A Systematic Approach to Anonymity

Sabah S. Al-Fedaghi

Computer Engineering Department, Kuwait University, Kuwait

Abstract. Personal information anonymity concerns anonymizing information that identifies individuals, in contrast to anonymizing activities such as downloading copyrighted items on the Internet. It may refer to encrypting personal data, generalization and suppression as in k-anonymization, 'untraceability' or 'unidentifiability' of identity in the network, etc. A common notion is hiding the "identities" of persons to whom the data refers to. We introduce a systematic framework of personal information anonymization by utilizing a new definition of private information based on referents to persons in linguistic assertions. Anonymization is classified with respect to its content, its proprietor (the person it refers to) or its possessor. A general methodology is introduced to anonymize private information, based on canonical forms that include a personal identity. The methodology is applied both to textual and tabular data.

1 Introduction

It is claimed that online anonymous communication is a strong human and constitutional right [19]. As in real life, people who work in cyberspace have legitimate reasons to employ anonymity to avoid the consequences of identity exposure. Anonymity has an important social function, as seen in such social phenomena as whistleblowing and hotlines (drug abuse). Anonymity also contributes to the general goal of controlling the use of private information. There are many motivations for interest in anonymous personal (private) information. Producing anonymous medical information is a policy objective in the USA [9] and the EU [22]. It is a very important research area aimed at providing the sharing and distribution of medical records while maintaining patient confidentiality.

From the technical point of view, the anonymization of private information adds one more level of security. There are situations that make known methods of cryptography undesirable, especially with the "legal limits" of technological protection of secrecy of communication. Anonymization technology that uses cryptographic methods to transform identifying information can be "de-anonymized" where encrypted records can be matched using such system as ANNA [10]. Consequently, developing new methodologies or refining existing ones to address hiding the nature of information, especially, in the privacy arena, is an important research objective.

Anonymization of personal data is of special significance in the area of health information systems. According to the U.S. Health Insurance Portability and

S. Al-Fedaghi S. (2005).
A Systematic Approach to Anonymity.
In *Proceedings of the 3rd International Workshop on Security in Information Systems*, pages 160-172
DOI: 10.5220/0002569401600172

Accountability Act of 1996, “anonymized data” refers to “[p]reviously identifiable data that have been deidentified and for which a code or other link no longer exists” [9]. Under the HIPAA Privacy Rule, one aspect of “deidentification” is that the health data does not include eighteen identifiers of persons which could be used alone or in combination with other information to identify the subject. These identifications include: names, telephone numbers, fax numbers, email addresses, social security numbers, URLs, etc. Also, data that “are separated from personal identifiers through use of a code” are termed as “coded data” and “[a]s long as a link exists, data are considered indirectly identifiable and not anonymous or anonymized” [9]. In the EU Data Protection Directive [6] [4], “anonymisation of personal data” is understood as erasing “person-identification” or converting identifiable data into non-identifiable data. Along the same line, the Germany data protection law specifies that “use [of personal data] shall be made of the possibilities of anonymisation and pseudonymisation where possible...” [7].

In general “anonymity” can be defined as the condition in which others do not know a person’s true identity. The term “anonymous” may be defined as the “condition of having a name that is unknown or concealed” [12]. We can see that there are several conceptualizations of the notion of ‘anonymous personal information’. It may refer to encrypting personal data, “generalization and suppression” of certain parts in the personal data, ‘untraceability’ or ‘unidentifiability of personal identity’ in the network, etc.

The notion of “effectiveness” may also influence what we mean by ‘anonymization’. For example, according to Walden [22], data is considered not ‘identifiable’ if the identification requires unreasonable amounts of efforts (EU Recommendation). “Achieving effective anonymisation may be a challenging task, from both a technical and compliance perspective” [22]. Sometimes data is considered neither personal nor completely anonymous (The Austrian data protection legislation).

