2

Anonymization Approach for Protect Privacy of Medical Data and Knowledge Management

Asmaa Hatem Rashid and Norizan Binti Mohd Yasin
Department of Information Science, Faculty of Computer Science and IT,
University of Malaya, Kuala Lampur,
Malaysia

1. Introduction

The evolution and development of information and technology have facilitated greater sharing and knowledge management of the collection of electronic information provided by data owners, including governments, corporations, and individuals. Such owners create significant opportunities for knowledge management and information retrieval, thus improving decision-making.Correspondingly; the increase in the use of the Internet and its applications in various aspects of life has led to the need to secure data in the medical and research fields, in government offices, corporations, and individual agencies in various fields. Two questions are addressed in the present study. First, why is there an increasing demand for data sharing and knowledge management? This increasing demand is reflected in the rate of demand for data sharing (Figure 1), which is the base reference data for all users (Gardner and Xiong 2009; El Emam et al. 2011).

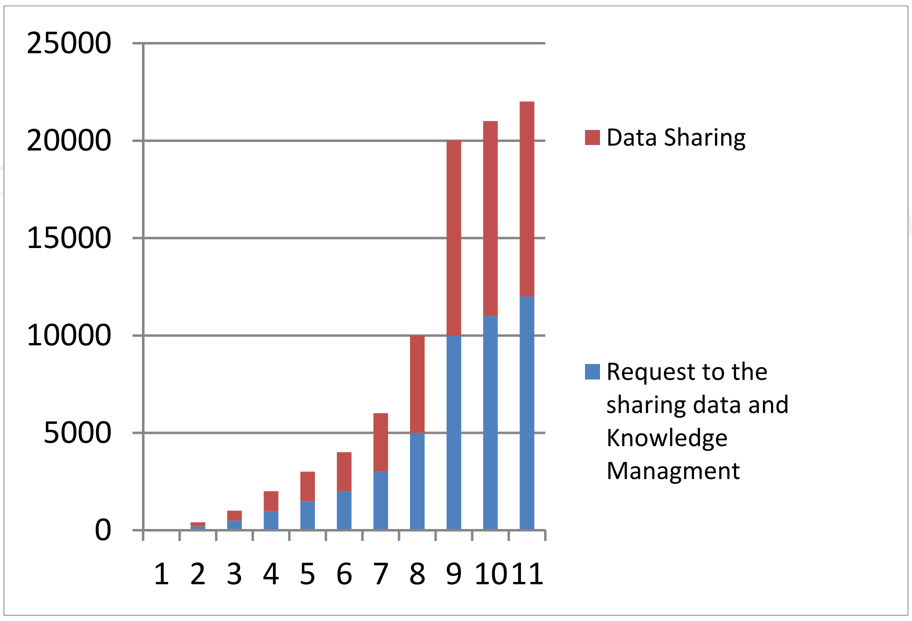

Fig. 1. Require to the Sharing Data and Knowledge Management.

Second, why is there an increase in the control, sharing, and managing of protected or sensitive data for knowledge management? As described in Figure 2, there is a rise in the demand for data sharing, which is the base reference data for all users ((Sweeney 2002)Gardner and Xiong 2009; El Emam, Jonker et al. 2011).

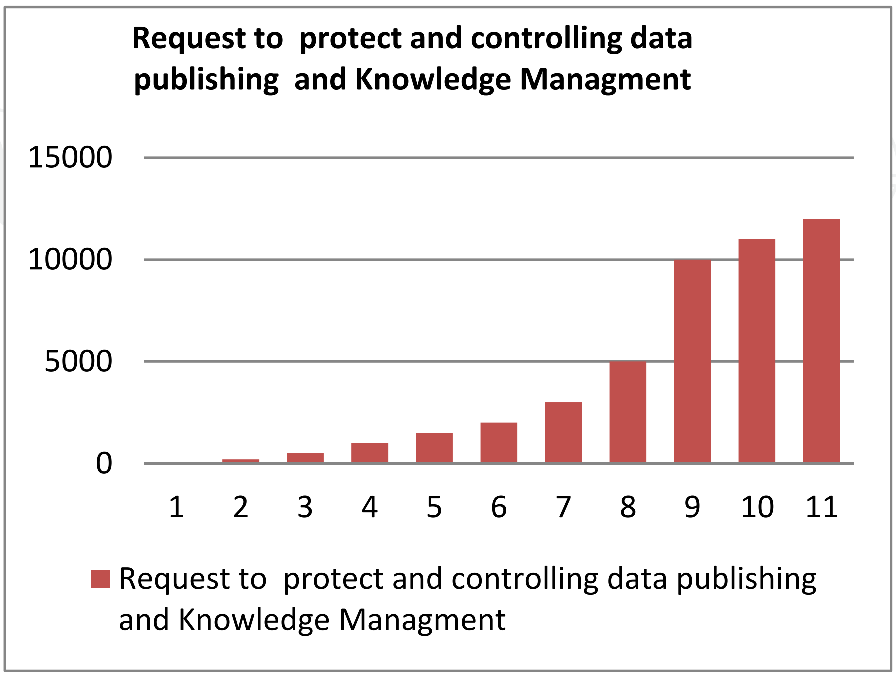

Fig. 2. Require to the protect and controlling data publishing and Knowledge Management.

