Comput Syst Sci & Eng (2005) 5: 377–386

International Journal of
Computer Systems
Science & Engineering

Integrity and dissemination control in administrative applications through information designators

Wouter Teepe

Department of Artificial Intelligence, University of Groningen, Groningen, The Netherlands. Email: wouter@teepe.com

When more and more information sources are being linked, it seems that it becomes ever more easy to track individuals in ways that are not deemed appropriate. However, increased linking of information does not need to imply increased dissemination of privacy-sensitive information. We present a new approach to linking information sources that allows the owners of the information to maintain a high level of control over the information they maintain. The key consideration is that it is in general not required to replicate information across multiple information systems. In fact, imprecise replication of information actually endangers the integrity of linked information systems. What is needed is an architecture that enables information systems to refer to undisclosed information in a secure and transparent way. In our approach, we introduce the *information designator* which is a generalisation of the pseudonym concept.

Keywords: Unlinkability, pseudonymity, anonymity, privacy, information integration, information designators

1. INTRODUCTION

Nowadays in the information society, a lot of information available from different sources is combined to deduce new information. This everyday practice reduces the amount of privacy and anonymity that individuals have. At the same time, the combining of information does not always seamlessly succeed, for example local changes at an information source may easily lead to unlinkability problems. Solving problems surrounding information integration properly is already so difficult [27, 13, 20], that it is no real surprise that issues such as privacy and anonymity are often no substantial part of the initial integration design, if they are included at all.

Information integration is done when a group of organisations decide to pool their information. Typically this is a tedious task in which unrelated, individual (relational) databases have to be combined in such a way that the databases jointly act as one. A query on the aggregate database must be seamlessly divided into subqueries which operate on the individual databases, and the results of these queries have to be merged into one query result.

To actually integrate the databases, the schemata of the databases are compared, and fields in different databases but with similar semantics are identified. For example, one database may relate students to courses, and another database may relate people to favourite sports: the student and people fields are then used to relate courses to favourite sports.

When tying databases together in this way, two problems frequently occur. First, it is difficult to make sure that all matches that should be found between different individual databases are actually established. This is typically due to different ways of encoding the same information in different databases. Second, where the individual databases may be internally consistent, the joint databases may very well be inconsistent.

The common denominator in addressing these problems is to expose more information. Making more information avail-

able allows for more matches to be found, and allows for inconsistencies to be detected. Thus it seems necessary to expose a lot of information in order to achieve proper information integration. From the privacy and anonymity perspective, a priori exposing a lot of information is out of the question. This suggests that information integration on the one hand, and confidentiality on the other hand, are not on comfortable terms.

We believe however, that both information dissemination control and proper information integration can actually be achieved by one and the same instrument. In this paper, we will present our solution. This solution is by no means a 'one size fits all' solution, nor is it easy to implement given the legacy of information systems. On the other hand, our solution is in the end rather elegant and effective, and we would like to present it as a proof of concept.

In the next section (section 2), we will further elaborate on information integration, and how its problems relate to ontologies and dissemination of information. Section 3 will present our new approach to these challenges, and the central concept of this approach: the information designator. Phenomena mentioned in section 3 will be illustrated in section 4, where we show an example of how both information integration and dissemination control are solved jointly. In section 5 we discuss the relevance of our approach and relate it to other research. And of course, we end with some conclusions.

2. AN ANALYSIS OF THE PROBLEMS OF INFORMATION INTEGRATION

In this section, we present our analysis of the fundamental challenges that must be faced when integrating information. These problems stem from the fact that some information may be modelled multiple times, but differently (section 2.1), and from the fact that information, once disseminated from its original source, is hard to control (section 2.2).

At this point, it is good to make some remarks on what we mean by 'information integration'. In the abstract sense, information integration is the act or process of making sure that information stored and maintained at separate locations and organisations, can be combined with ease.

Roughly, there are two ways to accomplish this goal. The first way is to take a number of information sources, and perform the difficult and tedious task of matching of the information at the different locations. This includes among others record matching, data re-identification, record linkage, and is what is traditionally understood when one refers to information integration [20, 13]. However, there is another, second way of achieving the goal of assuring the easy combination of dislocated information, which will be our approach. The main idea is to anticipate the combining of information at the moment the individual information sources are set up. In section 3, we will show how this can be done without assuming a trusted central authority and without disclosing information which may need to remain confidential. We consider such an approach an important step towards solving the problems of information integration, though it is somewhat nonstandard, if compared to the traditional meaning of information integration.

2.1 Overlapping ontologies

An ontology defines, for a single information source, what the information stored in the source represents, and how it is structured. Within the relational database paradigm, a database schema can be seen as the implementation of such an ontology. When information sources are combined, this is done by comparing the ontologies of the different sources. If the ontologies overlap sufficiently, or if it is possible to map parts of one ontology onto some parts of the other ontology, the information from the two sources can be linked.

The individual information sources are almost always stand-alone information systems by origin. Because of this origin, these systems store many kinds of information, since they have (had) to maximally support the owning organisation. For example, a university database typically stores a lot of details about students, like students' previous educations, birth dates, private addresses. This information is stored because at some moment in time the university will need it for some task.