In this paper, we propose a new framework for identifying and classifying private information anonymization. It includes setting it apart from other types of anonymization, identifying its categories and outlining a general methodology for applying it to different forms of information. The next section describes some current research in this area. In section 3, we review a newly proposed definition of private information. This definition forms the foundation of our contribution in this paper. In section 4, we analyze the relationship between the notion of anonymity and private information. Accordingly, we propose a classification of private information anonymization based on the content, proprietor, and possessor of private information. In section 5, we concentrate on a certain type of anonymization that is typically discussed in literature. In section 6, we propose a methodology of anonymizing textual private information and apply it to relational database tables. Finally, conclusions are drawn and directions for future work are discussed in Section 7.

2 Related Works and our Contribution

Anonymization permits data to be usefully shared or searched without revealing the individual’s identity. In the medical field, there is a great deal of interest in

anonymizing textual information. Sweeney's pioneering work [16] is based on removing the personal identifying information from the text so that the integrity of the information remains intact, even though the identity remains confidential. It includes developing an algorithm and software program called 'Scrub Extractor' that automatically extracts names, addresses, and other identifying information from the free text documents. Sweeney's recognition methodology aims at detecting information that can personally identify any person. One important issue that can be observed here, is related to the definition of "personal information." Is it the whole text, the paragraph, the sentence, the phrase or only the word that denotes the identity? We will answer this question in the next section. Sweeney also introduced the *DataFly* system that provides an additional level of anonymity [17]. Ruch et al. used syntactic and semantic knowledge to classify the tokens within a text [13]. N-gram type rules, finite state automata and a recursive transition network were used to encode the knowledge and extract patient identifiers. *Taira* et al. presented a methodology that manually tags all references to patient identifiers and context information [18]. The scheme searches for logical relations that are characterized by a predicate and an ordered list of one or more arguments. In most cases, the logical relation consists of three arguments; a head, a relation, and a value. In *Johnny underwent a pyeloplasty for uretropelvic junction stenosis…*the token *Johnny* is the logical relation head, *underwent* is the relation, and *pyeloplasty* is the value. In *Johnny is a 5 year old Caucasian male with Disease X*, the token (*5 year old* and *Caucasian*) modifies *male*, that syntactically modifies its head *Johnny* [18]. The identification detection problem is concerned with certain types of logical relations. All combinations of words in a sentence that can fill the roles (*i.e.,* head, relation, and value) of a given logical relation are considered. Other authors in the area of medical textual information worked on morpho-syntactic aspects of the term formation in medical language. For example, works in this area lead to the development of an encoding system for diagnoses and interventions based on a semi-automatic encoder with natural language entry and an interface [5].

Another important area of research in this direction is the notion of 'k-anonymity' [15]. The k-anonymization of a relational table, assumes that a table with a prime key that refers to a person is the personal information. Its main concern is anonymizing entries in the table in order to block any attempt to reach "identifiablity" that stems from these entries. Systems that use such techniques aim at protecting individual identifiable information and simultaneously maintaining the entity relationship in the original data. Still, the definition in these works of "personal information" is not clear. Implicitly, it is understood that the privacy aspect comes from associating the attribute name with the identifying key of the relation.

In spite of impressive efforts and results in this area, we claim that the topic of "private information anonymization" has not been systematized. Systematization here means systematically concentrating on the 'quality' of privacy in the general scheme of anonymization of information. It starts with the definition of 'private information'. Additionally, anonymization methods are usually focused on eliminating identities. This brute mechanism hides fine points of anonymizing private information of a person or private relations among persons. *John and Mary are in love* can be anonymized with respect to John (*Someone and Mary are in love*), with respect to Mary (*John and Someone are in love*) or with respect to the relationship between them (*John and Mary are in some type of relation*). Our proposed systematic

approach moves from a definition of private information to discriminating between types of anonymizing private information. While immediate benefits in terms of specific algorithms and technicalities are not introduced, the methodology provides a formal foundation to the topic. The next section is a brief review of a recent definition of private information that satisfies this requirement [1].

3 Private Information

Defining what is private information is a problematic issue. Privacy is usually said to be culturally defined notion. Wacks defines it as "those facts, communications or opinions which relate to the individual and which it would be reasonable to expect him to regard as intimate or confidential and therefore to want to withhold or at least to restrict their circulation" [21]. Several types of privacy have been distinguished in literature including 'physical privacy' and 'informational privacy' [8]. Recent results have shown 'private information' in true linguistic assertions about an identifiable individual. An ontological definition of private information can be developed from linguistic assertions in order to identify the basic units of private information.