The above questions imply that both data sharing and knowledge management are indeed on the rise. An increase in the exchange of data leads to the sharing of knowledge and drawing of conclusions based on real data. In turn, data sharing and knowledge management lead to more profit and benefits for service providers and customers. Furthermore, data sharing and knowledge management promote awareness and distribution of knowledge to support decision making in different sectors. Data sharing provides a single source of data to lessen the financial cost in collecting data from research and repeated operations, which require more time and effort. The second question relates to the increase in the control, sharing, and managing of protected or sensitive data and knowledge management. Controlling data exchange and ensuring the security of sensitive data for customers lead to increased trust between the service provider and the customer, promoting their strong relationship in the long run. Information sharing in the medical field supports many decision-making processes.

The strong relationship between patients and the hospital and the link among hospitals lead to better decisions on the management and health of patients. Management of such relationship or Patient Data Management can be assessed using the Customer Relationship Management (CRM) test.

One method that allows health information to be used and declared under existing legal frameworks is de identification. De identification refers to a set of methods that can be applied to data to ensure that the chance of assigning a correct identity to a record in the data is low (El Emam, Jonker et al. 2011).

In performing de-identification, we deduce from the relationship among the size of the data shared, knowledge management, and ways of controlling and sharing data in knowledge management.

A positive or a direct relationship is one that is shared by two parties in which a change in one variable is associated with a change in another variable in the same level. For example, an increase in the volume of data increases mutual knowledge and the control and sharing of data in knowledge management. Figure 3 describes the relationship between the size of data shared and the control and sharing of data in knowledge management (Gardner and Xiong 2009; El Emam, Jonker et al. 2011).

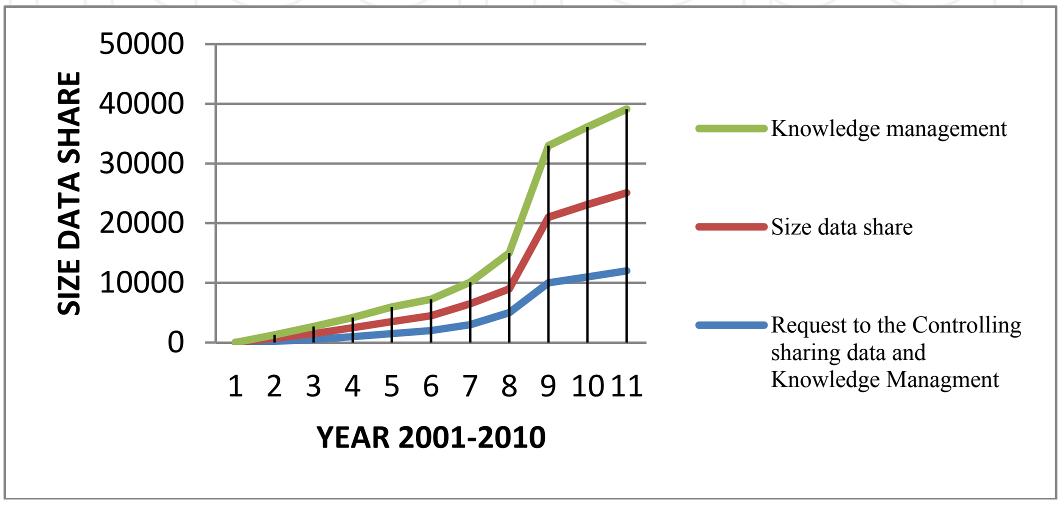

Fig. 3. The Relationship between Data Share Size and Controlling and sharing data of Knowledge Management.

The above motivated us to do work in this area, which will also hopefully encourage research on the control and sharing of data in knowledge management in other fields.

2. Related works

In the past few years, the issue of improving the control and sharing of data in knowledge management has attracted substantial interest among individual users and service providers such as research centers, companies, and governments. This growing interest confirms the importance of the subject and the sensitivity of the work and research involved. Most reviewers and researchers agree on the significant problems present in the control and sharing of data in knowledge management. The problems posed are under more than one research area. These include data confidentiality and privacy protection, with the review of suggested solutions to the problems involving confidential data for cryptographic information and hidden data (Chen and Chu 2008; Kantarcioglu, Jiang et al. 2008), among others. The second area involves data mining and data-mining algorithms to ensure privacy (Bertino, Fovino et al. 2005; Yeh and Hsu 2010), such as generalization and suppression techniques (Hintogdlu and Saygin 2010; Yang and Qiao 2010; Rashid 2010). Other research areas of data management and control of sharing of data ensure integration especially in medical information systems (El Emam, Jonker et al. 2011) and knowledge base systems, which will be the focus of the present study.