As a result the information sources subject to information integration tend to have a rather large ontology. It can even be argued that information integration happens because the ontologies grow so large that it is no longer viable for one single organisation to maintain all information within a stand-alone information system. Keeping track of how all information should be modelled, as well as actually obtaining all the information for a single, large stand-alone information system becomes very complicated when information from sources outside of the organisation have to be included.

It can be expected that in the example of the university database, inconsistencies will exist in the information that comes from outside of the organisation. Minor inconsistencies may arise from data-entry, bigger inconsistencies may arise form updating the information infrequently or not at all. Intricate inconsistencies may occur when the ontology does not have enough expressive power to facilitate the information that should be stored. On the other hand, it should be expected that the information in the university database concerning the university activities, such as course enrolments, grades and diplomas given, is essentially, if not by definition, correct.

An organisation which creates new information is probably the best suited organisation to model this information, and to maintain an ontology of this information. However, it is not unusual for such an organisation to maintain an ontology covering more than the *core business* of the organisation itself, but also to maintain a part of its ontology which is error-prone, and essentially a duplicate of many parts of many other ontologies of other organisations.

If the overlapping parts of the information sources' ontologies contain personal information, this means that this personal information is stored at several sites. If for whatever reason this information should be kept under some restricted disclosure regime, *all* sites storing this information should adhere to the restricted disclosure regime. Obviously, it may not be able to enforce this, which means that the information is kept private just insofar the weakest link does not disclose it. Information stored at only one site is easier to control, since there is only one party which has to adhere to a specific disclosure regime.

2.2 Information propagation

The reason for linking information sources, i.e. to perform information integration, is from the perspective of a participating organisation twofold. First, the organisation wants to *retrieve* authoritative information from external sources. When retrieving data, the desiderata are *availability* and *integrity* of the information. Second, the organisation wants to *publish* information, but possibly only to a restricted set of *consumers* for some restricted set of *application uses*. When publishing data, enforcing dissemination policies is the main challenge.[1]

To maximise integrity of information, best would be to verify the information at the authoritative source, as shortly as possible before actually using the information. Better could even be to just *fetch* the authoritative information at use-time. To prevent unwanted dissemination of information, best would be to verify that for each time the information is used, there is a legitimate reason to use this information. This can be achieved by requiring authorisation for each individual 'shipment' of information, and to make sure the information can only be used for the purpose stated in the authorisation procedure.

This leads to a central attitude in our approach: *don't propagate, but link*. Information should only be disclosed when it is really about to be used, and not at any time before that. At the very best, the disclosed information should be destroyed immediately after use.

This attitude may seem very unrealistic in two ways. First, it has to be properly defined what 'using information' actually means. If it is too widely defined, it does not really restrict dissemination. If it is too strictly defined, it prevents any sensible use of information. Information Designators, introduced and explained in the next section, will solve this apparent paradox. Second, one may question whether not propagating information would lead to unacceptable performance bottlenecks in the resulting information system. Assuring proper information granularity will minimise, if not circumvent this problem. Information Designators are the instrument that will offer us the flexibility to reason about information that is not *physically present*. This will dramatically lower the required capacity of information sources to disseminate information.

3. A JOINT APPROACH TO PRIVACY, ANONYMITY AND INFORMATION INTEGRATION

In this section, we will present our approach to solving information integration and dissemination control. First, we introduce the 'information designator' in section 3.1. Section 3.2 explains how, using information designators, information from various sources can be tied together, while these sources remain in control over their information. Moreover, in section 3.3 we explain how an organisation that provides information designators to others, can accurately manipulate which others can actually use the provided information designators, and to what extent.

3.1 Information designators

The central instrument in our approach is the *information designator*, which is a piece of information whose sole purpose it is to refer to other information without containing the other information, without any reference to a context. Let us define this new concept. Every designator contains an address at which a software agent, an *exchange agent*, can be contacted to translate the designator into the information it refers to. An exchange agent may place restrictions or conditions on the information requester before it translates a designator into the information it refers to.

An example of a designator could be *12345.67890*. If Bob were to ask Alice her home address, she could give Bob this designator. Bob then knows that if he wants to send postal mail to Alice's home, he must contact the exchange agent at *12345*[2] and hand over to the exchange agent the full designator *12345.67890*. In turn, if Bob meets the conditions set by the exchange agent, Bob will receive Alice's home address. The fact that the designator refers to Alice's home address, cannot be inferred from the designator itself. Bob only knows the designator has this semantics because Alice told Bob so. Alice should make sure that the exchange agent will answer Bob's call for information in the right way.

The process of Bob obtaining Alice's home address is now a two-step process. The *principal* step is the one in which Bob asks Alice her home address, and possibly after some combination of authorisation and agreeing on some terms, Alice hands over the information designator to Bob. From that moment on, until Bob contacts the exchange agent, the designator is something like an IOU of Alice to Bob, where the debt of Alice is the information that stands for her home address. Though Alice has granted Bob access to the information of her home address, she has still control over it. Alice can change her home address without any administrative burden to Bob. Also, Alice can *retract* her designator by instructing the exchange agent not to give Bob the information the designator refers (or: referred) to.

The second step is the *materialisation* step, in which Bob contacts the exchange agent. If Alice hasn't retracted the designator, and Bob meets the conditions set by the exchange agent, Bob will obtain the information that is Alice's home address.