Our basic ontological entities (things we talk about, subjects of predication) are individuals and non-individuals. We preserve the term 'individual' to denote a particular human being. Let Z denotes the set of ontological entities such that $Z = V \cup N$, where V and N are the sets of 'individuals' and 'non-individuals' respectively. We have three types of linguistic assertions:

(a) Non-individual assertions or 'assertions with zero private information'. That is, q is a zero (privacy) assertion if the set of ontological entities referred to by q is a subset of N.

(b) Individual (private) assertions, which, include two types:

Atomic Private Assertions: p is an atomic private assertion if p contains a single referent of type V.

Compound Private Assertions: p is a compound assertion if p contains more than one referent of type V.

The assertion *Spare part ax123 is in store 5*, is a zero assertion because it does not involve any individual (human). *Farmer John's house is burning* is an atomic assertion because it embeds a reference to a single identified individual. *Maria's preparing the document pleased John* is a compound private assertion because it embeds identities of two individuals. If an assertion is true, then it is said to be information, otherwise it is said to be misinformation. Consequently, there are zero information, atomic information, and compound information according to the number of referents.

We identify the relationship between individuals and their own atomic private information through the notion of *proprietorship*. Proprietorship of private information is different from the concepts of possession, ownership, and copyrighting. Any atomic private information of an individual is proprietary private information of its *proprietor*. A proprietor of private information may or may not be its possessor and vice versa. Atomic private information of an individual can be embedded in compound private information: a combination of pieces of atomic private information of several individuals. Two or more individuals may have the same piece of

compound private information because it embeds atomic private information from these individuals. But it is not possible that they have identical atomic private information, simply because they have different identities. Atomic private information is the "source" of privacy. Compound private information is "private" because it embeds atomic private information. Also, the concept of proprietorship is applied to compound private information, which represents "sharing of proprietorship" but not necessarily shared possession or 'knowing'. Some or all proprietors of compound private information may not "know" it.

Compound private information is privacy-reducible to a set of atomic assertions, but it is more than that. For example, *Maria's preparing the document pleased John* can be reduced to *Maria's preparing the document pleased someone* and *Someone's preparing the document pleased John*. However, compound private assertion is a "bind" that contains not only atomic assertions but also asserts something about its atomic assertions. Privacy-reducibility of compound information to atomic information means that "no known atomic information" of an individual implies "no known compound information" of that individual. Because, if the compound information is known, then its atomic assertions are known. Reducing a compound assertion to a set of atomic assertions refers to isolating the privacy aspects of the compound assertion. This means that, if we remove the atomic assertion concerning a certain individual from the compound assertion then the remaining part will not be a privacy-related assertion with respect to the individual involved.

Suppose we have the compound private information, *John saw Mary's uncle, Jim*. The privacy-reducibility process produces the following three atomic private assertions:

Assertion-1: *John saw someone's uncle.*
Assertion-2: *Mary has an uncle.*
Assertion-3: *Jim is an uncle of someone.*

Additionally, we can introduce the zero-information meta-assertion: *Assertion-1, Assertion-2, and Assertion-3 are assertions of one compound private assertion*, from which it is possible to reconstruct the original compound assertion. The methodology of syntactical construction is not of central concern here. In database modelling there are three (private information) databases of John, Mary and Jim, with one (non-private information) database that includes "pointers" that link the three private facts [2].

In releasing medical data for statistical analysis, reconstructing the original compound private information is not required. However, in certain applications, the reconstruction process is important. Compound private information is not a collection of atomic private information; and it is not "putting-together" connections. *V1 and V2 are in love* does not have this 'collectivity' meaning as in *V1 and V2 are London*. The latter, is pseudo compound private information. It is a collection of the atomic private information: *V1 is in London* and *V2 is in London*. *V1 and V2 are in London* is simply a simplified method of writing *V1 is in London* and *V2 is in London*.