In most fields, sharing of data needs the control and management of such data to ensure system integration (Gardner and Xiong 2009), such as patient data, without revealing any sensitive information that can identify a patient. There are several studies that focus on the management of data in medical applications to ensure system integration. However, this can result in information misuse. Nevertheless, there are many algorithms and methods that facilitate management of shared data using techniques such as removing sensitive characters from the information system. Such algorithms are used to prevent unauthorized access to the original data for illicit purposes. In the present research, the main problem is the identification of an algorithm that provides control and management of shared data. Updating current data can be useful for future purposes such as analysis and knowledge management to support decisions in different fields of medical applications; however, the updated data should still represent real cases.

The current study will address this main problem through analyzing and evaluating the following sub problem: There is no model that can identify the number of quasi-identifier characters in such a way that the shared data are managed and a new version of such data is always usable. There is a lack of trust among medical system providers in sharing data and managing knowledge. Aside from the lack of a centralized database to keep the collected data, the problem of case indexing is still left unresolved due to the inability to update data in such a way that it can be used for further analysis and studies. The lack of high-quality updated data and the possibility of errors that adversely affect the results of studies depend on the crucial task of updating the data. As such, the present study attempts to fill in these gaps.

There is a need to build models or design algorithms that manage the sharing of data to avoid misuse. The goal is to bring authenticity to the data system. Guided by recent studies from the years 2005 to 2011 on the control and sharing of data in knowledge management (Fung, Wang et al. 2010), the current work notes that most reviewers and researchers have focused on ensuring the privacy of sensitive data.

In other words, great concern has been directed on the control of data and its sharing to make it available to their owners. Some reviewers and researchers have even suggested the use of covert techniques which isolate data such as encryption technology. Different ways of protecting data have been dealt with in recent research. The methods previously introduced include information on how to spread and use data in research, decision making, scientific analyses, and other purposes (Fung, Wang et al. 2010). First, the concern is how to control data sharing and management and avoid the risk of publishing data that may lead to revealing the real data. Second, there is lack of unity among the collected data, and their sources vary as they are collected from various points such as governments, hospitals, companies, and so on. Third, the data collected may contain errors. How data are processed and formatted before access requires a high level of analysis techniques to extract and determine knowledge and relationships hidden. To identify the relationships among different data and their influence on the results, they must be accurate and correct, as one type of data relies on the results of the analysis.

Examples are the reasons for the spread of a particular disease in a particular area in the medical field, the losses incurred by a company after a change in business strategy, and the low standards of living in a society. The main objective of the present research is to control management and sharing of data in the medical field, which mainly involves "patient data." Our main objective is to propose means to preserve information. The secondary objectives, which relate to the removal of sensitive data, are as follows:

- To evaluate and identify the parameters that negatively affect the management of shared patient data, thus determining the reasons behind the decrease in trust between private and health information communities
- To evaluate and measure the efficiency of k-anonymization and generalization methods in privacy and misuse protection (El Emam, Jonker et al. 2011)
- To build a model that can help prevent shared patient data from being misused
- To test the information metric method using real medical information
- To ensure high-quality information in every stage of the model

Some research questions on the control and sharing of data in knowledge management are as follows: How can data be kept unidentified? How can shared data be managed, ensuring that these benefit the target communities? What indexing methods should be followed to facilitate accurate and fast indexing of a case? How should the effect of perturbation on scientific analysis be measured, and what is an acceptable effect?

3. Information security is not privacy protection

Through this research we would like to clarify the difference between information security and privacy protection, where there is a common area between two subjects where the confidentiality of the data associated with access control and authentication on the received data, which are the traditional areas associated with that in this area recipients of the information has the authority to receive that information.

The problem in this research is more complex and different for the confidentiality of information and very different from the principle of receiving data and how to protect data and which the recipient has the authority receipt.

As the general principle of this research is to release all the data so that the use of data sent or published in scientific fields, but must protect the identities of people who are the landlords of such data or other sensitive properties found in the data). Therefore,(Sweeney 2002).

The aim of this work presented in this research is located outside the traditional work on access and authentication control.

4. Relationship between CRM and control for sharing data

CRM is an integration of people, processes, and technology, which seeks to understand an organization's customers. It is an integrated approach to managing relationships by focusing on customer retention and relationship development.

CRM has been developed from advancements in information technology and organizational changes in customer-centric processes (Chen and Popovich 2003). Figure 4 illustrates the CRM model.

