Research Journal of Information Technology 3(4):133-138, 2011
ISSN: 2041-3114

Submitted: November 09, 2011 Accepted: December 09 , 2011 Published: December 30, 2011

Leveraging the Involvement of Data Subjects towards Controlling Their Own Personal Information

[1]Abdulrahman H. Altalhi, [2]Zailani Mohamed Sidek and [3]Norjihan Abdul Ghani
[1]Department of Information Technology, College of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[2]Information Security Department, Advanced Informatics School (AIS), University Technology Malaysia, 54100 Kuala Lumpur, Malaysia
[3]Information Systems Department, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

Abstract: Privacy has become an important component in systems that handle personal information. Data subjects have a right to control their own personal information. Hippocratic Databases represent a new era of database technology that has been proposed to fulfill personal information privacy protection requirements at the database level. However, the concept of "purpose" is introduced because the owner of the information has no control over his/her own personal information. The owner, also referred to as the data subject, not only needs to control to whom his/her information should be disclosed but should also be able to access his/her own personal information regardless of the purpose. For this reason, we concluded that there are two requirements when managing personal information: The owner should be able to control his/her own personal information and should be able to access it at any time and for any purpose. This Study introduces a new architecture that fulfills the above requirements, which we refer to as the owner-controlled architecture for Hippocratic Databases. First, we highlight the importance of controlling personal information in an information flow model, and then we explain the architecture that supports the proposed model.

Key words: Data access, Hippocratic databases, owner-controlled, personal information

INTRODUCTION

Data privacy of individuals is currently a challenging problem. Our current digital world, with e-commerce technology, not only enables individuals to conduct their business virtually at any time in any place but also provides the capability of storing various types of information that the users reveal during their activities. For this reason, many organizations rely on database systems as the key data management technology for various tasks, such as day-to-day operations and critical decision-making (Bertino *et al*., 2007). Increasing amounts of data about individuals is collected, stored and accessed from databases to complete the required tasks.

Because individuals are becoming more concerned about their privacy, they are becoming more reluctant to perform business and transactions online, and many companies are losing a considerable amount of potential profit. Easy access to private personal information may cause the misuse of data, a lack of control over the information and other problems. By demonstrating good privacy practices, many companies attempt to utilize information analysis and knowledge extraction to provide better services to individuals without violating individual privacy. In principle, each individual should own, maintain and control his own personal information, allowing access to those who need his/her information for certain purposes needed at a specific time. Privacy International, quoted here from one of the articles published by the Centre for Independent Journalism January 17th, 2008, said the following:

> "On the one hand, individual information is not protected and can be used virtually by anyone for any purpose, …"

In general, this Study discusses the importance of owner-controlled actions on personal information. The main objective of this Study is to introduce owner controlled architecture. The proposed architecture was established in implementing the Hippocratic Database in order to protect the personal information.

Corresponding Author: Abdulrahman H. Altalhi, Department of Information Technology, College of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia, Tel: +966-2-695-1317; Fax: +966-2-640-0000

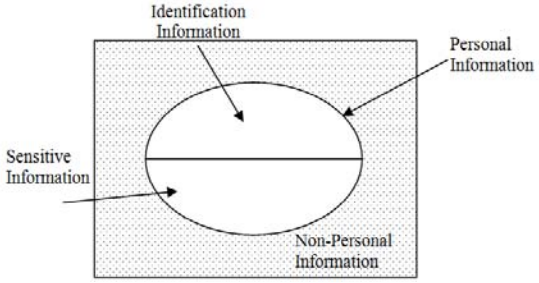

Fig. 1: Types of available information

LITERATURE REVIEW

This study is related to several topics in the area of privacy and security for data management, namely the identification of personal information and the databases that are tailored to privacy awareness. The amount of personal information that is stored in databases is rapidly growing (Bertino *et al*., 2007), and the awareness of privacy protection for data is rapidly increasing. To protect the privacy of data providers, Agrawal *et al*. (2002a) have emphasized that privacy protection should be included in the core capabilities of database systems and have proposed the concept of the Hippocratic database, which incorporates privacy protection in relational databases.