We have defined every piece of information that includes an identifiable person as private information. Nevertheless, such information can have different levels of sensitivity. "Sensitivity" in the context of private information refers to a special category of private topics that may disturb people. This definition of sensitive private information is related to the typical definition where sensitivity of information refers to the impact of disclosing information. Consider the case of Public Access to Court

Electronic Records, where the public is able to download and print court case files deemed to be "sensitive-but-not-confidential" by the courts in the Court Electronic Records (PACER) discussed in [11]. They include such information as "social security numbers, credit card numbers or medical information; they also can unearth personal filings such as divorce or bankruptcy cases." Privacy-rights advocates recommended that the system "electronically remove such personal information within public court filings that would be available online." Since our work in this paper concerns the mechanism of anonymization of private information, we will ignore the issue of what type of private information the anonymization is applied to.

An individual can have (process, possess, etc.) his/her own (proprietary) atomic private information or other's (non-proprietary) private information. A non-individual (company, government agency, hospital, etc.) can have only non-proprietary private information. We divide the atomic private information space of an individual into the following categories.

NProprietary Information: This type of information is the set of pieces of atomic private information of the others that are in possession of the individual or non-individual. If this private information is in the possession of an individual then he/she is not its proprietor.

Proprietary Information: This type of information is the set of atomic private information of the proprietor. It has two subsets:

Known: This is the set of atomic private information that is known by others (in possession of others).

Not Known (NKnown): This is the set of atomic private information that is only known by the proprietor and no one else.

The next section introduces our new contribution in this paper. We specify the notion of "private information anonymity" in terms of the definition of the private information given above.

4 Classification of Private Information Anonymity

Let the private information T be denoted as the triple: (NProprietary T, Proprietors, Possessors) where *Proprietors* is the set of proprietors in T and '*NProprietary T*' is a version of T produced from the original information following the concealment (e.g., removal, replacement, etc.) of the identifiers. In the communication context, 'Possessors' can be the sender and recipient of the message. In the relational database schema, a processor can be the view owner. For example, the possessor of a piece of information in EMPLOYEE (NAME, SALARY) is the finance department while it proprietor is the specific employee. The relationship between anonymity and the private T information can now be categorized in table-1.

Table 1. Categorization of different types of anonymization

	NProprietary T	Proprietors	Possessors
a	Unanonymous	Anonymous	Unanonymous
b	Unanonymous	Anonymous	Anonymous
c	Unanonymous	Unanonymous	Anonymous
d	Anonymous	Anonymous	Unanonymous
e	Anonymous	Anonymous	Anonymous
f	Anonymous	Unanonymous	Unanonymous
g	Anonymous	Unanonymous	Anonymous

'Unanonymous' NProprietary T means not hiding the version of T produced from the original information following the concealment of the identifiers. The case (a) represents the typical anonymity case where the possessor of private information (e.g., hospital) anonymizes the medical data before releasing it. In (b) the proprietor and possessor (e.g., source) are anonymized as in case of gossip, e.g., *According to an anonymous Hollywood source: A big movie star is an alcoholic*. In (c) the possessor (e.g., source) is anonymized as in the case of a "secret source" posting some private information on the network.

Anonymous NProprietary T means hiding the version of T produced from the original information following the concealment (e.g., removal, replacement, etc.) of the identifiers. This hiding of data may involve, for example, cryptographic methods used in anonymous data-matching technology. "To take a simple example, one-way hashing permits two owners of lists to encrypt their lists, compare them, and identify all of the items that are on both lists – without either one learning anything else about the contents of the other's list" [3]. Thus in (d) the private information is hidden, however, an external observer may know its proprietors. This is typical in network communication where the identities of the sender and the receiver are known but the content of the message that includes private information is not known. In (e) even the possessor (e.g., source) of the anonymous private information is not known. In (f) the external observer knows the possessors and proprietors but does not know the NProprietary T. For example, a person sends his/her CV to a company. The external observer knows that it is private information about a certain person (proprietor) and knows the sender and receiver but does not know the content of the private information (e.g., the proprietor's age). Also, this type of anonymity is reflected in such expressions as *Bob and Alice are talking about me, I wish I knew what they are*

saying. In (g) the external observer knows the possessor (e.g., sender) but does not know the content and the proprietor.