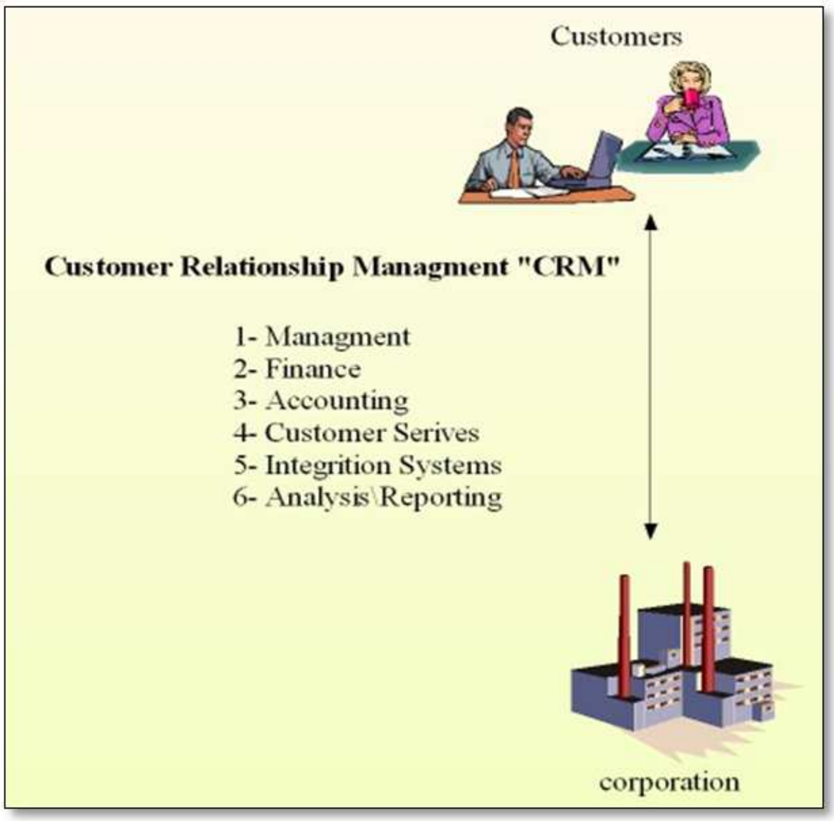

Fig. 4. Customer relationship management (CRM).

Corporations that successfully implement the CRM model gain customer trust and profitability in the long run. However, its successful implementation is unattainable in many organizations mostly because they do not understand that CRM requires organization-wide, cross-functional, and customer-focused business process reengineering (Chen and Popovich 2003).

Usually, CRM is applied in the business field but not in the medical one. The application of the CRM model can result in desirable results through linking the system and the approaches of one hospital with another.

A centralized database is developed, linked with the second database, and then with the third one, thus gathering data from various quarters. Linking of databases allows system integration and facilitates privacy of patient data. Algorithms are used to control sharing of data and knowledge management, as well as maintain privacy of patient data. Using CRM, research in the medical field is simplified by providing one data flow source. CRM encourages scientific research, supports the conclusions gained from the data, and saves patients' time and effort when seeking treatment. The approach also aids in finding the best treatment at the lowest cost and shortest time possible. The technology employing CRM is called Patient Relationship Management (PRM). (Figure 5) describe the Integration Hospital Database System.

Considering the advantages gained by CRM institutions, we therefore recommend the use of PRM. Through PRM, patients establish a connection with the organization (hospital) through an integrated approach of controlling patient data sharing and management. By

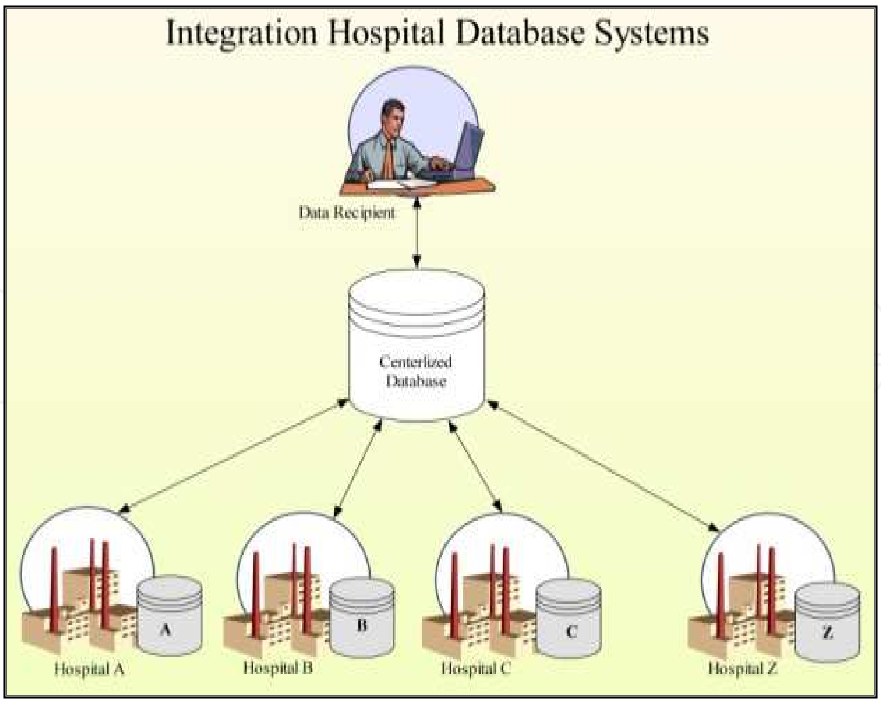

Fig. 5. Integration Hospital Database System.

examining CRM, we learn more about its importance and the advantages and benefits gained by the organizations which use them. We will focus on the processes carried out by the staff in the control of patient data sharing and management.

