Part III

Approaches

Linking information which stems from various sources, also called information integration, is difficult. Also, enforcing that linked information is kept secret seems impossible. We present the information designator, which is an information pseudonym, a concept that helps to solve both problems simultaneously.

Chapter 7

Information Designators

The discussion about the state of the art in computer security in general, and privacy protection in particular, divides the participants into optimists and pessimists. Consider for example the following question:

> Is security of exchanged information a solved problem?

If one looks at this question from the cryptography perspective, the answer tends to the positive. It is possible to store or communicate information in such a way that only the intended recipients can interpret the information. The *cryptographers* (those who design cryptographic schemes) are currently way ahead of the *cryptoanalysts* (those who try to break cryptographic schemes).

On the other hand, if one looks at the question from a civil liberties perspective, the answer would definitely tend to the negative. In practice, cryptographic techniques are only reluctantly applied to protect the privacy of individual citizens. Information about individuals from various sources is combined for commercial and 'homeland security' purposes, which are not always in the interest of the individual.[1]

The extent to which the privacy of individual citizens *should* be protected is a normative, if not political question, but to what extent it *can* be protected is a scientific question. This latter question will be the focus of this chapter. The trivial answer is that privacy can be protected by making sure that no information about individuals is communicated at all. Arguably, in the current society we have become so dependent on the automated processing of information that we cannot afford such a solution. Thus, the better question would be:

[1] The American Civil Liberties Union has a clear image of one of its worst nightmares, which can be found on `http://www.aclu.org/pizza/`. In a somewhat exaggerated movie, they tell a story of someone ordering a pizza on the phone, with the person handling the call looking through the client's medical and library files.

> Can privacy of citizens be protected without prohibiting the automated information processing we depend on?

In this chapter, we will demonstrate that it is possible to facilitate intricate, distributed information processing while at the same time protecting the privacy of the individuals involved. Thus, the commonly held belief that *privacy* and *information availability* are not on good terms, is not as rigid as it seems.

We cannot achieve secure information exchange by mere application of some cryptography. Cryptographic techniques often cannot be easily applied, therefore the privacy protection is mediocre in many information systems.

In 'traditional' cryptography, there is a very clear distinction between the *good guys* and the *bad guys*. The former can be fully trusted, the latter not at all. If discussing the exchange of privacy-sensitive information, it is fair to say that not every organization processing such privacy-sensitive information is intrinsically good. In fact, if Alice is afraid Bob might misuse the information, Bob does not belong to the good guys nor to the bad guys. Probably Bob belongs to the *so-so guys*: those not intrinsically bad, but not to be trusted more than strictly necessary. Cryptography assumes a clear distinction between the trusted and the untrusted, and therefore more than just cryptography is needed if privacy needs to be protected in a context where *so-so guys* exist. With this knowledge, we can answer the opening question of this chapter as follows:

> If we accept there are parties who are not unconditionally trusted, but at the same time need to process sensitive information, the security of this information is not a solved problem (yet).

It would be ludicrous to assume that if privacy were of no concern, all information from various sources could easily be combined. Solving problems surrounding information integration properly is already so difficult [RB01, DH05, GK05], 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.

We believe however, that both information dissemination control and proper information integration can actually be achieved by one and the same instrument. In this chapter, we will present our solution, the *information designator*. 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 Section 7.1, we will introduce the research field of information integration, and how its problems relate to ontologies and dissemination of information. Section 7.2 will present our new approach to these challenges, and the central concept of this approach: the information designator. Phenomena mentioned in Section 7.2 will be illustrated in Section 7.3, where we show an example of how both information integration and dissemination control are solved jointly. Section 7.4 will show how cryptography can be used to establish desirable properties of information designators. In Section 7.5 we discuss

student	course	grade
John Doe	expert systems	A
Joe Average	statistics	C
Jim Doolittle	statistics	E
...	...	...

name	birth date
J. Average	7/6/1946
J. Doe	3/31/1948
N. Chimpsky	11/21/1973
...	...

