patient data to electronic heath records (EHRs) that is occurring, the developing health IT infrastructure allows stakeholders across health industries to appreciate significant benefits such as remote care delivery and cohort recruitment. To further improve the effectiveness, efficiency, and quality of health services,

* Corresponding author.
E-mail addresses: luy4@student.unimelb.edu.au (Y. Lu), rsinnott@unimelb.edu.au (R.O. Sinnott).

http://dx.doi.org/10.1016/j.tele.2017.06.007
0736-5853/

the adoption of BI in the health sector is regarded as an opportunity to improve health care more generally (Mettler and Vimarlund, 2009). For instance, analysts are allowed to predicate epidemics of certain communities by exploring datasets with knowledge discovery techniques (Lopez et al., 2014). Healthcare BI could save costs for both providers and patients by removing redundancy and making more accurate decisions. An example of this is long-term care in the Netherlands (Spruit et al., 2014), where it was shown that healthcare business intelligence could promote knowledge discovery based on large datasets focused on chronic care.

According to recent studies, it has been recognized that patient-centred care is now the key to high-quality healthcare delivery (Dowsett et al., 2000; Kwan and Sandercock, 2004; Levesque et al., 2013; Pulvirenti et al., 2014). By conducting patient-focused approaches it is hoped to achieve maximum biomedical value by sharing information between professionals, patients and carers (Coulter and Collins, 2011). A key point in this approach is to maintain the interoperability of the systems where doctors and health researchers may exchange clinical decisions and population-based analysis results. To make this happen, it is necessary to build systems from a holistic perspective in making decisions and treatment planning. For instance, record linkage techniques in the biomedical domain allow a more complete picture of the health of the population than previously possible. Currently linkage techniques are widely employed in Australia such as Centre for Healthcare Record (CHeReL)[1] in New South Wales, SA-NT DataLink[2] in South Australia and Victorian Data Linkage (VDL)[3] in Victoria.

Challenges in the linkage-oriented systems include access control and data anonymity. Biomedical data containing patient demographics can be too sensitive to release. To eliminate malicious or non-malicious threats upon publishing EHRs to users, data custodians have adopted a range of solutions such as requiring informed consents, de-identification and reviewing the purpose of access data as applied. Technically, the authorization languages for restricting malicious operations on data should be interoperable among autonomous organizations. Due to the modularity and extensibility, XACML has been widely implemented for distributed security. Through specifying generic vocabularies as well as domain-specific facts, access requests for linkage data can be verified against XACML policies. However, implicit associations among concepts are typically neglected in manual data linkage operations, which may lead to "authorization missing" or data leakage dangers. Whilst data may not immediately be compromised, there is a danger of gradual erosion of privacy and the risk of re-identification of individuals through linkage with other data sources. To address this issue, this paper explores how semantic web technologies can be used to discover latent disclosure risks and mitigate the chance of privacy leakage.

2. Background

The ability to share clinical data in secure ways outside of the immediate health context is essential. In the field of healthcare and clinical research, online data for secondary use allows new discoveries on drugs, treatments, and clinical benchmarking more generally (Synnot et al., 2016). For instance, through comparing clinical records with existing metrics, hidden factors of complex diseases can be ascertained (Kontos et al., 2014). Meanwhile, applying new theories in disease prediction and effective control, it is expected to improve people's health and life quality (Abdelhak et al., 2014). Since it may involve personal privacy, researchers are required to use health data in an ethical and confidential way. Typically three procedures are required before accessing and using health data (O'Keefe and Connolly, 2010).

- *Informed consent*. Biomedical data usually contains both identifying and non-identifying information. Data custodians (hospitals/clinical institutions) are often required to collect consent letters from data subjects (patients). Upon receiving the confirmation, they can allow the research use of the sensitive information.
- *Anonymity*. To protect patient privacy, some health data is required to be anonymized for secondary use. Various degrees of anonymity are often conducted in accordance with the research nature, purposes, as well as the regions and even countries. For instance, comparing influenza data and HIV/AIDS data may have different levels of restrictions and sensitivities on anonymisation.
- *Access control*. Data access should be authorized through evaluating security policies, where domain-specific knowledge is often required. Thus what roles are required to access what data and in which context. Rules are typically formalized to minimize potential security risks. These policies tend to be statically defined and inflexible to more dynamic linkage scenarios.

2.1. Data anonymity

Data anonymity is legislated by many countries to guide and mentor data processing in research activities. For instance, the National Statement on Ethical Conduct in Human Research (2007) has categorized data into individually identifiable, re-identifiable and non-identifiable. In addition, they specify in what contexts, which type of data is allowed to be collected, stored and published. In the U.S., privacy issues of health data are regulated by Health Insurance Portability and Accountability Act 1996 (HIPPA) in which de-identified data levels and guidelines on different levels of anonymity have been stressed.

[1] Centre for Health Record Linkage (CHeReL). http://www.cherel.org.au/.
[2] SA-NT DataLink. https://www.santdatalink.org.au/.
[3] Victorian Data Linkage (VDL). https://www2.health.vic.gov.au/about/reporting-planning-data/victorian-data-linkages.

Please cite this article in press as: Lu, Y., Sinnott, R.O. Semantic privacy-preserving framework for electronic health record linkage. Telemat. Informat. (2017), http://dx.doi.org/10.1016/j.tele.2017.06.007