The following questions arise in the process: How can we control the process of participation and patient data in the medical field? As these data will be collated into a database, this storage combines data from different sources. After the gathering of such data is their organization to form a a knowledge base. We then determine the degree of data privacy involved and the limits to which they can be disclosed so that the identity of the owners of the data can also be protected.

We clarify the process of sending information, for example from a hospital to a central database system, to determine if the divisions have reached integration. To preserve and control the data in this case, we use LeFevre et al.'s (2006) generalization technique. The generalization of the domain values of relational characters to more general values uses the process of distribution of data.

The technique converts the data from private to public while still preserving its usefulness. We delete sensitive information about patients such as their name, identification number, and other details that should be removed, then apply the rules of law agreed upon between the hospital and the patients. Figure 6 provides an example of the generalization technique processing of "patient data" (LeFevre, DeWitt et al. 2005). The general technique includes domain and value generalization hierarchies for zip code (a, b), birth date (c, d), and gender (e, f) (LeFevre, DeWitt et al. 2005).

Using the generalization technique enables us to control the data to be shared and to send it to a central database. After analysis and processing, we provide a knowledge base of real data and results on a scientific basis to provide general information for research and other measures that need the results based on real data.

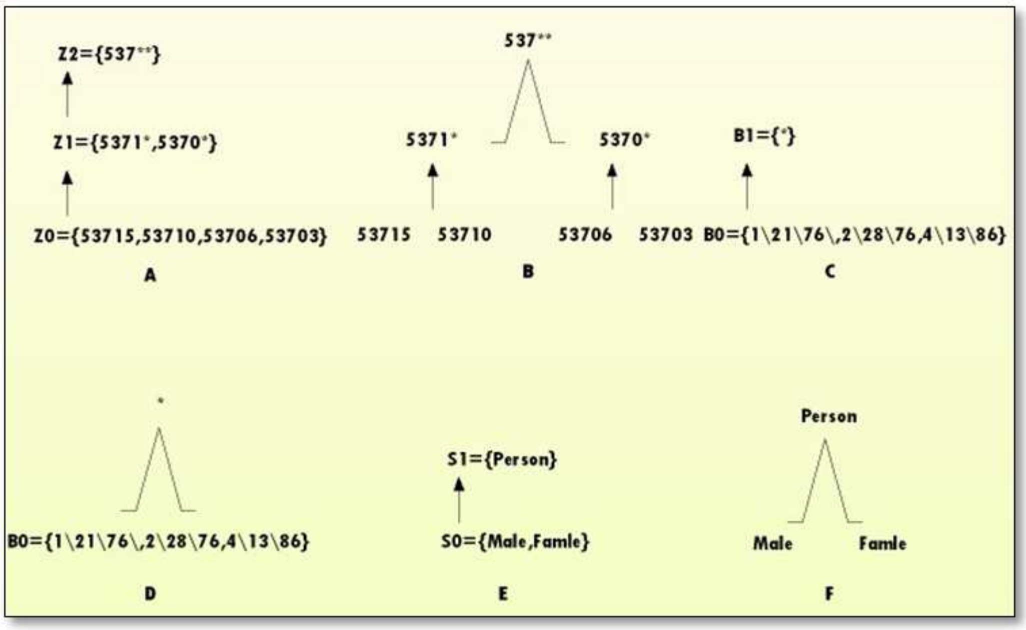

Fig. 6. Domain and value generalization hierarchies for Zip code (a, b), Birth Date (c, d), and Sex (e, f)(LeFevre, DeWitt et al. 2005).

Privacy is addressed by preventing distribution rather than integrating privacy constraints into the data sharing process. Privacy-preserving integration and sharing of research data in health sciences have become decisive in enabling scientific discovery as cited in Sharing Scientific Research Data (Clifton, Kantarcio lu et al. 2004).

5. State-of-the-art privacy preserving

We briefly review the most relevant areas below and discuss how our work levels up with current state-of-the-art systems.

5.1 Privacy preservation in data publication.

The preservation of privacy when publishing data for centralized databases has been examined intensively in recent years. One thread of work aims at devising privacy principles such as k-anonymity and subsequent principles that address problems, which in turn serve as criteria for judging whether a published data set enables privacy protection (Sweeney 2002; Nergiz and Clifton 2007). Another body of work has contributed to the development of an algorithm that transforms a data set to meet one of the privacy principles (dominantly k-anonymity). However, most of these works have focused only on structured data (Li, Li et al. 2007; Xiao and Tao 2007; Gardner and Xiong 2009).

5.2 Medical text de-identification

In the medical informatics community, there have been efforts in deidentifying medical text documents (Sweeney 2002; Zhong, Yang et al. 2005; Gardner and Xiong 2009). Most of them use a two-step approach which extracts the identifying characters first and then removes or masks the attributes for deidentification purposes. Most of them are specialized for specific document types, for example, pathology reports only (Zhong, Yang et al. 2005; Gardner and Xiong 2008). Some systems focus on a subset of Health Insurance Portability and Accountability Act (HIPAA) identifiers, for example, name only (Aramaki, Imai et al. 2006; Gardner and Xiong 2009), whereas others focus on differentiating protected health information (PHI) from non-PHI (Gardner and Xiong 2009). Most importantly, most of these studies rely on simple identifier removal or grouping techniques, and they do not take advantage of recent research developments that guarantee a more formalized notion of privacy while increasing data utility.