The use of this kind of mapping allows for changing of the information referred to without the need to update references [29]. This would allow telecom operators to redistribute phone numbers, or the city council of Tel Aviv to rename the 'Malchei Yisrael Square' into the 'Yitzhak Rabin Square' without introducing inconsistencies into databases where these numbers or names are referred to.

This flexible use of designators has benefits for both the users of information and the providers of information. The users of information have access to the information they need, but they do not need to worry about the housekeeping of this information. Barring unforeseen exceptions, the users are guaranteed access to the information. At the same time, the providers of information are given greater control over the dissemination of the information, and can individually audit the use of the information.

1 This may all seem straightforward, but an important implication is that it is rarely if ever the case that an organisation would want to *directly* alter information that is within the realm of another organisation.

2 This could be a phone number, IP address, or something else that allows setting up acommunication channel in an automated way.

The architecture presented here could be considered to be a peer-to-peer data management system (PDMS), like the Piazza PDMS [22]. However, the PDMSs we know of lack the concept of an information designator, and do not distinguish between *raw* information, and a reference to such information. In fact, techniques used in Web Services [1] and PDMSs are generally a vehicle to ease the problem of schema integration, whereas the information designator is a means to *bypass* the problem of schema integration.

There are a number of solutions to the problem of providing services to retrieve data without providing direct access to data [1, 9, 28], but these are generally very tailor-made to fit just a few specific applications in isolation. The information designator, on the other hand, is not bound to any specific application or protocol. If desired, the same information designator can be re-used in multiple applications and protocols.

3.2 Dependency and (un)linkability

It may seem that by using information designators, the users of information are subject to possible arbitrary behaviour of the providers of information. For example, the providers might choose to instruct their exchange agents to further deny any information to the users. We do not believe this scenario is any more likely to happen than in a context where another mechanism for information integration is used. Even stronger, we believe the *possibility* to retract designators on an individual basis may well happen to be essential for many a participant in an information integration project. More organisations will be willing to provide information, because they have the option to retract the information in the case of an unlikely or unforeseen event.

Using information designators makes existing informational dependencies of organisations explicit. If an organisation depends for some task on information from another organisation, this will inevitably lead to an infrastructure in which designators are used whose corresponding exchange agents operate under the auspice of the organisation depended on.

The information designator approach has the very interesting property that if it is fully applied, there are no overlapping ontologies. Different organisations *provide* information under their own, simultaneously provided ontology. If this information is used, the provided ontology will be used. If this information is related to information from some other ontology, it will be related by means of a designator in the one ontology, pointing to information in the other ontology. Technically this means that instead of multiple information sources storing identical information, there is one information source that stores the original information, while other information sources store references (information designators) to this original information. In this sense, designators are the *glue* between ontologies, that allows ontologies to be disjoint, but integrated at the same time.

Disjointness of ontologies is an extremely useful feature from both the information integration and from the privacy and anonymity perspective. It effectively makes it impossible for conflicting information on one subject to be established, which seriously limits the class of possible inconsistencies that can arise from linking information. At the same time, information can be linked without automatically disclosing a part of the linked information: information normally made public can be kept private.

3.3 Operations on designators

One could wonder whether introducing designators actually improves privacy and anonymity, by reasoning that the designators themselves will fulfil the role of identifying information; that a person is not identified by his or her name, but by the designator that refers to his or her name. This would indeed be the case, if for each piece of information, there would only be one designator referring to it. If multiple parties would have this same designator, they could recognise that the information they individually have is about the same person or artifact.

However, it is *nowhere necessary* that each piece of information has only one designator pointing to it. In fact, the introduction of designators would have little to offer on the privacy and anonymity front if each piece of information would have its unique corresponding designator. An organisation handing out designators could in fact every time it hands out a designator, create an extra 'fully anonymous' designator for the information it needs to point to.[3] In this scenario, the organisation handing out designators knows for sure that the designators it handed out cannot be combined in any way to find matches between designators.

There are excitingly many policies between strictly unique designators on the one hand and fully anonymous designators. Here, we will mention just a few. Designators to the same piece of information could be the same, if given to the same requesting organisation, or if given to an organisation in some given group, thereby allowing the organisation or group of organisations to compare their designators. It is totally at the discretion of an organisation handing out designators to decide whether its designators will have these properties. Also, it could provide these properties to some users of information, and not to others. The closer the policy is to strictly unique designators, the more recombination possibilities there are that need no consent of the organisation that handed out the designators.

An organisation handing out designators does not have to fully decide on its policy when it starts handing out designators. For example, it could by default hand out only fully anonymous designators, and upon special request exchange some of the designators for designators that can be recombined in some specific way. A user, or group of users could for example ask the specific question if within a specific set of their designators, some refer to the same information. The organisation handing out designators could in turn translate the given specific set into other designators in such a way that only within this set duplicates can be detected.

Depending on policy decisions, the extent to which designators are valuable to users can be varied in a very precise

3 Creating an extra designator every time a designator is handed out will not have any serious impact on the required storage capacity of the exchange agent. This can be achieved for example by designators that actually are encrypted versions of a 'master designator', of which the exchange agent is the only agent knowing the decryption key. For more examples of designator obfuscation, see appendix A.

way. Organisations handing out designators can choose to make their designators on a per-user and per-transaction basis, homomorphic to the information the designators refer to.

