A secure architecture for the pseudonymization of medical data

Bernhard Riedl, Thomas Neubauer, Gernot Goluch
Secure Business Austria
riedl, neubauer, goluch@securityresearch.ac.at

Oswald Boehm, Gert Reinauer, Alexander Krumboeck
Braincon Technologies, Vienna, Austria
o.boehm, g.reinauer, a.krumboeck@bct.co.at

Abstract

As aging and very expensive programs put more pressure on health and social care systems, an increase in the need for electronic healthcare records can be observed, because they promise massive savings and better clinical quality. However, patients and commissioners for data protection have legitimate concerns about the privacy and confidentiality of the stored data. Although the concept of pseudonymization allows an association with a patient only under specified and controlled circumstances, existing approaches have major vulnerabilities. This paper provides a new architecture for the pseudonymization of medical data that combines primary and secondary use in one system and thus provides a solution to vulnerabilities of existing approaches.

1 Introduction

The health care sector is at a crossroads. Aging and very expensive programs put more and more pressure on health and social care systems. As the health care sector accounts for 9% of Gross domestic product (GDP) in Europe and the share is growing at 6% p.a. [4], most governments are challenged by the balancing act of sustaining a high quality health care system by simultaneously reducing costs. Due to the cost pressure on the health care system [14] an increase in the need for electronic healthcare records (EHR) [9] could be observed in the last decade, because the EHR promises massive savings by digitizing diagnostic tests and images. EHRs could improve communication, access to data and documentation and thus, promises to lead to better clinical and service quality. A study by the nonprofit research organization Rand Corp. found that adopting the EHR could result in more than $81 billion in annual savings in the US, if 90% of the healthcare providers used it [4]. In addition, the EHR would provide the succeeding advantages:

- The expenses of access to manually created and delivered traditional paper-based records is considerable and causes unacceptable delays in processing patients. The EHR on the one hand minimizes the costs for the retrieval of medical data and on the other hand maximizes the availability of patient related data.
- The EHR allows the reduction of adverse drug events (ADR) accounting for about $175 billion a year in the US [3] and for the very high number of more than 200.000 cases of death a year in the US [4]. The EHR provides health care providers with decision support systems and guidelines for drug interactions.
- The EHR enables the collection of both patient care and clinical research data. This improves the efficiency of clinical trails and the medical treatment, because studies could be carried out faster and with a significant number of samples. In addition, disease specific registries could be established [1, 15, 13] that allow the monitoring of diseases and therefore provide experts with more appropriate data, e.g. for the development of new medication or treatment methods.

The health care sector is driven by the need to control costs and quality and thus forces the development and use of central repositories of health and personal information concerning patients. However, patients and commissioners for data protection have legitimate concerns about the privacy and confidentiality of the stored data. On the one hand it is the patient's right to demand privacy and on the other hand the disclosure of medical data can create serious problems for the patient. Insurance companies or employers could use this information to deny health coverage or employment. The disclosure of sensitive data, such as the history about

Second International Conference on Availability, Reliability and Security (ARES'07)
0-7695-2775-2/07 $20.00 © 2007 IEEE

substance abuse or HIV infection, could result in discrimination or harassment. The discussion of privacy is one of the fundamental issues in health care today and a trade-off between the patient's need for privacy as well as the society's needs to improve efficiency and reduce costs of the health care system. These concerns raise the need for data storage that guarantees data privacy and keeps the access to health data under strict control of the patient. Pseudonymity is an approach that provides a form of traceable anonymity and requires legal, organizational or technical procedures, consequently the association can only be accomplished under specified and controlled circumstances. A pseudonymous record or transaction is one that cannot – in the ordinary course of events – be associated with a particular individual [19]. Existing approaches for pseudonymization [13, 12, 20] have several drawbacks that pose a major threat to the privacy and confidentiality of stored patient data. Others have not yet been realized for the field of patient related privacy issues [2, 5, 6, 7].

In this paper, we propose a new architecture for the pseudonymization of medical data that provides the following contributions:

- Our architecture allows the authorization of other persons, such as health care providers or relatives, to access defined data of the EHR on encryption level. This is the most secure state-of-the-art authorization technology as alternative approaches, e.g. role-based access models [17, 23], may be compromised.
- We provide a backup mechanism that allows to recover the access to the health care records if the security token carrying the keys (e.g. a smartcard) is lost or stolen. Without using such a fall-back method, the link between the patient's identification and his medical data would be lost forever.
- To gain unobservability in data, our approach eliminates data profiling by using different pseudonyms for every anamnesis case and the possibility to only establish a link by knowing a certain key which is used to calculate a certain pseudonym.
- Secondary use without the option to establish a link between the patient and her data by anyone that was not the patient or authorized by the patient.