5.3 Information extraction

Extracting atomic identifiers and sensitive characters (such as name, address, and disease) from unstructured text such as pathology reports can be seen as an application of the named entity recognition (NER) problem (Neumann 2010). NER systems can be roughly classified into two categories, both of which are applied in medical domains for deidentification. The first uses grammar-based or rule-based techniques (Gardner and Xiong 2008). Unfortunately, such hand-crafted systems may take months of work by experienced domain experts, and the rules will likely change for different data repositories. The second category uses statistical learning approaches such as support vector machine (SVM)-based classification methods. However, an SVM-based method such as that introduced by Sibanda and Unuzer (Sibanda and Uzuner 2006) only performs binary classification of the terms into PHI or non-PHI. It does not also allow statistical deidentification which requires knowledge on different types of identifying characters.

6. Novelty and technical contribution

In the following, we explain the novelty and technical contributions of the survey to data privacy through the control and sharing of data in knowledge management. We focus on six aspects of technical contributions, which we consider to be the most interesting (Xiao 2009).

6.1 Personalized privacy preservation

We examined the work of (Xiao and Tao 2006) on the publication of sensitive data using generalization, the most popular anonymization methodology in the literature. The existing privacy model for generalized tables (that is, noisy microdata obtained through generalization) exerts the same amount of protection on all individuals in the data set without catering to their concrete needs. For example, in a set of medical records, a patient who has contracted flu would receive the same degree of privacy protection as a patient suffering from cancer, despite the willingness of the former to reveal his/her symptoms directly (mainly because flu is a common disease) (Xiao and Tao 2006). Motivated by this, we propose a personalized framework that allows each individual to specify his/ her preferred privacy protection in relation to his/her data. Based on this framework, we devised the first privacy model that considers personalized privacy requests. We also developed an efficient algorithm for computing generalized tables that conform to the model. Through extensive experiments, we show that our solution outperforms other generalization techniques by providing superior privacy while incurring the least possible information loss (Xiao and Tao 2006).

6.2 Republishing dynamic data sets

Data collection is often a continuous process, where tuples are inserted into and deleted from the microdata as time evolves. Therefore, a data publisher may need to republish the microdata at multiple times to reflect the most recent changes. Such republication is not supported by conventional generalization techniques because microdata are assumed to be static (Xiao and Tao 2007). We address this issue by proposing an innovative privacy model called m-invariance which secures the privacy of any individual involved in the republication process, even against a rival who exploits the correlations between multiple releases of the microdata. The model is accompanied by a generalization algorithm whose space and time complexity are independent of the number n of generalized tables that have been released by the publisher. This property of the algorithm is essential in the republication scenario, where n increases monotonically with time (Xiao and Tao 2007).

6.3 Complexity of data anonymization

We have presented the first study on the complexity of producing generalized tables, which conform to ℓ-diversity, the most commonly adopted privacy model. We note that achieving ℓ-diversity with minimum information loss is NP-hard for any ℓ larger than two and any data set that contains at least three distinct sensitive values. Considering this, we developed an O(ℓ.d)-approximation algorithm, where d is the number of QI characters contained in the microdata (Xiao 2008). Aside from its theoretical guarantee, the proposed algorithm works fairly well in practice and considerably outperforms state-of-the-art techniques in several aspects (Xiao 2008).

6.4 Transparent anonymization

Previous solutions for data publication consider the idea that the rival controls certain prior knowledge about each individual. However, they overlook the possibility that the rival may also know the anonymization algorithm adopted by the data publisher. Thus, an attacker can compromise the privacy protection enforced by the solutions by exploiting various characteristics of the anonymization approach (Xiao 2008). To address this problem, we propose the first analytical model for evaluating the disclosure risks in generalized tables under the assumption that everything involved in the anonymization process, except the data set, is public knowledge. Based on this model, we developed three generalization algorithms to ensure privacy protection, even against a rival who has a thorough understanding of the algorithms. Compared with state-of-the-art generalization techniques, our algorithms not only provide a higher degree of privacy protection but also satisfactory performance in terms of information distortion and overhead estimation (Xiao 2008).

6.5 Anonymization via anatomy

While most previous work adopts generalization to anonymize data, we propose a novel anonymization method anatomy which provides almost the same privacy guarantee as generalization does. However, it significantly outperforms it in terms of the accuracy of data analysis on the distorted microdata (Xiao and Tao 2006). We provide theoretical justifications for the superiority of anatomy over generalization and develop a linear time algorithm for anonymizing data via anatomy. The efficiency of our solution was verified through extensive experiments.