4. AN EXAMPLE: THE DATA MINING BOOKSHOP

The information designator is more than a theoretical concept. In fact, we have built a prototype system which demonstrates several of the above-mentioned properties. The prototype illustrates an example of information integration and information exchange which would, without information designators, either be impossible or seriously infringe privacy. We present the prototype here for three purposes: (1) to stress that information designator systems *can actually be built* [23], (2) to show how an information designator system works internally, thereby illustrating the theory explained in the previous section, and (3) to give an application example which demonstrates how information designators help in protecting privacy and maintaining unlinkability.

4.1 Organisational setting

Our example is about information flow between four organisations. The first organisation is the civic *authority*, which has the task to maintain the municipal inhabitants register, which contains inhabitants' names, birth dates, and residence addresses. The second organisation is the local *school*, whose students live in the domain of the authority. The school keeps record of its students, their results, their course enrolments and required literature for courses. The third organisation is the local *bookshop*, which is located conveniently next to the school. The bookshop wants to provide for the literature demands from the school students, but does not want to overstock. The fourth organisation is the *publisher*, which publishes the books that are used in the courses of the school. The publisher maintains information about books and their details, such as titles, author(s) and ordering information.

There are relations between the information maintained by these organisations. The students of the school are all registered at the authority. Contrary to the publisher and the bookshop, the school has the right to access some of the information stored and maintained by the authority. The books the school recommends for their various courses, are all published by the publisher. The publisher is fairly liberal in allowing access to the information about its books, however, it has some extra information for its known resellers, one of which is the bookshop.

The bookshop has a strong desire not to overstock books, and at the same time the school wishes all their students to have their obligatory books when the term starts. As a result, the school depends on the behaviour of the bookshop, and the bookshop depends on information from the school. A very naive way to solve this dependency would be that the school gives the bookshop full access to the school administration. This would obviously lead to unacceptable privacy infringements, even if the school would limit the access to things like course enrolments (and hide exam results). A slightly less naive solution would be that the school gives the bookshop an update of the expected number of required books once in a while. However, these updates are just snapshots. It would be ideal for the bookshop to directly look in the administration of the school at the moments relevant for the bookshop. If this would not infringe the privacy of the students, the school would probably see no or only little problems in such a solution.

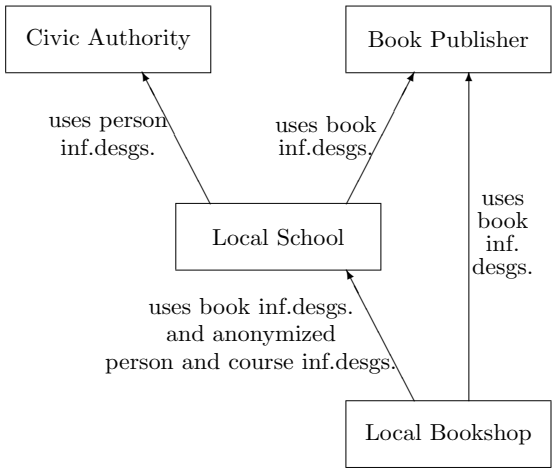

Figure 1 An information dependency graph containing the four organisations of the example. The organisations and their information demands are described in section 4.1. An arrow from organisation A leading to B means that A is interested in information maintained by B. For example, the bookshop depends on (desires) information from both the school and the publisher.

Figure 1 shows how the four organisations relate to one another with respect to their information needs.

The example may seem a perfect case for setting up a Web Service framework. However, whereas a Web Service framework would offer a means for exchanging information, the use of information designators offers a means for assuring mutual information integrity and consistency while keeping almost all information confidential. The confidentiality and integrity is not manually crafted into the architecture, it is a mere consequence of using information designator technology.

4.2 Designators in action

The information that is maintained by the organisations is summarised in Table 1. The table shows the schemata of the local databases. These could be plain vanilla relational databases, in which the 'person' field contains a string which denominates the person's name. This is however not the case. All fields contain *information designators*. Some designators are created by the local organisation, like the designators stored in the 'course' fields. The content of these fields is fully defined by the school; the school creates the designators that refer to the various courses offered by the school. Some other designators are *foreign*, they originate from outside the organisation. The 'person' designators are created by the authority, and the school's 'person' fields are an example of fields which will be filled by such foreign designators. In this way, the school database is linked to the database of the authority. A similar link exists to the database of the book publisher. The 'names', 'birthdates',

Table 1 The schemata of the information that is maintained by the authority, the school and the publisher. The fields written in *italics* contain designators from an external organisation. The fields in **bold** contain raw data, that is, information which not a designator. The fields written in normal font, are designators which are locally defined.

Providing organisation	*Table name*	*Field 1*	*Field 2*
authority	names	person	**name**
authority	birthdates	person	**date**
school	students	–	*person*
school	courses	course	**name**
school	enrolments	course	*person*
school	literature	course	*book*
publisher	book_details	book	**details**

'courses' and 'details' are the only tables also containing raw data that is not encoded via a designator.