We distinguish "private information anonymity" from communication anonymity where the issue is hiding the sender and recipient identities. 'Private information anonymity' refers to anonymizing the content and not the act of communicating. The communicated information in 'communication anonymity' is not necessarily private information. If it is private information, then the two notions may overlap each other.

For example, *I love you* is anonymized private information that refers to its proprietors by the labels "I" and "you" (case (a) or (b)). However if we know that the sender is Bob and the receiver is Alice then the message is no longer anonymous because Possessors $\subseteq$ Proprietors. If the message is *He loves her* and Possessors $\cap$ Proprietors = $\varnothing$ such as an external observer knows that Bob's message is directed to Alice (case (a)) then the private information is anonymous even though the privacy of the communicating act is not. The privacy of the possessors (the communicating parties) is different from the privacy of the message. The sender and recipient can be non-individuals. If they are individuals they can be non-proprietors. If they are proprietors then "private information anonymity" and communication anonymity become identical topic.

Interestingly, the proprietors of private information can be its possessors. We can view the anonymization (a) to (g) under this condition as follows:

(a) This situation can be described as:

T$\rightarrow$(NProprietary T, anonymous proprietor, possessor)

The symbol $\rightarrow$ denotes transforming the original data T into the triple on the left. In this case, the proprietor/possessor anonymizes only his/her identity as the proprietor of the information as in the situation of a person releasing anonymous private information and hides the fact that it is about him (e.g., a politician announces starting an investigation of a scandal in his/her campaign but does not mention the fact that the scandal involves him/her).

(b) This situation can be described as:

T$\rightarrow$(NProprietary T, anonymous proprietor, anonymous possessor)

In this case, the proprietor/possessor anonymizes his/her identity as the proprietor and source of the information as in the situation of a whistleblower releasing private information that involves him/her without identifying him/herself.

(c) This situation can be described as:

T$\rightarrow$(NProprietary T, proprietor, anonymous possessor)

In this case, the proprietor/possessor anonymizes only the fact that he/she possesses the information as in the situation of a person releasing his/her private information and he/she hides this fact (e.g., celebrities secretly releasing their own information to the tabloids). In the Internet, this type of anonymization involves anonymizing of the Internet Protocol (IP) address as the source of a message that contains the person's identity.

A more interesting situation is, not only when the proprietor is the same as the possessor, but also when the private information is in his/her NKnown. That is, a person who anonymizes his/her private information, which no one knows, but him/herself. The whole set NKnown is a set of pieces of this type of anonymous private information. There are very elaborate techniques that try to achieve this type of anonymization. For example, using a blind signature to create anonymous e-

money, thus providing a cash-like payment mechanism has the property that the user of the cash can remain anonymous; i.e., not exposing part of his/her informational space, NKnown.

5 Private Information anonymity

The previous section categorized all types of anonymization related to private information. In the rest of this paper, we concentrate on private information anonymity of type (a), which is of special importance in the area of health information systems. The aim of anonymization here is to provide the sharing and distribution of private information while maintaining individual confidentiality. So in a straight de-identification of a patient's record, the possessor is unanonymous, the record is unanonymous, but the proprietor is anonymous. The U.S. Health Insurance Portability and Accountability Act of 1996 is mainly applied to anonymization of previously identifiable data in possessions of others. Accordingly, we are interested in private information protection that involves anonymity and is achieved through severing the association of the content of the information from its proprietor.

Accordingly, we define private information anonymization in the sense of type (a) where the issue of anonymizing the possessor is ignored, as follows:

Definition: Private information T is said to be anonymized if it is transformed into non-proprietary information '*NProprietary T*'. The anonymization of private information of T involves the transformation: T→(NProprietary T, Proprietors), where *Proprietors* is the set of proprietors in T and '*NProprietary T*' is the anonymous version of T.