Moreover, compared to existing approaches, our concept does not depend on a patient list, which reflects the association between the patient's identification and medical data or a breakable algorithm, because both pose a weak point of any architecture. Instead, we based our architecture on a layered or hull structure. To get access to a certain key, which can be used to generate a pseudonym every user has to conduct different encryption and decryption operations. As long as the patient keeps her keys secure, it is not possible to associate a certain patient with her medical data.

2 Background

Todays health care systems have to face a trade-off between patient's privacy and the need for data access of other stakeholders. The "Healthcare Team", as Jamens Pope called all people providing medical services for a patient [14], needs access to the patient's data to do their job. This is called primary use of data. In this context, the electronic health record (EHR) that allows to store patient data in a centralized system, would be a cost-reducing and quality-improving factor for the health care sector. Furthermore, research institutions could get access to the data for developing new medical treatment or conducting long-time studies without compromising the patients' privacy. Their output, gained by the secondary use of medical data, would be a further cost-decreasing factor for the health care system. Another group of stakeholders represent social insurance companies that are interested in gaining unrestricted access to the patient's medical data. It is obvious that using the EHR could save time and therefore money by handling data more efficiently. Moreover, the quality of services could be improved by giving the option of comparing actual data with historical information.

However, as a centralized system holding sensible data is an interesting target for attackers, it is necessary to provide security for the information in the system and protect the privacy of the participants, even if the system has been compromised. Privacy can be gained in a two-step-process. First of all, it is necessary to identify all information that can be uniquely associated with a certain person. In the next step this identification data is separated from the remaining medical data. In other words, this means that every patient's data is divided into three groups: (a) the identification data, like the name, a social insurance number or the address, which we separate from (b) an entry in the medical database for every anamnesis and (c) a special case of medical data, the emergency data, like the blood-group or everyday medication. The next paragraphs present an overview of known approaches used for providing patients' privacy.

The first technique we want to mention is anonymization, which is the removal of the identifier from the data. In the medical case it means deleting the patient's identification data and leaving the anamnesis data [16, 11, 21] for secondary use. As this approach does not establish a link between the anonymized data and its associated individual, it is the most secure way for granting privacy [16, 11, 21]. Although this approach is often used in research projects due to its simplicity, it has the major drawback that patients cannot be informed about actual findings of a certain study (such as new developed medical treatment or major changes

in the healing progress). A technique similar to anonymization is called depersonalization, which means the removal of as much information as needed to conceal the identity of a patient [16].

Another approach for granting privacy issues is called pseudonymization. In Greek a pseudonym is a false name, used for example by authors which do not want to share their identity. In the field of information engineering it means a technique where identification data is transformed into a and afterwards replaced by a specifier which can not be associated with the identification data without knowing a certain secret. Algorithms for calculating the pseudonym can be based on encrypting or hashing techniques [8]. If the latter is applied, the only way of assuring reversibility is to store a list where all pseudonyms are kept [13, 12, 2, 5]. The usage of a list is a weak point in the architecture of existing systems for pseudonymization, because if an attacker gains access to this list, he is able to establish a link between the identification data and the medical data of a specific patient. Encryption provides a more secure alternative for building pseudonyms. For using encryption with a symmetric algorithm, a secret key, for the asymmetric alternative a key-pair (secret or private key/public key), is needed. Either way it is necessary to assure secure storage of the secret key. As the applied keys may be stored on security tokens, such as smartcards, which can be lost, stolen, destroyed or compromised, there is a need for backup mechanisms that allow restoring the key, e.g. by sharing the secret key with other persons or systems (cf. [12, 13, 20, 2, 5]. Otherwise, if the patient is not in possession of his key anymore, the medical data is lost forever, because the pseudonym cannot be reversed anymore.