In any information systems, especially web-based information systems, various types of information are required to accomplish a task. This requirement includes not only personal information but also non-personal information. A substantial amount of work has been performed to provide a good solution for data privacy and protection. We first study the notion of personal information that was introduced by Al-Fedaghi and Thalheim (2009) and Guarda and Zannone (2009). Guarda and Zannone (2009) differentiate four different types of data that are involved during processing: personal data, sensitive data, identification data and anonymous data. From this classification and the definition given, sensitive data and identification data is a different thing. Meanwhile, Al-Fedaghi and Thalheim (2008) defines four types of information in the system: information, personal information, personal identifiable information and personal non-identifiable information. Metzler and Harker (2008) defines two types of Personal Identifiable Information (PII), which included moderately sensitive pieces of information, such as gender, name and e-mail address, and highly sensitive information, such as credit card numbers, passport number and identity card numbers. The difference between these two types of PII is that moderately sensitive PII is a type of PII that requires less protection, whereas highly sensitive PII requires more strict security.

Based on Al-Fedaghi and Thalheim (2008), personal identifiable information is any information that can be used to identify a person, regardless of whether it is sensitive or not. The sensitivity of personal information refers to the degree of its personal information privacy. Different people have different levels of sensitivity. For example, an email address is highly sensitive information for person A but it may be less sensitive for person B.

Westin (1967) stated that there are three statements on how people agree or disagree about PI privacy concerns:

- Consumers have lost all control over how PI is collected and used by companies
- Most businesses handle the PI they collect about consumers in a proper and confidential way
- Existing laws and organizational practices provide a reasonable level of protection for consumer privacy today

This research highlights the concept that different types of information which require different levels of privacy. Moreover, Agrawal *et al*. (2002a) introduced a new concept of database technology called Hippocratic Databases. Agrawal *et al*. (2002a) presented a privacy preserving database architecture called Strawman, which is based on the access control of notion purposes and opens up database-level research on privacy protection technologies.

PERSONAL INFORMATION

Data are important in any transaction, in either off-line transactions or online transactions. In transactions, users submit their personal information to obtain services provided by the company; meanwhile, organizations need personal information to conduct their business. There is a need for both the users and the organizations to agree on how the data are to be collected, used, stored and manipulated.

We believe that identification and sensitivity are two different concepts that should be considered in personal information privacy. With this understanding, we define and categorize a new classification of information: personal information and non-personal information. Personal information is any information that is related to a person, such as name, address, telephone number, and credit card number. Personal information is related to the general information about a person; meanwhile, non-personal information is any information that cannot be associated with a person, such as information about a company's details. Figure 1 shows the types of available information. There are four types of information, as defined below:

- **Personal information:** Information that is related to a person and is available on a system. There are two types of personal information:
  - **Identification information:** A subset of personal information that can be sensitive and can be associated with any identified person or is identifiable to a person. Identification information is personal information that permits the direct identification of the person, such as DNA, credit card number and identity card number.
  - **Sensitive information:** A subset of personal information that can be sensitive but cannot be associated with any identified person or is not identifiable to a person. Sensitive information can also be any information that discloses information about the person's racial or ethnic origin, religious, philosophical or other beliefs, political affiliations, as well as information about a person's health, such as his/her health history.
- **Non-personal information:** A set of non-personal information available in a system.

Identification information: Identification information, or *IdInfo*, defined as any linguistic expression that directly identifies a referent(s) of the type individual. Examples of identification information are an identity card number, a credit card number, and a passport number. If *S(x)* is a sentence that contains identification information, where *x* is refers to this information, and there is no referent (R), the sentence is still considered as identification information although there is no referent stated. The reason for this categorization is that *IdInfo* is directly identified to a person. The following is an example:

The credit card number is 1234 5678 9012

The above sentence is considered as *IdInfo* although we have explained that every piece of personal information must have at least one referent. However, in this case, there is no referent within the information. Because we know that a credit card number is unique and that one credit card number belongs to one person only, this number can be used to directly identify a person. This structure implies that, if we know the credit card number, we will automatically know to whom it belongs.

Sensitive information: Sensitive information, or *SnInfo*, is defined as any linguistic expression that is sensitive to a referent(s). Sensitive information can be any data that discloses information about a racial or ethnic origin, religion, philosophical or other belief, political or personal opinion, or membership of parties as well as personal data that discloses one's health, such as a health history. If *S(x)* is a sentence that contains sensitive information, where *x* refers to a person, at least one referent (R) should exist so that the information is considered to be sensitive information. Sensitive information can also be any personal information that cannot be used to directly identify a person but is considered as private information, and some people may need to keep this information confidential. The following is an example:

Ahmad's annual income is SAR 50,000.00

The above sentence is considered to be sensitive information. Sensitivity is quite difficult to explain because different people may have a different level of sensitivity on the same subject. Some people do not mind releasing or disclosing their personal information, such as health or financial, but other people do not want this information disclosed. In many situations, sensitivity depends on the context.