This type of anonymization of private information is different from any other notions of protecting the privacy of personal activities. For example, in the news it is reported, *Rural/Metro, an ambulance and fire service company in Scottsdale, Arizona, sued four individuals who ... The defendants were four individuals, known as John/Jane Does 1-4...* [20]. Using the labels "John/Jane Does 1-4" instead of the identities of the involved individuals is what we call "anonymizing private information". On the other hand, network privacy technologies that utilize anonymization to prevent abuse is not, in general, private information anonymization. A protocol that allows anonymous communication between two entities protects the privacy of "communication" but does not necessarily protect "private information." Hence, if a user uses anonymous services to protect his/her "privacy" in such activity as downloading non-private information (e.g., copyrighted music), then such a measure is not in the domain of private information anonymity.

6 Methodology of Anonymization

We develop a general methodology to anonymize private information through identifying atomic assertions. We define a canonical form (I, A) for any atomic private assertion where I is the identity of the proprietor and A is *NProprietary* of the

assertion, the zero-privacy version of the atomic assertion. The method is applied first to textual data and then to relational databases.

Suppose that the private text under consideration is T. An anonymization algorithm can be specified as follows:

1. Identify (T_1, T_2,... , T_n) in their order of occurrences, where T_i is either an atomic or compound assertion.
2. If T_i is an atomic assertion then it is represented by its canonical form (I_i, A'_i), where I_i is an identifier of the proprietor of T_i and A'_i is a zero-privacy assertion version of T_i.
3. If T_i is a compound assertion then let (C_1, C_2,... , C_k) be its set of corresponding atomic assertions, in their order of occurrences. For each C_i replace it with its canonical form (I_i, C'_i), where I_i is an identifier of the proprietor of C_i. and C'_i is a zero-privacy assertion version of C_i.
4. Replace T by T′ which is a sequence of canonical forms such that each atomic assertion is replaced by one form (I_i, A'_i) and each compound assertion is replaced by the sequence of forms ((I_1, C'_1),... , (I_k, C'_k)).
5. Factor out the identifiers in T′, thus producing two lists I and Z where: I is the list of identifications in T′ after the deletion of assertions and Z the list of assertions in T after the deletion of the identifiers.

Example: Consider the text: "*Mary is in London. John saw Mary's uncle, Jim, in Paris.*" The canonical form of the atomic assertion *Mary is in London* is (Mary, *Someone is in London*). The privacy-reducibility process produces the following three atomic private assertions from the compound assertion *John saw Mary's uncle Jim in Paris*:

(1) *John saw someone's uncle in Paris*. Its canonical form is (John, *Proprietor saw someone's uncle in Paris*).

(2) *Mary has an uncle*. Its canonical form is (Mary, *Proprietor has an uncle*).

(3) *Jim is an uncle of someone*. Its canonical form is (Jim, *Proprietor is an uncle of someone*).

Thus, T′ is: [(Mary, Proprietor is in London), [(John, Proprietor saw someone's uncle in Paris), (Mary, Proprietor has an uncle), (Jim, Proprietor is an uncle of someone)]].

Factoring out the identifiers produces the following lists:

N = ([Mary, [John, Mary, Jim], Mary, [Mary, Jim])

Z= [(*Proprietor is in London*), [(*Proprietor saw someone's uncle in Paris*), (*Proprietor has an uncle*), (*Proprietor is an uncle of someone*)]]

This can be rewritten as: *X is in London, Y saw X's uncle, Z, in Paris.* Reconstructing the original data from the anonymized data is not necessary in some applications (e.g., releasing medical data for statistical analysis). As we can see, one difficulty of such a process is to maintain the connections between atomic assertions that facilitate constructing combined assertions. Compound assertions such as *Jack and Jill love John* are pseudo-compound assertions; in contrast to a 'real' tri-proprietors compound assertion such as *Jack, John, and Jill love each other*. The former assertion needs to be formulated as the compound assertions *Jack loves John* and *Jill loves John.* These compound assertions can now be anonymized as *Jack loves someone* and *Jill loves someone.* A meta-statement that ties these two assertions

together reconstructs the original assertion if that “someone” is the same person in the two assertions. The relational database model makes such a task easier.