The bookshop desires a summary which states how many copies of each book can be expected to be sold. Executing the following global SQL query would provide this information. The bookshop should make sure this query is executed, and parties providing necessary information cooperate sufficiently.

```
SELECT COUNT(DISTINCT person),details
        FROM enrolments
        JOIN literature USING (course)
        JOIN book_details USING (book)
GROUP BY book;
```

To execute this query, access is needed to the 'enrolments' and 'literature' tables from the school, and to the 'book_details' table from the publisher. There are essentially two ways to execute the query. First, the query could be divided into two subqueries. The first subquery is executed by the school, its results are sent to the publisher, which performs the second subquery, and the merged result is forwarded to the bookshop. This solution works, but for more complex queries, it will become quite difficult to divide the query into subqueries. Also, the intermediate query results could leak information. The second solution could be to grant the bookshop read access to the required tables.

If these tables were plain vanilla relational databases, access to these tables would have disclosed detailed information about the interests and advances of named students. This would be a very obvious example of privacy violation. However, if the following three conditions are met, the privacy conditions are much improved.

1. The information in the tables does not contain sensitive information.
2. The information in the tables cannot be used to retrieve sensitive information.
3. The information in the tables cannot be combined with external tables to infer sensitive information.

We will show how designators can be used to make sure the tables of the school satisfy these properties. First, by using designators, it is made sure no raw identifiable data is stored in the tables, hereby meeting condition 1. Satisfying condition 2 is somewhat more complicated, but well doable. It should be ensured that though the designators can be used by *the school* to retrieve information, this cannot be done by others. In fact, it can be expected the authority would only grant the school access to its information in case it can make sure the school will not leak the information. The solution to condition 2 lies in the authority, which can create designators specially for use by the school in such a way that others, such as the bookshop, cannot materialise the designators. How this is done technically and in an efficient way is shown in appendix A.1.

Condition 3 can be met by making sure the designators given to the bookshop do not match designators referring to sensitive information the bookshop may have found elsewhere. Thus, the designators given to the bookshop should be unlinkable. However, the internal correspondences between the tables should remain intact. In our example, if a student occurs multiple times in the enrolments table, all these occurrences should be replaced by the same designator. Yet what the *actual content* of the designator is, is irrelevant and may therefore be altered. A way to create such designators on the spot is shown in Appendix A.2.

If all three conditions are met, there is no problem in granting the bookshop full access to the 'enrolments' and 'literature' tables as maintained by the school. The publisher grants the bookshop access to its 'book_details' table and everything is solved. That is, everything is solved from the privacy and unlinkability perspective, while still giving the bookshop a royal amount of freedom in accessing the information it desires to have.

Still, there is a lot to optimise. Of course, the bookshop might retrieve the full contents of the 'enrolments' and 'literature' tables, and perform the *joins* by itself, but it is easy to see that this would require a high amount of communication. It may well be the case that using subqueries and executing subqueries at various different locations is resource-wise a more optimal solution.

4.3 Observations about the use of subqueries

The approach to the question whether or not to use subqueries when assessing a global query may seem unusual. First, we found subqueries difficult, information-leaking instruments. So instead, we granted access to all information sources, but we ensured nothing sensitive was left in these information sources. Then, we observed that though operating correctly, our solution would be very inefficient so we re-allowed the use of subqueries.

However, in making a detour away from and back to the use of subqueries, we have ensured a very important property. Namely, we have obtained that any result from any subquery cannot be linked to sensitive information, because the information it stems from cannot be linked to sensitive information. Thus, we have a guarantee about the unlinkability of the subquery results. Not only have the tables from the authority not been accessed during query execution, also the subquery results and query result offer nothing that might help in getting access to the authority's tables.

The alternative to this detour would be that for each query it would need to be assessed whether the answer would somehow leak too much information. In this assessment,

answers received from previous queries should be taken into account. This easily would become complex, not to say unmanageable. The designator approach is liberal in the sense that any query which can be resolved using the 'obfuscated tables' is allowed, and restrictive in the sense that any query which cannot be resolved in this way is not allowed. In effect, linking information across organisations, and hiding information from other organisations can go hand in hand in an elegant and easy way.

The detour has in fact something more to offer. Since subquery results cannot contain sensitive information, global queries may be divided into subqueries in any way that happens to be resource-wise the most optimal. The subqueries could be executed by the organisations offering the information (e.g. the school), but also be executed by mobile agents on behalf of the information users (e.g. the bookshop).

5. DISCUSSION AND RELATED WORK

The use of information designators that we introduce in this paper allows information systems to fulfil many different roles at the same time. They can simultaneously be a transaction system, a public information system, subject to data mining, and still hide the information contained. Moreover, integrity can be guaranteed to an extent higher than normal for information integration systems.

Two important properties of the information designator system enable the seamless combination of these roles. First, the information system can supply different user 'views' of the information it has, but these views are only mutually comparable if the providing information system explicitly allows and enables this. Second, the information contained in these views (i.e., in the returned records) is not interpretable without the explicit cooperation of the providing information system.