Personal information flow model: The Personal Information Flow Model (PIFM) was introduced by Al-Fedaghi (2006); it consists of four main modules, or phases: creating, collecting, processing and disclosing personal information. Thus, PIFM provides a systematic method for understanding related notions and explains a wide variety of cases by illustrating the relationship between different actors on personal information. The PIFM consists of four main phases, which include informational privacy entities and processes. These four phases are:

- Creating personal information
- Collecting personal information
- Processing personal information
- Disclosing personal information

This model complements other descriptions, such as the data protection EU directive as an explicit representation of personal information flow in reality. This EU directive lumps together all of the processing of personal data, including collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, and erasure or destruction.

Dorsey (2000) introduced different types of categories that can be applied to personal information: retrieving information, evaluating/assessing information, organizing information, analyzing information, presenting information, securing information, and collaborating with information. In the context of personal information privacy, this categorization can be applied to several phases, such as creating, collecting, processing, controlling and disclosing the personal information.

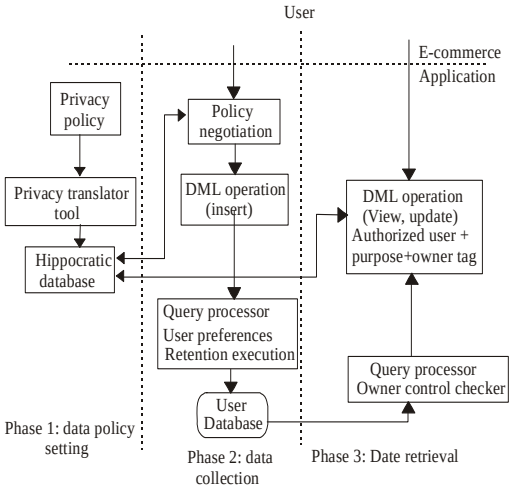

Fig. 2: The enhancement of PIFM introduced by Al-Fedaghi (2006)

However, to protect the personal information, there is a need to control the personal information disclosure. For this reason, we have decided to add one more phase between the processing and the disclosing of personal information, to achieve this objective.

In a web-based environment, personal information is disclosed by the data subject and is used by organizations. The organization will collect, store, and manipulate information to fulfill their organization's needs. Because of the importance of data privacy and the subjects' involvements towards their own personal information, we add another phase called Controlling the personal information before is closing the personal information phase, as shown in Fig. 2. This model, which was introduced by Norjihan and Zailani (2008), stated that any personal information should be disclosed to only authorized users who have a specific purpose and only for a limited time.

Creating personal information: Creating personal information is the first phase of the PIFM. Personal information can be created by two parties: a proprietor or a non-proprietor (e.g., medical diagnostic procedures performed by physicians) or can be deduced (e.g., data mining that generates new information from existing information) (Al-Fedaghi, 2006). Figure 1 show that personal information can be created at the points that are labeled 1, 2 and 6. Any atomic personal information of an individual is proprietary personal information of its proprietor. Once the personal information has been created, it can be either used (point 5) or collected (point 4), or it can go to a controlling phase before disclosing it (point 3). "Uses" means that the personal information is being used in a decision making process. Point 3 stated that the personal information should be controlled before it is disclosed, which means that the personal information will only be disclosed if it passes the fourth phase.

Collecting personal information: After the personal information is created, it can be collected at point 4. Personal information is collected from various sources and for various purposes of collection. The collected personal information can be either kept as records for future use (point 8), used (point 10), processed (point 9) or moved to a controlling phase (point 7).

Processing personal information: The processing phase of personal information involves storing (point 11), using (point 12) and mining (point 14) the personal information. Personal information is processed based on the purpose for which it was collected. In addition, personal information can be controlled (point 13).

Controlling personal information: The previous model introduced by Al-Fedaghi (2006) is modeled without a "controlling personal information phase". In this paper, we extend this work by adding this phase. In this new era of the Internet, it is important to control personal information before disclosing it. This phase will check the personal information before it enters the last phase of disclosing it. Figure 1 shows that all of the personal information is controlled at points 3, 7 and 13 before deciding whether the personal information can be disclosed.

Disclosing personal information: Disclosing personal information means that the personal information will be released to insiders or outsiders. Personal information is only disclosed if there is authorization to do so.