A relational database of private information can be envisioned as set of zero, atomic or compound assertions. A private database methodology has been studied in [2]. Suppose that a relational database HOSPITAL includes the following relations:

PATIENT (NAME, APPOINTMENT-NUMBER, AGE, ...)
DOCTOR (NAME, APPOINTMENT-NUMBER, PHONE, ...)
APPOINTMENT (APPOINTMENT-NUMBER, TIME, ROOM. ...)
A-P-D (APPOINTMENT-NUMBER, DOCTOR-NAME, PATIENT-NAME)
TEST (NUMBER, DESCRIPTION)
T-P (LABTEST-NUMBER, PATIENT-NAME)

Each relation in this schema represents a set of assertions. Relations that include private information are: PATIENT, DOCTOR, T-P and A-P-D. PATIENT stands for the atomic private assertions: *The name of the patient is X*, *The patient has appointment Y*, *The age of the patient is M*. The relation A-P-D stands for the compound private assertion *The appointment number of Dr. X with patient Y is Z.* TEST stands for the zero-privacy assertion *The description of lab test number X is Z.* Notice that we use *X*, *Y*, *M*, and *Z* to denote arbitrary values in the database.

Hence, we can produce a textual version of the relational schema and apply the same methodology of de-identification described previously. This is interesting theoretically; but there is no practical reason that motivates such a process. In general, transforming data between the textual and tabular forms based on atomic/compound assertions may have some application in the database design field. Notice that in contrast to the textual data, the order of the assertions is immaterial.

Any Atomic assertion has two components: referent-part and Zero-part. For example in *He is shy*, ‘*He*’ is the referent-part and the predicate *is-shy* is the Zero-part. The k-anonymization method goes one more step by anonymizing the zero-part of the assertion. So if the AGE of a person is suppressed the resultant atomic assertion is *Someone’s age is some number*. In this type of suppression there is a complete loss of information. This concept creates further categorization of types of anonymization in table-1. In some applications, it is required to anonymize the zero-part of the assertion instead of the referent part. This type of anonymization can be observed in the published news of allegations against a person without disclosing their contents.

7 Conclusion

This paper introduces a systematic approach to define anonymization in the context of private information. The concept of “private information anonymization” is distinguished from other related notions. It is also classified according to its content, its proprietor and its possessor. A general algorithm is introduced to recognize and anonymize private information. The method is based on canonical forms of linguistic assertions that include a personal identity. The algorithm is applied both for textual data and tabulated data as in the case of relational databases.

For textual information our methodology introduces the basic principle of the approach. Of course a great deal of improvements and refinements can be introduced on the basic methodology such as developing a more sophisticated mechanism for the "straight" list I. Sweeney's detection and replacement machinery can be applied at different levels of this methodology. A great deal of research is needed at the linguistic level to develop a "Private Information Analyzer". Many statements and words in natural language do not contain private information. Reading the text to identify private information is a tedious process since it may be scattered through out the text with a lot of privacy-unrelated text in between. A Private Information Analyzer will assist people in locating and analyzing private information in documents. It will perform various tasks such as finding all the occurrences of private information in a text, ranking pieces of private information according to their sensitivity, suggesting possible replacement to reduce the level of sensitivity etc. It may be used alone or it may be connected to a knowledge-based system so that privacy found in the text can be embedded in the system.

It is still difficult to provide a fully automated understanding of unrestricted natural language, because of its involved theoretical complexities. We are currently, exploring a modified subset version of controlled English called ClearTalk to facilitate our analysis of private assertions. Skuce has provided semi-automated tools for analyzing unrestricted language dealing with knowledge extraction and document-based knowledge [14]. ClearTalk is generated from English text by utilizing these tools and a human editor. The main objective here is to create logically equivalent structures that are intelligible to both people and computers.

For the relational database model, our method is being developed as part of the Enhanced Privacy Information System [2].