Existing approaches for the pseudonymization of medical data can be differentiated by a) the way the pseudonym is created and shared, b) the security techniques that are used and c) by the owner of the secret. Pommerening et al. [13, 12] propose an architecture for the pseudonymization by using a two-staged process. First of all, the patient's identification data is transformed by an algorithm, called PID service, into a unique patient identifier. Secondly, the pseudonym is generated by encrypting this ID with a symmetric key. If there is the need to unveil the pseudonym, a) the system decrypts the pseudonym and uses the gained input to b) calculate the identification data with the usage of the PID service. Another option is to store the output of the PID service and the original identification data in a patient list. Both only slightly different approaches offer the possibility to an attacker to gain illegal access to the algorithm or the patient list by compromising the system, e.g. by a social engineering attack [22, 10] on persons with an administrative role in the system. Statistics show that half of the attacks on a system are conducted or supported by insiders. Hence, on the one hand an attacker could break into the system by taking possession of the patient list or by gaining illegal access to the PID service. On the other hand an attacker could bribe an insider to sell the secret of the patient list or enable illegal access to the PID service.

A similar architecture was proposed in a patent by Thielscher et al. [20]. Their system is set up by security tokens on smartcards. The options to calculate the pseudonyms are a) with a patient's smartcard alone or b) with the combination of a patient's and a health care provider's smartcard. In the case of lost or stolen smartcards, these security tokens can be replaced by the usage of a patient list, which is stored offline. In other words, their concept is to operate a pseudonymization server without a network connection. This approach assures better security, but an attacker could still gain illegal physical access to the server or conduct a social engineering attack. Moreover, the pseudonymization server has to be updated regularly with new or changed data in order to issue new smartcards. The procedure of transporting data via other channels than a network, for example dvd-rom or a mobile hard disk is slow and more expensive, because it needs a person or a machine to conduct this work. From a security view, the approach of Thielscher et al. protects the privacy of a certain patient better compared to the approach of Pommerening et al., because an attacker has no possibility to break inside the patient list from outside, but the system is still compromisable. Within the upcoming section, we propose a new architecture for protecting the privacy of medical data, which solves the problems of the architectures mentioned above.

3 Pseudonymization Architecture

The goal of our architecture is to gain the optimal trade-off between maximum security on the one hand and usability and performance on the other hand. Although we do not rely on a patient list inside the system, we still provide a mechanism for recovering lost or destroyed keys. Moreover, basing the authorization on encryption techniques allows us to avoid profiling of data. Firstly we want to outline the roles and components in our architecture. We continue with a presentation of the design principles and security methods we applied.

Our architecture (cf. Figure 1) consists of the following roles and components:

- A central system (e.g. server, etc.) (St) which provides access to a central storage which itself is divided (e.g. logical, physical) into two separate storage systems (e.g. databases, etc.), where one is related to identification data (AID) and the other one is related to data, which should be pseudonymized (PMD) as well as the associated pseudonym PSN,
- a central logic (Lo) that provides an interface between

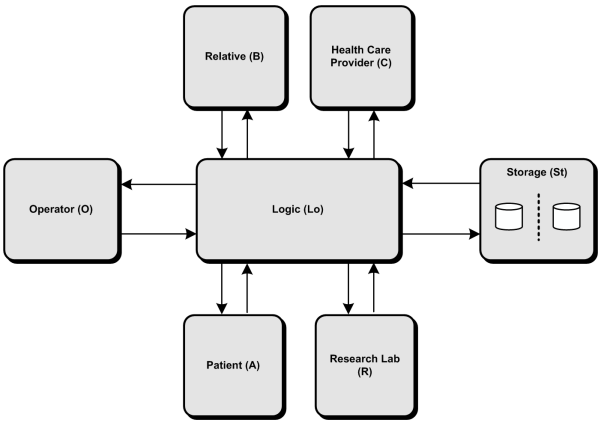

Figure 1. Architecture

the central storage and the clients for the purpose of saving and loading the data,

- the patient (A) who has full access to the AID and PMD on the central system via the central logic by using a security token (e.g. smartcard with a PIN, etc.),
- the relative (B) who shares the identification data with a specific patient and therefore has the same rights by default (can be changed by a role based access model),
- the health care provider (C) who shares one or more entries in the pseudonymized database with the patient,
- the research lab (R) that just has access to the PMD on the server system via the central logic for the purpose of analysis needed for improving the efficiency of clinical trails, the medical treatment, or medication and
- the operator(-team) (O) which may hold secrets on behalf of the system. In other words this role assures that if a patient loses or destroys his smartcard, the access to the system can be restored by a team of operators.