TABLE 7.1: Two relational tables which can be combined to relate courses to birth dates.

the relevance of our approach and relate it to other research. And of course, we end with some conclusions.

7.1 Information Integration and its Challenges

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 modeled multiple times, but differently (Section 7.1.1), and from the fact that information, once disseminated from its original source, is hard to control (Section 7.1.2).

Information integration is done when a group of organizations decides 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 if 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 names to their birth dates: the student and name fields can then be used to relate courses to birth dates. (See Figure 7.1.)

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 available 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.

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 organizations, can be combined with ease and without introducing inconsistencies.

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 the information at the different locations. This includes among others record matching, data re-identification, record linkage, and this is what is traditionally understood when one refers to information integration [GK05, DH05]. 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 7.2, 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.

7.1.1 Overlapping Ontologies

An ontology defines, for a single information source, what the information stored in the source represents, and how it is structured [AvH04]. 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 organization. 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 organization to maintain all information within one stand-alone information system. Keeping track of how all information should be modeled, 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 organization have to be included.

It can be expected that in the example of the university database, inaccuracies will exist in the information that comes from outside of the organization.

Minor inaccuracies may arise from data-entry, bigger inaccuracies may arise from updating the information infrequently or not at all. Intricate inaccuracies may occur when the ontology does not have enough expressive power to facilitate the information that should be stored. When inaccurate information from various sources is combined, this will almost inevitably lead to inconsistencies.

It should be expected that the information in the university database concerning the core university activities, such as course enrollments, grades and diplomas given, is essentially, if not by definition, correct.

An organization which creates new information is probably the best suited organization to model this information, and to maintain an ontology of this information. However, it is not unusual for such an organization to maintain an ontology covering more than the *core business* of the organization 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 organizations.

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 be impossible 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.

7.1.2 Information Propagation

The reason for linking information sources, i.e., to perform information integration, is twofold from the perspective of a participating organization. First, the organization wants to *retrieve* authoritative information from external sources. When retrieving data, the desiderata are *availability* and *integrity* of the information. Second, the organization 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.[2] These aims and interests of the participating organizations are depicted in Figure 7.1.

To maximize integrity of information, it would be good 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 authorization for each individual 'shipment' of information, and to make sure the information can only be used for the purpose stated in the authorization procedure.

[2] It is rarely if ever the case that an organization would want to *directly* alter information that is within the realm of another organization.

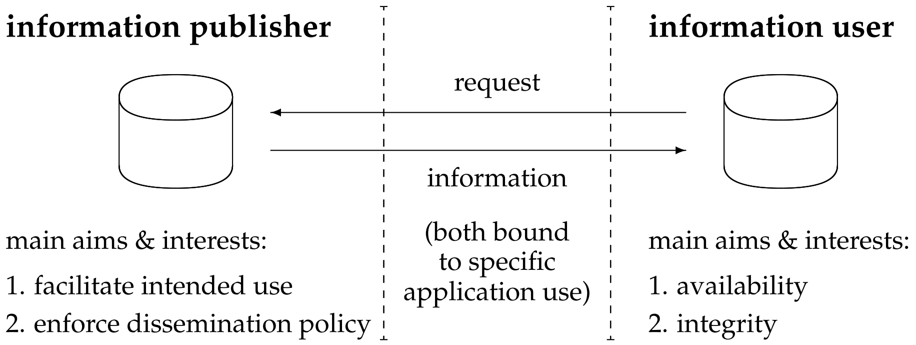

FIGURE 7.1: The main aims and interests for organizations participating in information integration. Virtually every organization in an information integration setting is both *user* of some externally published information, and *publisher* of some other information. This figure depicts the interests for one organization in the role of *information publisher* and another in the role of *information user* with respect to one 'piece' of information.

This leads to a central adage 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 adage 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. For example, if counting the existence of a piece of information (such as when counting students in a room) is regarded as 'using' information, counting a student would lead to disclosure of his personal information. If 'using information' 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 minimize, 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 may lower the capacity of information sources to disseminate information, but it will give the information holder much more control over who has access to what information.