References

1. Al-Fedaghi, S.: The 'Right to Let Alone' and Private Information, Proceedings of the 7th *International Conference on Enterprise Information Systems*, Miami, USA (2005).
2. Al-Fedaghi S., Fiedler G., and and Thalheim B.: Privacy Enhanced Information Systems, Proceedings of *The 15th European-Japanese Conference on Information Modelling And Knowledge Bases:* Tallinn, Estonia (2005).
3. Baker, Stewart: Testimony Before the National Commission on Terrorist Attacks Upon the United States, December 8 (2003). http://www.911commission.gov/hearings/hearing6/witness_baker.pdf
4. Baker, S., Kees Kuilwijk, Winnie Chang and Daniel Mah: ANONYMIZATION, DATA-MATCHING AND PRIVACY: A CASE STUDY, (2003) http://www.911commission.gov/hearings/hearing6/witness_baker.pdf .
5. Baud RH, Lovis C, Ruch P, Rassinoux AM.: A Toolset for Medical Text Processing. Medical Informatics Europe, Hannover 2000, MIE 2000, IOS Press (2000).
6. EU Directive 1995/46/EC: On the protection of individuals with regard to the processing of personal data and on the free movement of such data, *Official Journal of the European Communities*, No. L 281, 23.11.(1995).
7. Federal Data Protection 1990-2001, Act of 20 December 1990 (Federal Gazette I, p. 2954, 2955), amended 18 May 2001; at § 3a.

8. Floridi di Luciano: Information Ethics: On the Philosophical Foundation of Computer Ethics, ETHICOMP98 *The Fourth International Conference on Ethical Issues of Information Technology*, (1998), http://www.wolfson.ox.ac.uk/~floridi/ie.htm.
9. HIPAA: Glossary of Common Terms, Health Insurance Portability and Accountability Act of 1996, http://healthcare.partners.org/phsirb/hipaaglos.htm#g3.
10. Levy, Steven: Geek War on Terror, Newsweek, (2004), http://www.msnbc.msn.com/id/4486823/.
11. Olsen, Stefanie: Privacy advocates question Net access to court docs, *ZD Net*, (2001), http://zdnet.com.com/2100-11-527611.html?legacy=zdnn.
12. RFC 2828: Internet Security Glossary, Internet RFC/STD/FYI/BCP Archives, May (2000). http://www.faqs.org/rfcs/rfc2828.html
13. Ruch, P., R. H. Baud, A-M Rassinoux, P. Bouillon, and G. Rober: Medical document anonymization with a semantic lexicon. In Proc. of the AMIA Fall Symposium, (2000) 729-733.
14. Skuce, D.: Integrating web-based documents, shared knowledge bases, and information retrieval for user help, *Computational Intelligence* 16:1 (2000).
15. Samarati, P.: Protecting Respondents' Identities in Microdata Release, IEEE Transactions on Knowledge and Data Engineering, November/December, 13:6 (2001).
16. Sweeney, Latanya: Replacing Personally-Identifying Information in Medical Records, the Scrub System. In: Cimino, JJ, ed. Proceedings, *Journal of the American Medical Informatics Assoc.* Washington, DC: Hanley & Belfus (1996) 333-337.
17. Sweeney, Latanya: Guaranteeing Anonymity when Sharing Medical Data, the DataFly System, In Proc. of the AMIA Fall Symposium, (1997) 51-55.
18. Taira, Ricky K., Alex A. T. Bui, and Hooshang Kangarloo: Identification of Patient Name References within Medical Documents Using Semantic Selectional Restrictions, Proceedings of the AMIA 2002 Annual Symposium, (2002).
19. Teich, A.; Mark S. Frankel, Rob Kling, and Ya-ching Lee: ANONYMOUS COMMUNICATION POLICIES FOR THE INTERNET: RESULTS AND RECOMMENDATIONS OF THE AAAS CONFERENCE, *The Information Society* 15:2 (1999), http://www.indiana.edu/~tisj/readers/full-text/15-2%20teich.pdf .
20. Terraciano, Jeffrey: Can John Doe Stay Anonymous?, *Wired News*, Feb. 21 (2001), http://www.wired.com/news/politics/0,1283,41714,00.html.
21. Wacks, R.: 'The Poverty of "Privacy"', The Law Quarterly Review (1996) 73–95.
22. Walden, Ian: ANONYMISING PERSONAL DATA, *Int J Law Info Tech* 2002 10: (2002) 224-237. http://research.imshealth.com/IMS%20HEALTH_valueofdata_files/Anonymision%20and%20European%20Data%20Protection%20Directive.pdf .