The architecture is based on a layered model with a minimum of three security-hulls representing the authorization mechanism (cf. Figure 2). Every hull includes one or more different symmetric keys or asymmetric key pairs. A key K_N of a certain hull H_N is encrypted with a key K_{N+1} of the hull H_{N+1} enveloping hull K_N. We identify each entry PMD in the database, which represents the concealed data hull, by the usage of a pseudonym PSN, which is calculated based on the most inner key, for example by using this key for encrypting several defined attributes of the patient's identification data. To establish a link for a certain patient and an entry in the concealed hull, a patient, or a system on behalf of the patient, conducts the suceeding operations.

The user starts by authenticating against her most outer hull key, for example by entering a pin to authenticate against a smartcard. The key of the authentication layer, which is, in our example, stored on a smartcard, can be used to decrypt the encrypted secret key K_A, in the next inner hull, the user permissions hull. After gaining access to K_A the patient selects an anamnesis and decrypts the key which is related to the selected anamnesis and is able to establish a link to the data by calculating the pseudonym again.

As it is necessary to authorize different stakeholders to access all or portions of the data, we propose the possibility of sharing pseudonyms. Exemplarily a patient is able to authorize a health care provider C to access a certain anamnesis by sharing the pseudonymization key. Moreover, we offer the option to grant a person, for example a relative, access to all of the data by sharing the secret key K_A of the user permissions hull.

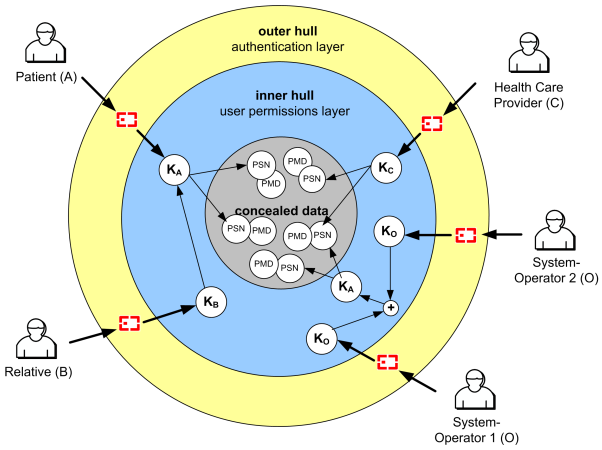

Figure 2. Layered model representing the authorization mechanism

In the next section we provide a detailed view of the functions and permissions of the different roles in our architecture and expose possible attacking scenarios.

3.1 Authorization on Encryption Level

As role-based access models can be compromised [17, 23] we implemented all major parts of the system (e.g. the authorization techniques) on encryption level. In other words authorization is given by sharing certain keys between users in the inner hull/user permissions layer and revoked by deleting or changing them. Every user possesses a security token (e.g. smart-card), where the outer key pair is located. After authenticating against the system, the Logic transfers the encrypted inner public key of a certain user to the user, who decrypts this key with her outer private key.

Afterwards it is possible to access every secret which has been encrypted before with the inner public key. Moreover, at this point of time there are several other keys in the system which can be accessed by the user. This assures that no attacker can commit for instance a buffer-overflow or an elevation of privileges attack to benefit of a security flaw [17, 23].

It is still possible to share given keys with other users without the notice of the user which gave the original permission to share a certain key. Hence, it is necessary to share complete keys, which also means full access to a secret with as little users as possible. Moreover, shared secrets should be divided between more persons which have to act together to unveil a certain secret, to minimize the risk of misuse. In our architecture, the inner key pair of a patient may be shared with other users to allow, e.g. a relative, a custodian or nursing stuff access to the patient's data. Furthermore, this concept allows to assure, in case of a lost or destroyed outer key pair, that the secrets inside are still restorable. If for example a patient loses her smartcard, there would be no possibility to gain access to her pseudonyms again and the data would not be accessible any more. If desired by the patient, her inner key pair can be shared with one or more users she defines which have full access to the secret.

We define a number of n operators with k operators owning a share to unveil the secret key of a certain user. The association between the operators is first encrypted with the logic key and afterwards encrypted with the operators inner public key. Hence, the operator, who was randomly assigned to a certain user, herself does not know the association to a specific user, unless she knows the logic key. Moreover, no operator knows who her counterparts are. On the other hand this information is hidden from the system, too, as long as the operators do not use their private keys to decrypt the association.