7.2 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 7.2.1. Section 7.2.2 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 7.2.3 we explain how an organization that provides information designators to others, can accurately manipulate which others can actually use the provided information designators, and to what extent.

7.2.1 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 and without any reference to a context. 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`[3], 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, as follows:

- The *principal* step is the one in which Bob asks Alice her home address, and possibly after some combination of authorization 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 'I owe you' (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

[3] This could be a phone number, IP address, or something else that allows setting up a communication channel in an automated way.

by instructing the exchange agent not to give Bob the information the designator refers (or: referred) to.

- The second step is the *materialization* 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. 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.[4]

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.

The architecture presented here could be considered to be a peer-to-peer (P2P) data management system (PDMS), like the Piazza PDMS [HIM+04]. 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 the web services and the semantic web [ACKM04] 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.

7.2.2 Dependency and (Un)linkability

It may seem that by using information designators, the users of information are subject to possible arbitrary behavior 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 that 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 an essential requirement for many organizations to participate in an information integration project. More organizations will be willing to provide information, because

[4] Thus, because the 'raw data' such as a street name is separated from the concept of what it represents in the data structure, it is possible to perform database transactions on the 'raw data' without even touching the database records that link to the 'raw data'. From the perspective of complex database transactions and the frame problem, this is an interesting feature [Rei95].

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 organizations explicit. If an organization depends for some task on information from another organization, this will inevitably lead to an infrastructure in which designators are used whose corresponding exchange agents operate under the auspice of the organization depended on.

The information designator approach has the very interesting property that if it is fully applied, there need not be overlapping ontologies. Different organizations *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.[5] At the same time, information can be linked without automatically disclosing a part of the linked information: information 'normally' (otherwise) made public can be kept private.

7.2.3 Operations on Designators

One could wonder whether introducing designators actually improves privacy and anonymity, by reasoning that the designators themselves will fulfill 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 recognize 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 organization handing out designators could in fact every time it hands out a designator, create an

[5] The claim is somewhat weak, and this is on purpose: there might be other classes of inconsistencies we have not thought of. As we cannot prove to prevent *all* types of inconsistency, we will not claim so.

extra ‘fully anonymous’ designator for the information it needs to point to.[6] In this scenario, the organization 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 organization, or if given to an organization in some given group, thereby allowing the organization or group of organizations to compare their designators. It is totally at the discretion of an organization 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 organization that handed out the designators.

An organization 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 organization 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 way. Organizations 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.

7.3 An Example: the Datamining 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 it would seriously infringe privacy. We present the prototype here for three purposes:

1. to stress that information designator systems *can actually be built* [Hid04],

[6] 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 Section 7.4.

2. to show how an information designator system works internally, thereby illustrating the subject matters explained in the previous section, and
3. to give an application example which demonstrates how information designators help in protecting privacy and maintaining unlinkability.

7.3.1 Organizational Setting

Our example is about information flow between the following four organizations.

Civic Authority This organization has the task to maintain the municipal inhabitants register, which contains inhabitants' names, birth dates, and residence addresses.

Local School The students of the local school live in the domain of the civic authority. The local school keeps record of its students, their results, their course enrollments and required literature for courses.

Local Bookshop This organization is located conveniently next to the local school. The local bookshop wants to provide for the literature demands from the local school students, but does not want to overstock.

Book Publisher This organization publishes the books that are used in the courses of the local school. The book publisher maintains information about books and their details, such as titles, authors and ordering information.

There are many relations between the information maintained by these organizations. The students of the local school are all registered at the civic authority. Contrary to the book publisher and the local bookshop, the local school has the right to access some of the information stored and maintained by the civic authority. The books the local school recommends for their various courses, are all published by the book publisher. The book 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 local bookshop.