OWNER-CONTROLLED ARCHITECTURE FOR HIPPOCRATIC DATABASES

The previous section discusses the emerging trend of Hippocratic Databases as a new dimension of databases that incorporates privacy protection needs. In this research, we apply the concept of Hippocratic Databases to enforce privacy protection towards personal information stored in a database. We also introduce a technique on how an information owner should be able to control their own personal information while it is maintained in a database. Owner-controlled means that the information owner should be able to access and update (where necessary) their own personal information regardless of the purpose. This consideration is important because Hippocratic Databases introduce the concept of "purpose" when users access their information.

Confidentiality involves the sharing of information. "Confidential" information generally refers to any personal information that is kept in confidence such that

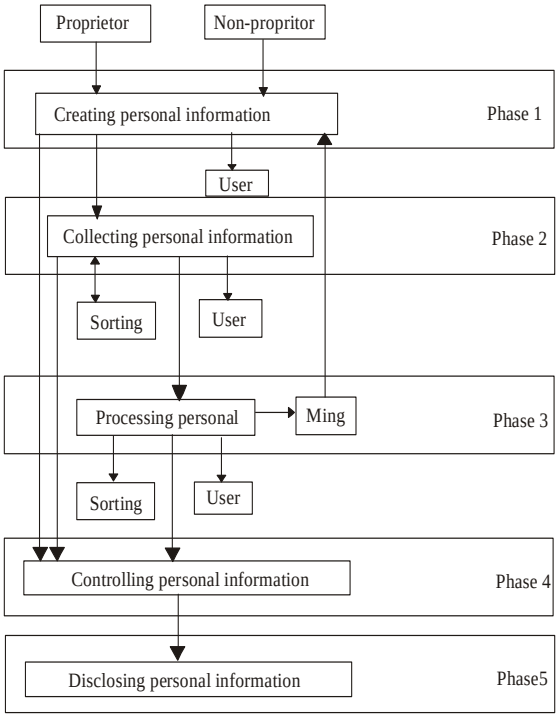

Fig. 3: Owner-controlled architecture in Hippocratic database

its revelation requires the consent of its owner. Confidentiality implies the protection of other people's secret information through the control of access to information and its release, according to certain agreements between the organization and the owner. A credit card number, identity card number and telephone number should be considered to be confidential information. Afyouni (2006) addresses confidentiality through two important aspects of information security:

- The prevention from unauthorized access to private information
- The process of safeguarding confidential information and controlling the disclosure of private information only to authorized users by means of classifying the information

Personal information should be kept only by the owner him/herself, and he/she should be able to control the disclosure of it to ensure its privacy. However, in web-based applications, this information should be disclosed to fulfill the transaction requirements. Although the personal information is disclosed, it cannot typically be accessed by unauthorized users for security and privacy reasons. Therefore, there are three main issues that must be considered:

- personal information should not be accessed by unauthorized users
- only required personal information will be disclosed
- personal information cannot be provided to those who do not need the information

The purpose of our proposed architecture is then to assert ownership control towards his/her own personal information while it is stored in the database. We have already discussed that the PI should only be disclosed to any people who are authorized to view it. In addition, we also claim the following:

- The owner of the PI should have control of his/her own personal information
- The owner of the PI should be able to know who, where, how and to what extent his/her PI will be used
- The owner of the PI should have a direct involvement towards his/her own PI in the following situations:
  - No matter what the purpose of the accessibility
  - Either it is known or unknown to authorized users
  - Either within the retention period or not, and
- Only the owner of the PI can control his/her PI

Figure 3 shows the Owner-Controlled Architecture (Norjihan and Zailani, 2009), which demonstrates a way to enable the owner to control his or her own personal information. This architecture was inspired by the Strawman Architecture, which was introduced by Agrawal *et al.* (2002a). Two databases are required in this architecture, and we separated them between the Hippocratic database and the user database. Because the Hippocratic database uses a meta-process mechanism, it is better to create a separate database to store privacy and privilege levels. In this case, we refer to it as a Federated database rather than an integrated database. Basically, at the application level, the processes of sending, checking and verifying the privacy using "purpose" and "authorized-user parameter" happens transparently. In this architecture, there are three phases:

Phase 1: Privacy Policy Setting
Phase 2: Data Collection
Phase 3: Data Retrieval

Phase (1) privacy policy setting: The organization first designs a privacy policy and then uses the Privacy Translator Tool to translate the privacy policy and to generate the privacy metadata tables. In general, a policy can be specified using any privacy policy specification language. Mapping the privacy policy into its database equivalent results in two tables, called the Privacy-policy Table and the Privacy-authorization table. This database acts as a centralized reference for privacy and privilege levels. The mapping from the privacy policy to the privacy-policies table makes use of automated tools called P3P IBM Tool. Before creating the privacy-authorization table, we must know who should have access to what

data. These two tables will be kept as privacy metadata tables in a Hippocratic Database.