As a result, an information system can choose to allow extensive analysis of its information, without disclosing sensitive records within this information [26]. This is useful in applications where it is undesirable for individual records to be disclosed (this would for example harm someone's privacy) but at the same time it is not a problem to produce and use accurate aggregate statistics of the information [14]. Simultaneously, administrative information exchange about such details between organisations remains possible.

An information designator can be seen as a pseudonym for information. Where pseudonyms are typically associated with people [5–7], there is no conceptual problem in using codewords to denominate a piece of information which does not refer to a person. In this perspective, a pseudonym is just a special case of an information designator. Moreover, we have generalised the idea of using multiple pseudonyms for one person to using multiple designators for one piece of information. The decision when information designators should and can be materialised is of course essentially a policy issue which has to reflect the opinions of the participants involved. Identity escrow schemes [25], and threshold-based privacy solutions [24], can be seem as special cases of solutions possible with our approach.

Information designators offer a mechanism to reason about information that is not physically present. If properly authorised, it is possible to retrieve the information that an information designator refers to. However, it is also possible to retrieve only *some* properties of the information designator at hand. In an insurance company for example, the claim experts normally see the names of the clients, because these are part of the portfolio, and are needed for subsequent steps in the claim handling process. For establishing a good judgement, the claim expert does not need the name of the client; it may even be argued that he will judge more fairly if he *does not know* the name of the client at hand. Using designators, it would be relatively easy to create workflow systems that hide all information but the information relevant in the specific step of the workflow system [33].

Reasoning about information without disclosing raw data is also subject of our other work [34, 35], in which we present protocols for comparing secrets for equality without disclosing the contents of the secrets [15]. In [34, 35], we assume two agents, both possessing 'raw data', and these agents are interested in comparing their raw data mutually without disclosing it in case the data is not equal. In this article, we demonstrate that it is also possible to compare information that is not even present at any of the two agents involved. However, the organisation that owns the information compared has to deliberately allow this comparison. We see possibilities for applying the protocols for comparing raw data to information designators, allowing very counterintuitive but useful information combination solutions, which we will investigate in our future work.

In [16, 17], cryptography is used to protect the contents of databases on a record level and field level, which has some similarities to our approach. However, in [16, 17], no cooperation from the information provider is required to materialise raw data. Our approach allows the information provider to refuse materialisation of data, which is a means of control *after* information has been disclosed in the form of information designators.

Other approaches choose to protect the privacy of the *users* against analysis of their queries by the information provider (private information retrieval) [8], or to distrust the information provider to inspect the information it stores [32]. Although these are not primary goals of our approach, we believe that similar concepts could be implemented in information designator systems. Indeed, when an organisation stores designators which it cannot materialise, this organisation is seriously limited in analysing and linking its data and the queries it receives from users.

The database representations suggested in our work form a radical departure from some of the basics of relational databases [10]. First, the tables of the database are no longer filled with actual raw data, but with some kind of 'global pointers', i.e. information designators. These designators point to information which is *vertically fragmented* over distributed information providers [11, 3, 2]. The ontologies of these providers do *not* overlap,which is dramatically different from most uses of ontologies [21, 36], and also noticeably different from the ontology use in the semantic web community [12].

6. CONCLUSION

The world described in this paper is totally different from the world we live in. We are used to information systems storing raw data, and replicating data all over the place. The information designator approach is technically not yet

sufficiently fleshed out to be applied to large-scale production-quality information systems. Also, lack of integration with existing legacy systems and lack of a critical mass of information systems using information designators, are currently prohibitive for a widespread adoption.

It is not our goal to present an instantly applicable technique. We want to demonstrate that information integration on the one hand, and privacy, unlinkability, confidentiality and related considerations on the other hand, can go hand in hand. In our presented information designator approach, goals like fluent information integration, information exchange and tight dissemination control can be satisfied simultaneously.

In line with this, we believe that the apparent trade-off between privacy and availability of information may not be as vigorous as commonly believed. The strong common belief in this apparent trade-off is a result of using information systems in which raw data is exchanged. Therefore, we believe abandoning information systems which mainly manipulate raw data may be the way to overcome the apparent delusion that information exchange and privacy can not be simultaneously established.

The fact that information designator systems demonstrate these desirable properties is a sufficient justification for elaborate research on technical details on the one hand, and on its implications on privacy and anonymity on the other hand.

ACKNOWLEDGMENTS

The author would like to thank Rineke Verbrugge, Jan-Willem Hiddink, Pieter Dijkstra and the anonymous referees for the discussions and feedback on the subjects addressed in this article.

A. METHODS FOR RESTRICTING DESIGNATOR USES

In these appendices we sketch tentative solutions for creating designators on the fly which satisfy various properties. We assume some basic understanding of cryptography and security protocols [31, 4, 18, 19]. All examples are about three organisations, *A*, *B* and *C*. *A* is always the organisation handing out a designator to B, sometimes also to *C*. Most of the solutions we present assume (deterministic) asymmetric encryption with signatures (e.g. RSA [30]).

Organisation *A* internally uses designators, which we will refer to as *master designators*. The designator it hands out to organisation *B*, will be called a *user-bound designator*. Organisation *A* has a secret, *S*. The public and private keys of *A* are pk_A and pk_A^{-1}, and similarly the public and private keys of *B* and *C* are pk_B, pk_B^{-1}, pk_C and pk_C^{-1}, respectively. The methods described in these sections can easily be combined within one step, if necessary. The purpose of showing these methods is to show that it can be done, and roughly how, omitting the deepest technical details. We do not claim that these ways of solving the problems are necessarily the best, or most efficient.