The local bookshop has a strong desire not to overstock books, and at the same time the local school wishes all their students to have their obligatory books when the term starts. As a result, the local school depends on the behavior of the local bookshop, and the local bookshop depends on information from the local school. A very naive way to solve this dependency would be that the local school gives the local bookshop full access to the local school administration. This would obviously lead to unacceptable privacy infringements, even if the local school would limit the access to things like course enrollments (and hide exam results). A slightly less naive solution would be that the local school gives the local bookshop an update of the expected number of required books once in a while. However, these updates are just snapshots. It would be ideal

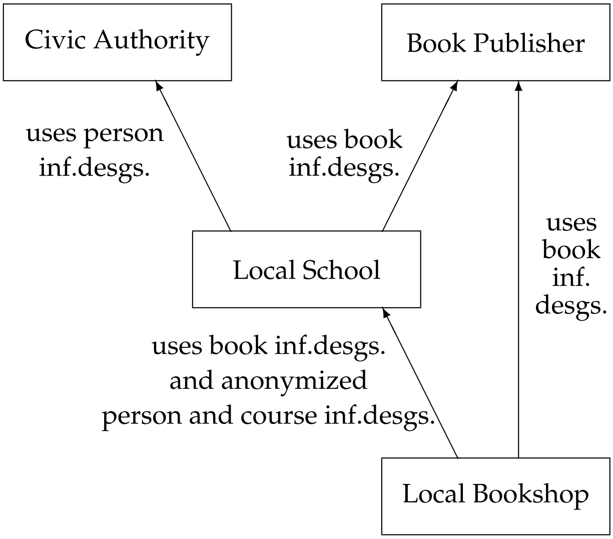

FIGURE 7.2: An information dependency graph containing the four organizations of the example. The organizations and their information demands are described in Section 7.3.1. An arrow from organization A leading to B means that A is interested in information maintained by B. For example, the local bookshop depends on (desires) information from both the local school and the book publisher. 'Information designators' is abbreviated to 'inf.desgs.'.

for the local bookshop to directly look in the administration of the local school at the moments relevant for the local bookshop. If this would not infringe on the privacy of the students, the local school would probably find such a solution fairly unproblematic.

Figure 7.2 shows how the four organizations relate to one another with respect to their information needs.

The example may seem a perfect case for setting up a Web Service framework [ACKM04]. However, a Web Service framework would offer only a means for exchanging information, while 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.

7.3.2 Designators in Action

The information that is maintained by the organizations is summarized in table 7.2. The table shows the schemata of the local databases. These might 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

providing organization	table name	field 1	field 2
civic authority	names	person	**name**
civic authority	birthdates	person	**date**
local school	students	-	*person*
local school	courses	course	**name**
local school	enrollments	course	*person*
local school	literature	course	*book*
book publisher	book_details	book	**details**

TABLE 7.2: The schemata of the information that is maintained by the civic authority, the local school and the book publisher. The fields written in *italics* contain designators from an external organization. The fields in **bold** contain raw data, that is, information which is not a designator, and therefore readily interpretable. The fields written in normal font, are designators which are locally defined.

contain *information designators*. Some designators are created by the local organization, like the designators stored in the 'course' fields. The content of these fields is fully defined by the local school; the local school creates the designators that refer to the various courses offered by the local school. Some other designators are *foreign*, they originate from outside the organization. The 'person' designators are created by the civic authority, and the local school's 'person' fields are an example of fields which will be filled by such foreign designators. In this way, the local school database is linked to the database of the civic authority. A similar link exists to the database of the book publisher. The 'names', 'birthdates', 'courses' and 'details' are the only tables also containing raw data that is not encoded via a designator.

The local bookshop desires a summary which states how many copies of each book can be expected to be sold. Executing the global SQL query shown in Figure 7.3 would provide this information. The local bookshop should make sure that this query is executed, and that parties providing necessary information cooperate sufficiently.