Phase (2) data collection: The Data Collection phase is a phase where personal data are collected and stored in a database. Before the user provides any information, we must check whether the privacy policy is acceptable to the user. During the policy negotiation, users will perform a negotiation with the system, and users can only provide information if they agree with the policy. Assuming that the policy negotiation is successful, the DML Operation (Insert) is sent through the system to update the user data. The Query Processor will examine the user preferences, and review the retention before the data are stored in a database. All of the updated information will be maintained in the User Database.

Phase (3) data retrieval: Phase 3 begins when a user submits a query through an application. For an easier explanation, we have divided the users into two types, an owner and a non-owner. An owner is a user who owns the information inside the database; non-owners are all of the other users except for the owner. This categorization is important because the retrieval process is different between the owner and the non-owner. As usual, in the Hippocratic Database, queries are submitted to the database along with their intended purpose, which is either to view or to update the data. Before executing the queries, the system will check through the privacy metadata table to know the purpose, authorized users and owner tag of the query. The Query Processor will execute the query and check the Owner Control Checker to identify whether the query is from an owner or not. For the owner, it should be tagged together with the owner tag because not all of the information can be accessed by owners. For example, in a healthcare system, the "treatment" purpose is to be set up by a doctor. However, a patient should be able to access any information about the entries, including what types of treatment that he/she obtains and the name of the doctor who provides the treatment. In other words, regardless of the purpose, an owner should be able to access all of his/her information.

CONCLUSION

The world is currently shifting from off-line systems to on-line systems. Increasing amounts of personal information are collected, stored, manipulated and disclosed. Thus, there is an increasing need to protect personal information from incidents or mishaps. The Hippocratic database system is a good start in protecting personal information and is built at the database level as well as at the application level. In our solution, the owner of the PI, of course, has control of their own personal information. The architecture proposed here can help organizations to gain more trust from the owner of the personal information.

REFERENCES

Afyouni, H.A., 2006. Database Security and Auditing: Protecting Data Integrity and Accessibility. Thomson Course Technology, Massachusetts, Boston.

Agrawal, R., J. Kiernan and R. Srikant, 2002a. Hippocratic Database. Paper presented at the 28th International Conference on Very Large Data Bases, Hong Kong, China.

Al-Fedaghi, S.S., 2006. Personal Information Flow Model for P3P. Paper Presented at the W3C Workshop on Language for Privacy Policy Negotiation and Semantics Driven Enforcement, Ispra, Italy.

Al-Fedaghi, S.S. and B. Thalheim, 2008. Databases of Personal Identifiable Information. Proceedings of the 2008 IEEE International Conference on Signal Image Technology and Internet Based Systems.

Al-Fedaghi, S.S. and B. Thalheim, 2009. Personal information databases. Inter. J. Comp. Sci. Inf. Security, 5(1): 11-20

Bertino, E., J.W. Byun and A. Kamra, 2007. Database Security. In: Petkovic, M. and J. Willem, (Eds.), Security, Privacy and Trust in Modern Data Management, 18: 472.

Dorsey, P., 2000. What is PKM? Retrieved from: http://www.milikin.edu/webmaster/seminar/pkm.html, (Accessed on: July 20, 2008).

Guarda, P. and N. Zannone, 2009. Towards the development of privacy-aware systems. Inf. Software Technol., 51(2): 337-350.

Metzler, M. and P. Harker, 2008. Personally Identifiable Information (PII): A White Paper on Information Security, Version 1.6.

Norjihan, A.G. and M.S. Zailani, 2008. Controlling your personal information disclosure. Proceedings of the 7th WSEAS International Conference on Information Security and Privacy, pp: 23-27.

Norjihan, A.G. and M.S. Zailani, 2009. Owner-Controlled Towards Personal Information Stored in Hippocratic Database. International Conference on Computer Technology and Development, pp: 227-231.

Westin, A., 1967. Privacy and Freedom. Atheneum, New York.