6.6 Dynamic anonymization

We propose dynamic anonymization which produces a tailor-made anonymized version of the data set for each query given by users; the anonymized data increases the accuracy of the query result. Privacy preservation is achieved by ensuring that no private information is revealed despite combining all anonymized data (Xiao 2008). For example, even if the rival obtains every anonymized version of the data set, he/she would not be able to infer the sensitive value of any individual. Through extensive experiments, we show that compared with existing techniques, dynamic anonymization significantly improves the accuracy of queries on the anonymized data (Xiao 2008).

7. Models to control the publication of data

After reviewing the models used in previous research and determining the results of the present study, we make a comparison between the results of the sample and those of different disciplines. Some of the findings were categorized under the confidentiality and privacy of data, whereas others were categorized under the control of post-data.These models are as follows. After searching and accessing a number of studies, we found models used to protect data and those that manage the privacy of data. This step helped us develop and improve the privacy and dissemination of data that can be used in various disciplines while maintaining the same degree of privacy needed (Bugliesi, Preneel et al. 2006; Fung, Wang et al. 2010). the following table describes Models to control publishing of data.

NO.	MODEL NAME
1	k-Anonymity
2	Multi R k-Anonymity
3	ℓ Diversity
4	Confidence Bounding
5	(a; k)-Anonymity
6	(X; Y)-Privacy
7	(k; e)-Anonymity
8	(€;m)-Anonymity
9	Personalized Privacy
10	t-Closeness
11	£, Presence
12	(c; t)-Isolation
13	E-Differential Privacy
14	(d; y)-Privacy
15	Distributional Privacy

Table 1. Models to protect and controlling data publishing (Fung, Wang et al. 2010).

In today's information society, given the unprecedented ease of finding and accessing information, protection of privacy has become a very important concern. In particular, large databases that include sensitive information (e.g., health information) have often been available to public access, frequently with identifiers stripped of in an attempt to protect privacy. However, if such information can be associated with the corresponding people's identifiers, perhaps using other publicly available databases, then privacy can be seriously violated.

For example, (Sweeney 2002)pointed out that one can find out who has what disease using a public database and voter lists. To solve such problems, (Samarati and Sweeney 1998)have proposed a technique called k-anonymization. In this research, we study how to enhance privacy in carrying out the process of k-anonymization.

8. The framework of the proposed model for controlling and managing data sharing

The framework suggested by the present work consists of three stages. As explained in Figure 7, the first stage is when the provider sends data from different databases into an expert database. At this stage, the problem is how to preserve the confidentiality of data sent to the main database. We assume that the connection between the data provider and the centralized database is characterized by trust (Rashid 2010).

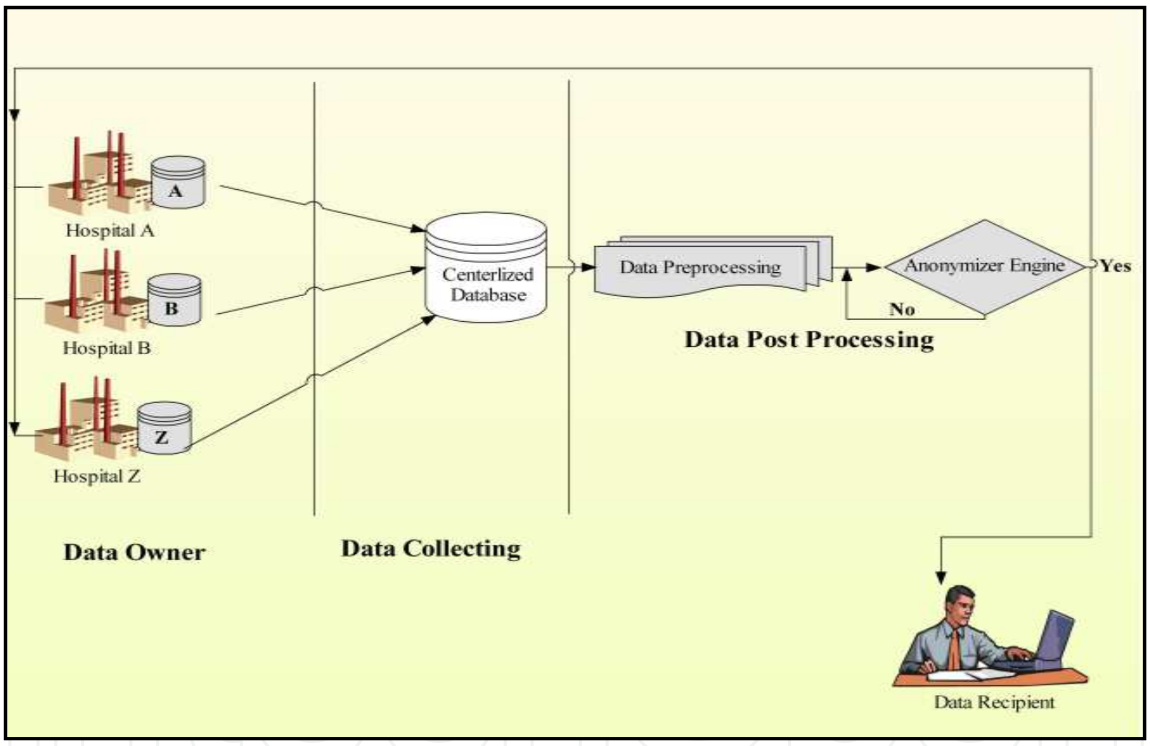

Fig. 7. the Proposed Model to protecting and controlling and manage data sharing.