To execute this query, access is needed to the 'enrollments' and 'literature' tables from the local school, and to the 'book_details' table from the book 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 local school, its results are sent to the book publisher, which performs the second subquery, and the merged result is forwarded to the local 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 local 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 ob-

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

FIGURE 7.3: A global SQL query, which would provide the local bookshop with the information it desires: details of each of the books needed by the students of the local school, and the number of required copies of each of these books. To obtain this desired table, the 'enrollments' and 'literature' table of the local school are consulted, as well as the 'book_details' table of the book publisher.

vious 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 local school satisfy these properties. First, by using designators, it is ensured that 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 made sure that though the designators can be used by *the local school* to retrieve information, this cannot be done by others. In fact, it can be expected that the civic authority would only grant the local school access to its information in case it can make sure the local school will not leak the information. The solution to condition 2 lies in the civic authority, which can create designators especially for use by the local school in such a way that others, such as the local bookshop, cannot materialize the designators. How this is done technically and in an efficient way is shown in Section 7.4.

Condition 3 can be met by making sure the designators given to the local bookshop do not match designators referring to sensitive information the local bookshop may have found elsewhere. Thus, the designators given to the local 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 enrollments 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 Section 7.4.

If all three conditions are met, there is no problem in granting the local bookshop full access to the 'enrollments' and 'literature' tables as maintained by the local school. The book publisher grants the local 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 local bookshop a royal amount of freedom in accessing the information it desires to have. The local bookshop can get up-to-date information at any moment it wishes.

Still, there is a lot to optimize. Of course, the local bookshop might retrieve the full contents of the 'enrollments' 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. Therefore, the ideal approach should be liberal in allowing queries to be divided into subqueries. Our approach is such an approach, and this will be the focus of the next section.

7.3.3 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 that 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 civic authority not been accessed during query execution, also the subquery results and query result offer nothing that might help in getting access to the civic 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 organizations and hiding information from other 'third' organizations 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 organizations offering the information (e.g. the local school), but also be executed by mobile agents on behalf of the information users (e.g. the local bookshop).

7.4 Methods for Restricting Designator Uses

In Section 7.3.2, we have assumed that it is viable to ensure certain properties of designators, such as that it is impossible to recombine designators in specific ways. In this section, we sketch tentative solutions for creating designators which satisfy these properties.

All examples are about three organizations, A, B and C. Organization A (the *information publisher*) is always the organization handing out a designator to B, sometimes also to C (the *information users*). Most of the solutions we present assume (deterministic) asymmetric encryption with signatures (e.g. RSA [RSA78]). When a cryptographic hash function $H(\cdot)$ is used, it is assumed that it is a correlation-free non-incremental cryptographic hash function (see Chapter 3 and Section 3.6).

Organization A internally uses designators, which we will refer to as *master designators*. The designator it hands out to organization B will be called a *user-bound designator*. Organization A has a private secret, S. The public and private keys of A are $+K_A$ and $-K_A$, and similarly the public and private keys of B and C are $+K_B$, $-K_B$, $+K_C$ and $-K_C$, respectively.

The methods described in thesis section 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 ones.

1. *Designators that can only be materialized by a specific user*

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

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

   $$D^B = \{D, +K_B, \text{access-specification}, S\}_{+K_A}$$

   where access-specification may be some extra information restricting the access of B to D. The user-bound designator D^B is given to B. Nobody but A can decrypt D^B. If at some moment later in time B wishes to materialize the designator, it has to send $\{D^B\}_{-K_B}$ (a signed copy of the designator D^B)[7] to A. In turn, A will decrypt D^B (using his private key

[7] It has to be made sure that A can decrypt D^B before verification of the signature, since the public key $+K_B$ required for verification is stored within D^B. A 'two-step' signature scheme can facilitate this.

$-K_A$), and verify whether the signature matches the public key $+K_B$ found in the decrypted D^B. If either decryption fails, or the signature cannot be verified, or the secret is not present, or the access-specification is not met, then A will refuse to present the materialization of D.

If D^B_M falls into the hands of a third organization, say C, this third organization cannot materialize the designator since C is unable to forge B's signature.