To assure maximum security, we base our technique to share keys with the operator-team on the threshold scheme by Shamir to share secrets securely [18]. In a scenario with three operators holding shares for a secret and a minimum of two have to act together to access the secret, an attacker, maybe a compromised operator who wants to get access to users data, needs to convince in worst case $n - 1$ other operators and take possession of the logic key, to commit a successful attack. This method is not 100% secure, though it can be considered safe in our opinion. As an enhancement on the organizational level, there is still the possibility to use different offices for the operators to prevent arrangements between the operators.

As pseudonyms are generated by encrypting the identification data of a certain patient, the usage of the same keys for encryption of identification data leads to the building of the same pseudonyms. Hence, if pseudonyms are used too often, the risk of profiling arises. Within the upcoming section, we deal with the issue and explain the possible options to introduce countermeasures.

3.2 Unobservability

One vital point of our system is that the data itself does not have to be encrypted. We assure privacy by securing the link between the patient's identification data and her anamnesis data with the encryption of her identification data with a pseudonymization key. It is possible to share all pseudonymization keys, which are located in the most inner hull, as well. For example to authorize a health care provider, a certain user may decrypt one of her pseudonymization keys and encrypt it with the inner public key of the mentioned health care provider. A pseudonymization key can be used for n times or a time period t, which could be the interval of the diagnose until the healing of a disease. If the usage already extends n or t has elapsed, the system provides a new key to every participant. As the key can be passed on without notice of the data owner (the patient) n can also be set to 1 and in that case the system issues a key separately for every instance and every dataset.

The latter approach eliminates data profiling completely, because it is not possible to establish a connection between more pseudonyms and therefore avoids the possibility of guessing the identity of a certain user by combining more data sets together, because there is no concealed association. Hence, if health care providers store only anamnesis data which cannot be related to a certain patient, in other words any attribute which could be associated with a patient would be stored with the identification data and not with the anamnesis data, unobservability is guaranteed, because every anamnesis dataset has one pseudonym for one user or more different pseudonyms for every user that has access to that certain dataset. It is only possible to build these pseudonyms by knowing the pseudonymization key and in order to know the keys one has to be authorized. This approach uses more storage space and cpu time, but as in most security-related architectures this represents the trade-off between security and usability regarding the performance. If pseudonymization keys are not shared between different users, there is also a secure option to revoke authorization, even without the notice of the user who shares a secret. This can be done by deleting the pseudonym in the specific anamnesis dataset, because the encryption of the patient's identification data with the revoked pseudonymization key leads to a pseudonym which is not associated with a dataset any more. Hence, the patient is in full control of the data.

4 Conclusion

The 'Healthcare Team' needs access to different portions of information of a patient. It is our goal to introduce a new type of an EHR architecture to combine primary and secondary use of medical data with the most efficient degree of security. Our solution provides an approach to reuse medical data in research institutions with the option to inform the patient about results, follow-up studies or evolved medical treatment. Moreover, we assure that even if an attacker is in possession of the whole database, it is not possible to establish a link between a specific patient's identification and her medical data without the usage of the patient's or another authorized person's security token.

Moreover, our solution enables that patient to authorize other persons, for instance a health care provider for a second opinion or relatives. The proposed system grants that the patient remains in full control of her data as she can revoke a given authorization without the other person's attendance. Hence, she is able to decide to which information each individual or organization has access. Furthermore, we introduced the concept of a hull structure that allows to exchange keys in different hulls independently from one another. As a result, keys on security tokens or inside the system can be replaced easily, e.g. if state-of-the-art evolves and new techniques should be adopted. Another contribution of our approach is that we include a secure backup mechanism. This fall-back mechanism can be set-up by administrative persons inside the system or defined persons outside the system, for example a relative. These persons share the patient's secret to assure the redundancy of possible lost or destroyed security tokens and therefore avoid the risk of lost data. If operators hold the secret on behalf of the user, we apply the four-eyes-principle and segregation of duties in our approach.

Within the next months, we plan to publish the details concerning our architecture, as presented in our pending patent, including case studies and a prototype. To detail the authorization technique, we will propose an additional role-based access model which will reflect the tailored needs of the specific stakeholders, which could not be represented by the authentication on encryption approach. This will decrease the possibility of frauds which may occur, if shared secrets are distributed to other person without the notice of the patient.

Acknowledgment

This work was performed at Secure Business Austria, a competence center that is funded by the Austrian Federal Ministry of Economics and Labor (BMWA) as well as by the provincial government of Vienna.

References

[1] The Austrian Cancer Register.