The second stage is when the expert receives data in the database and recreates (reprocesses) these before sending to the anonymizer engine that applies the k-anonymization and generalization technique. Thus, the second stage is designed for preparing and obtaining data.

The third stage applies data mining algorithms such as analysis, which should identify the hidden relationships among various data and extract results supporting scientific research and decision making. The last stage is the publication of the results on the Web site.

The interface used (published data) and the last version (results) should appear in a simple style to ensure understanding by the recipient (Rashid 2010).

9. Discussion

The purposes of the study on field data anonymization and knowledge management are to allow the release of scientifically useful data in a form that protects the privacy of its subjects and publish knowledge based on real data. Implementing these goals requires more than simply removing personal identifiers from the data because an attacker can still use auxiliary information to infer sensitive individual information.

Additional perturbation is necessary to prevent such inferences, and perturbation of the data in a way that preserves their analytical utility is of significant importance. The great challenge in producing an anonymization scheme is the provision of adequate privacy protection while minimally affecting the analytical utility of the data, which is difficult to doing general and even more difficult to do with high dimensional data.

We previously introduced the observation that anonymity is not required to operate the original data source and proposed that transformation to a prudently chosen source can yield the proper combination of privacy protection, analytical utility, and computational efficiency of anonymization. Studies on data privacy protection have yielded basic criteria with which the degree of privacy required can be measured while maintaining the scientific usefulness of data analysis. The benefits of research and its importance in scientific development and the collection of analytical data based on real medical cases help promote and disseminate knowledge that could assist in the processes of scientific research and raise the level of understanding of the beneficiaries of the research results.

The following questions were raised during the course of our research: Why there is a growing demand for data exchange and knowledge management? How can the demands for controlled data sharing and knowledge management be met? Future research to yield real data can provide answers to these questions and encourage data providers to allow the exchange of personal data for scientific purposes while preserving privacy and sensitive data.

10. Conclusion

This work demonstrates the effort needed to set up a policy framework for the control and sharing of data in knowledge management in the medical field. Data sharing can help guide the nation’s adoption of health information technologies and improve the availability of health information and the quality of health care. The proposed control and sharing of data in knowledge management uses the k-anonymization model and generalization technique. The efficiency of these processes has been confirmed through the study and analysis of all processes involved and recent scientific research in the same domain. The control and sharing of data in knowledge management of medical information secure data between health care consumers and providers. The broad use of the proposed system has the potential to improve health care quality and prevent medical errors, thus increasing the efficiency of the care provided and reducing unnecessary health care costs. Moreover, the proposed system would increase administrative efficiency, expand access to affordable care, improve people’s health, and provide relevant data to support scientific research.

11. References

Aramaki, E., T. Imai, et al. (2006). *Automatic deidentification by using sentence features and label consistency.*

Fung, B., K. Wang, et al. (2010). "Privacy-preserving data publishing: A survey of recent developments." *ACM Computing Surveys (CSUR)* 42(4): 1-53.

Gardner, J. and L. Xiong (2008). *HIDE: An integrated system for health information de-identification*, IEEE.

Gardner, J. and L. Xiong (2009). "An integrated framework for de-identifying unstructured medical data." *Data & Knowledge Engineering* 68(12): 1441-1451.

LeFevre, K., D. J. DeWitt, et al. (2005). *Incognito: Efficient full-domain k-anonymity*, ACM.

Li, N., T. Li, et al. (2007). *t-closeness: Privacy beyond k-anonymity and l-diversity*, IEEE.

Nergiz, M. E. and C. Clifton (2007). "Thoughts on k-anonymization." *Data & Knowledge Engineering* 63(3): 622-645.

Neumann, R. G. (2010). "Information Extraction." *Architecture* 2: 05.11.

Samarati, P. and L. Sweeney (1998). *Generalizing data to provide anonymity when disclosing information*, ASSOCIATION FOR COMPUTING MACHINERY.

Sibanda, T. and O. Uzuner (2006). "Role of Local Context in De-identification of Ungrammatical, Fragmented Text." *North American Chapter of Association for Computational Linguistics/Human Language Technology (NAACL-HLT)*.

Sweeney, L. (2002). "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty Fuzziness and Knowledge Based Systems* 10(5): 557-570.

Xiao, X. (2008). "Privacy Preserving Data Publishing: A Research Summary."

Xiao, X. (2009). "Privacy Preserving Data Publishing: A Research Summary."

Xiao, X. and Y. Tao (2006). *Personalized privacy preservation*, ACM.

Xiao, X. and Y. Tao (2007). *M-invariance: towards privacy preserving re-publication of dynamic datasets*, ACM.

Zhong, S., Z. Yang, et al. (2005). *Privacy-enhancing k-anonymization of customer data*, ACM.