2. *Designators that cannot be recombined by multiple users*

   Consider an organization 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 receives two designators D^B_1, D^B_2, it can infer whether they refer to the same information by verifying whether D^B_1 itself is equal to D^B_2. However, if C receives designator D^C_3, B and C should not be able to find out whether D^C_3 is equal to either D^B_1 or D^B_2 (without cooperation of the organization 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:

$$D^B = \{D, B, S\}_{+K_A}$$

   If the designator never needs to be looked up by organization A, the following simpler solution would also suffice:

$$D^B = H(D, B, S)$$

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

3. *Designators that cannot be recombined over time*

   Consider an organization A that would like to allow users to analyze the structure of the information at a specific moment in time, but does not want to allow the users to analyze how the structure evolves over time. For example, in the local bookshop scenario, the local school would like to prevent the local bookshop from finding out how long students are studying at the local 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 time-dependent designators D^t in the following way:

$$D^t = \{D, t, S\}_{+K_A}$$

where t is the moment in time when the designator is created. Essentially, t is a time interval, and some care must be taken in choosing the size of this time interval. To be useful, t should not be too small, because otherwise too little designators from the same time frame would exist to make *any* snapshot inferences. 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 never needs to be looked up by organization A, the following simpler solution would also suffice:

$$D^t = H(D, t, S)$$

Note that this solution does not require a global clock, but only a local clock for A.

The space requirements (i.e., size) of designators are only limited. A designator which is constructed using a cryptographic hash function is trivially bounded in length, with current cryptographic hash functions only a few hundred bits. A designator which is constructed using an encryption step is bigger than the designator it encapsulates by a constant factor. A designator never reveals the length of the information it designates.

7.5 Discussion and Related Work

The use of information designators that we introduce in this chapter allows information systems to fulfill many different roles at the same time. They can simultaneously be a transaction system, a public information system, subject to datamining, 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:

1. The information system can supply to different users different 'views' of the information it has, but these views are only mutually comparable if the providing information system explicitly allows and enables this.

2. 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 [LP00]. 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 [ESAG02]. Simultaneously, administrative information exchange about such details between organizations remains possible.

An information designator can be seen as a pseudonym for information. While pseudonyms are typically associated with persons (as in [Cha81, Cha85, Cha92]), 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 generalized 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 materialized is of course essentially a policy issue which has to reflect the opinions of the participants involved. Identity escrow schemes [KP98] and threshold-based privacy solutions [JLS02] can be seen 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 authorized, 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 judgment, 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. Similar considerations apply to tasks like the judging of job applications. 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 [TvdRO03].

Reasoning about information without disclosing raw data is also subject of Chapters 8–10 of this thesis, in which we present protocols for comparing secrets for equality without disclosing the contents of the secrets [FNW96]. In Chapters 8–10, we consider 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 that chapter we demonstrate that it is also possible to compare information that is *not even present* at any of the two agents involved. However, the organization that owns the information compared has to deliberately allow this comparison.

Thus, for the sake of the protocols of the next chapters, the information designators could be considered as ‘raw data’. This allows for example two organizations, who have pools of ‘anonymous data items’, to compute the intersection of these pools *without identification of the data items themselves*. Such a rather counter-intuitive computation may have a number of applications, such as privacy-respecting informed policy-making.

In [FGR92, FLW91], 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 [FGR92, FLW91], no cooperation from the information provider is required to materialize raw data. Our approach allows the information provider to refuse materialization 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) [CGKS98], or to distrust the information provider to inspect the information it stores [SWP00]. Although these are not primary goals of our approach, we believe that similar concepts could be implemented in information designator systems. Indeed, when an organization stores designators which it cannot materialize, this organization is seriously limited in analyzing 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 [Cod70]. 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 [CK85, BKK95, Bon02]. The ontologies of these providers do *not* overlap, which is dramatically different from most uses of ontologies [Gua98, UG96], and also noticeably different from the ontology use in the semantic web community [DMDH02].

7.6 Conclusion

In this chapter, we have described a way of structuring and linking information that is totally different from the way that information is structured and linked nowadays. Nowadays, it is common that information systems store raw data, and replicate data almost abundantly. 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 the presented information designator approach, goals like fluent information integration, information exchange and tight dissemination policy enforcement 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 rigid 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 part of the way to overcome the misunderstanding that information exchange and privacy can not be simultaneously established.