A.1 Designators that can only be materialised by a specific user

Consider an organisation *A* that would like to hand out designators to its own information to organisation *B*, granting *B* access to the information maintained by *A*. At the same time, *A* wants to make sure only *B* can materialise the designators. However *A* lacks the capacity to maintain a record of each designator individual it hands out, since this would require storage space for each designator handed out, and require computation time to look up each designator in this storage at the time of materialisation.

Now, if *A* wants to grant *B* access to the information referred to by the master designator D_M, it hands out the user-bound designator D_M^B:

$$D_M^B = \{D_M, pk_B, \textit{access-specification}, S\}_{pk_A}$$

where access-specification may be some extra information restricting the access of *B* to D_M. D_M^B is given to *B*. Nobody but *A* can decrypt D_M^B. If at some moment later in time *B* wishes to materialise the designator, it has to send $\{D_M^B\}_{pk_B^{-1}}$ (a signed copy of the designator D_M^B)[4] to *A*. In turn, *A* will decrypt D_M^B, and verify whether the signature matches the public key found in the decrypted D_M^B. If either decryption fails, the signature cannot be verified, the secret is not present, or the access-specification is not met, *A* will refuse to present the materialisation of D_M. If D_M^B falls into the hands of a third organisation, say *C*, this third organisation cannot materialise the designator since *C* is unable to forge *B*'s signature.

A.2 Designators that cannot be recombined by multiple users

Consider an organisation *A* that wants to hand out designators to both *B* and *C*, but wants to prevent that *B* and *C* can combine their information. Designators should be unique with respect to the information they refer to, but only within the realm of one single user. Thus, if *B* sees two designators D_1, D_2, it can infer whether they refer to the same information by verifying whether D_1 itself is equal to D_2. However, if *C* sees designator D_3, *B* and *C* cannot find out whether D_3 is equal to neither D_1 nor D_2 without cooperation of the organisation that handed out the designators, namely *A*.

If *A* wants to create such a user-bound designator to *B*, it hands out the following designator to *B*:

$$D_M^B = \{D_M, B, S\}_{pk_A}$$

If the designator does not need to be looked up ever by organisation *A*, the following simpler solution would also suffice:

$$D_M^B = h(\{D_M, B, S\})$$

where $h(\cdot)$ is a one-way cryptographic hash function.

Because all steps in the generating the user-bound designator are deterministic, uniqueness of designators is preserved as long as the requesting users (i.e. *B*) remains the same. However, if both *B* and *C* get a designator which refers to the information D_M refers to, these designators will not be mutually comparable.

4 Some technical tricks have to be applied to allow the message to be decrypted before verifying the signature. These tricks are easy to incorporate, but are beyond the scope of this paper.

A.3 Designators that cannot be combined over time

Consider an organisation *A* that would like to allow users to analyse the structure of the information at a specific moment in time, but does not want to allow the users to analyse how the structure evolves over time. For example, in the bookshop scenario, the school would like to prevent that the bookshop finds out how long students are studying at the school. Thus, designators should only be uniquely referring to information if these designators are all obtained at the same moment in time.

To enforce this property, *A* can create out time-dependant designators D_M^t in the following way:

$$D_M^t = \{D_M, t, S\}_{pk_A}$$

where t is the moment in time when the designator is created. To be useful, the resolution of t should not be too high, because otherwise too little designators from the same time frame would exist to make *any* snapshot inferrals. So essentially, t is a time interval. Depending on the application domain, the interval could be as long as a minute, day, week or possibly even a longer period of time. If the designator does not need to be looked up ever by organisation *A*, the following simpler solution would also suffice:

$$D_M^t = h(\{D_M, t, S\})$$

where $h(\cdot)$ is a one-way cryptographic hash function.

REFERENCES

1 **G. Alonso, F. Casati, H. Kuno and V. Machiraju.** *Web Services: Concepts, Architecture and Applications*. Springer Verlag, 2004.

2 **P. A. Boncz.** Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications. *PhD thesis*, Universiteit van Amsterdam, Amsterdam, The Netherlands, May 2002.

3 **P.A. Boncz, F. Kwakkel and M.L. Kersten.** High performance support for OO traversals in Monet. In *Proceedings British National Conference on Databases (BNCOD96)*, page 20. Centrum voor Wiskunde en Informatica (CWI), ISSN 0169-118X, 1995.

4 **Michael Burrows, Martín Abadi and Roger Needham.** A logic of authentication. *ACM Transactions on Computer Systems*, 8(1):18–36, February1990.

5 **David Chaum.** Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 24(2):84–88, February 1981.

6 **David Chaum.** Security without identification: transaction systems to make big brother obsolete. *Communications of the ACM,* 28(10):1030–1044, October 1985.

7 **David Chaum.** Achieving electronic privacy. *Scientific American*, pages 96–101, 1992.

8 **Benny Chor, Oded Goldreich, Eyal Kushilevitz and Madhu Sudan.** Private information retrieval. *IEEE Symposium on Foundations of Computer Science*, pages 41–50, 1995.