[2] J. Biskup and U. Flegel. Transaction-based pseudonyms in audit data for privacy respecting intrusion detection. In *RAID '00: Proceedings of the Third International Workshop on Recent Advances in Intrusion Detection*, pages 28–48, London, UK, 2000. Springer-Verlag.

[3] F. R. Ernst and A. J. Grizzle. Drug-related morbidity and mortality: Updating the cost-of-illness model. Technical report, 1995.

[4] F. R. Ernst and A. J. Grizzle. Drug-related morbidity and mortality: Updating the cost-of-illness model. Technical report, 2001.

[5] U. Flegel. Pseudonymizing unix log files. In *InfraSec '02: Proceedings of the International Conference on Infrastructure Security*, pages 162–179, London, UK, 2002. Springer-Verlag.

[6] O. Jorns, S. Bessler, and R. Pailer. An efficient mechanism to ensure location privacy in telecom service applications. In *Net-Con 2004, Mallorca, Spain*, 2004.

[7] O. Jorns, O. Jung, J. Gross, and S. Bessler. A privacy enhancement mechanism for location based service architectures using transaction pseudonyms. In *2nd International Conference on Trust, Privacy, and Security in Digital Business, Copenhagen, Denmark*, 2005.

[8] A. Lysyanskaya, R. L. Rivest, A. Sahai, and S. Wolf. Pseudonym systems. In *Proceedings of the Sixth Annual Workshop on Selected Areas in Cryptography (SAC '99)*.

[9] S. Maerkle, K. Koechy, R. Tschirley, and H. U. Lemke. The PREPaRe system – Patient Oriented Access to the Personal Electronic Medical Record. In *CARS 2001 Computer Assisted Radiology and Surgery, H.U. Lemke, et al. (Eds.), Excerpta Medica International Congress Series, Elsevier, Amsterdam, Netherlands*, pages 849–854, 2001.

[10] K. Maris. The Human Factor, 2005.

[11] A. Pfitzmann and M. Koehntopp. Anonymity, Unlinkability, Unobservability, Pseudonymity, and Identity Management A Consolidated Proposal for Terminology. In *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2005.

[12] K. Pommerening. Medical Requirements for Data Protection. In *IFIP Congress, Vol. 2*, pages 533–540, 1994.

[13] K. Pommerening and M. Reng. Secondary use of the Electronic Health Record via pseudonymisation. *Medical Care Compunetics 1*, -:441–446, 2004.

[14] J. Pope. Implementing EHRs requires a shift in thinking. PHRs–the building blocks of EHRs–may be the quickest path to the fulfillment of disease management. *Health Management Technology*, 27(6):24,26,120, 2006.

[15] J. Powell and I. Buchan. Electronic Health Records Should Support Clinical Research. *Journal of Medical Internet Research*, 7:e4, 2005.

[16] A. Rector, J. Rogers, A. Taweel, D. Ingram, D. Kalra, J. Milan, P. Singleton, R. Gaizauskas, M. Hepple, D. Scott, and R. Power. Clef - joining up healthcare with clinical and post-genomic research, 2003.

[17] R. Russell, D. Kaminsky, R. F. Puppy, J. Grand, D. Ahmad, H. Flynn, I. Dubrawsky, S. W. Manzuik, and R. Permeh. *Hack Proofing Your Network (Second Edition)*. Syngress Publishing, 2002.

[18] A. Shamir. How to share a secret. *Commun. ACM*, 22(11):612–613, 1979.

[19] K. Taipale. Technology, Security and Privacy: The Fear of Frankenstein, the Mythology of Privacy and the Lessons of King Ludd. *International Journal of Communications Law & Policy*, 9:–, 2004.

[20] C. Thielscher, M. Gottfried, S. Umbreit, F. Boegner, J. Haack, and N. Schroeders. Patent: Data processing system for patent data, 2005.

[21] D. Thomson, L. Bzdel, K. Golden-Biddle, T. Reay, and C. A. Estabrooks. Central Questions of Anonymization: A Case Study of Secondary Use of Qualitative Data. *Forum Qualitative Social Research*, 6:29, 2005.

[22] T. Thornburgh. Social engineering: the "Dark Art". In *InfoSecCD '04: Proceedings of the 1st annual conference on Information security curriculum development*, pages 133–135, New York, NY, USA, 2004. ACM Press.

[23] T. Westran, M. Mack, and R. Enbody. The last line of defense: a host-based, real-time, kernel-level intrusion detection system. In *submitted to IEEE Symposium on Security and Privacy*, 2003.