9 **Tim Churches and Peter Christen.** Some methods for blindfolded record linkage. *BMC Medical Informatics and Decision Making,* 4(9), June 2004.

10 **E. F. Codd.** A relational model of data for large shared data banks. *CACM*, 13(6):377–387, 1970.

11 **G.P. Copeland and S. Khoshafian.** A decomposition storage model. In S.B. Navathe, editor, *Proceedings of the 1985 ACM SIGMOD International Conference on Management of Data*, Austin, Texas, May 28-31, 1985, pages 268–279. ACM Press, 1985.

12 **A. Doan, J. Madhavan, P. Domingos and A. Halevy.** Learning to map between ontologies on the semantic web. *In The Eleventh International WWW Conference*, Hawaii, 2002.

13 **AnHai Doan and Alon Y. Halevy.** Semantic integration research in the database community: A brief survey. *AI Magazine,* 26(1):83–94, Spring 2005.

14 **A. Evfimievski, R. Srikant, R. Agrawal and J. Gehrke.** Privacy preserving mining of association rules. In *Proc. of 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (KDD), July 2002.

15 **R. Fagin, M. Naor and P. Winkler.** Comparing information without leaking it. *Communications of the ACM,* 39(5):77–85, 1996.

16 **J. Feigenbaum, E. Grosse and J.A. Reeds.** Cryptographic protection of membership lists. *Newsletter of the International Association for Cryptologic Research*, 9(1):16–20, 1992.

17 **J. Feigenbaum, M. Liberman and R. Wright.** *Cryptographic protection of databases and software*, 1991.

18 **Li Gong, Roger Needham and Raphael Yahalom.** Reasoning About Belief in Cryptographic Protocols. In Deborah Cooper and Teresa Lunt, editors, *Proceedings 1990 IEEE Symposium on Research in Security and Privacy,* pages 234–248. IEEE Computer Society, 1990.

19 **S. Gritzalis, D. Spinellis and P. Georgiadis.** Security protocols over open networks and distributed systems: formal methods for their analysis, design, and verification. *Computer Communications,* 22(8):697–709, May 1999.

20 **Michael Grüninger and Joseph B. Kopena.** Semantic integration through invariants. *AI Magazine,* 26(1):11–20, Spring 2005.

21 **N. Guarino.** Formal ontology and information systems. In N. Guarino, editor, Proceedings of the *1st International Conference on Formal Ontologies in Information Systems*, pages 3–15, Trento, Italy, June 1998. IOS Press.

22 **Alon Y. Halevy, Zachary G. Ives, Jayant Madhavan, Peter Mork, Dan Suciu and Igor Tatarinov.** The Piazza peer data management system. *IEEE Transactions on Knowledge and Data Engineering*, 16(7):787–798, July 2004.

23 **Jan-Willem Hiddink.** Informatie als waardegoed. *Master's thesis*, Rijksuniversiteit Groningen, August 2004.

24 **Stanislaw Jarecki, Patrick Lincoln and Vitaly Shmatikov.** Negotiated privacy (extended abstract). In *Proceedings of the International Symposium of Software Security (ISSS)*, pages 96–111, 2002.

25 **Joe Kilian and Erez Petrank.** Identity escrow. *Lecture Notes in Computer Science*, 1462:169–185, 1998.

26 **Yehuda Lindell and Benny Pinkas.** Privacy preserving data mining. *Lecture Notes in Computer Science,* 1880:36–47, 2000.

27 **Erhard Rahm and Philip A. Bernstein.** A survey of approaches to automatic schema matching. *VLDB Journal: Very Large Data Bases,* 10(4):334–350, 2001.

28 **Pradeep Ravikumar, William W. Cohen and Stephen E. Fienberg.** A secure protocol for computing string distance metrics. In *Proceedings of the Workshop on Privacy and Security Aspects of Data Mining*, pages 40–46, November 2004.

29 **R. Reiter.** On specifying database updates. *Journal of Logic Programming,* 25(1):53–91, 1995.

30 **R. L. Rivest, A. Shamir and L. Adleman.** A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21(2):120–126, 1978.

31 **B. Schneier.** *Applied Cryptography*. John Wiley & Sons, New York, 1996.

32 **Dawn Xiaodong Song, David Wagner and Adrian Perrig.**

Practical techniques for searches on encrypted data. In *IEEE Symposium on Security and Privacy*, pages 44–55, 2000.

33 **W. Teepe, R.P. van de Riet and M. Olivier.** Workflow analyzed for security and privacy in using databases. *Journal of Computer Security*, 11(3):353–363, 2003.

34 **Wouter Teepe.** New protocols for proving knowledge of arbitrary secrets while not giving them away. In Peter McBurney, Wiebe van der Hoek, and Michael Wooldridge, editors, *Proceedings of the first Knowledge and Games Workshop,* pages 99–116, Liverpool, July 2004. Department of Compute Science, University of Liverpool.

35 **Wouter Teepe.** Proving possession of arbitrary secrets while not giving them away, new protocols and a proof in GNY logic. Synthese, to appear, 2005.

36 **M. Uschold and M. Grüninger.** Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 11(2):93–155, 1996.