Privacy Issues of Provenance in Electronic Healthcare Record Systems

Tamás Kifor[1], László Z. Varga[1]
Sergio Álvarez[2], Javier Vázquez-Salceda[2], Steven Willmott[2]

[1] Computer and Automation Research Institute, Kende u. 13-17, 1111 Budapest, Hungary
{tamas.kifor, laszlo.varga}@sztaki.hu
http://www.sztaki.hu/
[2] Universitat Politècnica de Catalunya, Jordi Girona Salgado 1-3, E - 08034 Barcelona, Spain
{salvarez, jvazquez, steve}@lsi.upc.edu
http://www.upc.edu/

Abstract. Cooperation techniques modelled by agent systems and standardised electronic healthcare record exchange techniques help the reunification of the different pieces of the therapy of a single patient executed in a distributed way at different places, but currently these models and techniques are ad-hoc and based on the information provided by the patient. In the organ transplant demonstration application of the Provenance project we propose the usage of the novel provenance techniques to provide better healthcare services for patients by providing a unified view of the whole health treatment history. While this is good to improve the medical processes, it also introduces new privacy risks, because the agent with the provenance information knows much more about the patient than any other agent in the system. In this paper we are going to investigate the privacy aspects of introducing provenance into healthcare information systems and propose methods against the new risks.

1 Introduction

The applications of the agent paradigm for Healthcare information systems increase day by day [1]. Agents can support communication and coordination not only between organizations but even among all members of a medical team, allowing the sharing of information and providing distributed decision making support. Agents can also be used to adapt medical services to patients' needs (*personalization*). Moreover, the flexible way in which agents operate is suited to the dynamic situations in the open and changing environment in which healthcare information systems are expected to operate. In distributed scenarios, modelling the application components as agents with some degree of autonomy easily reflects the decentralized nature of the network of healthcare institutions and can be considered as the natural extension to the notion of encapsulation in systems that are owned and developed by different authorities.

Although the agent paradigm is well suited to modelling healthcare information systems, sometimes the distributed nature of healthcare institutions themselves hinders the treatment of patients, because the indivisible healthcare history and therapy of the patient is allocated to independent and autonomous healthcare institutions. Cooperation techniques modelled by agent systems and standardised electronic healthcare record exchange techniques help the reunification of the different pieces of the therapy of a single patient executed at different places, but currently these are based on ad-hoc methods and the information provided by the patient. The Provenance project [2] studies the provenance of electronic data in service oriented architectures in order to enable users to trace how a particular result has been produced by identifying the individual and aggregated services that produced a particular output. In the organ transplant demonstration application of the Provenance project we propose the usage of the novel provenance techniques to provide better healthcare services for patients by providing a unified view of the whole health treatment history.

Privacy[1] is especially critical issue in healthcare applications. Patients and practitioners entrust sensitive information to the different agents of the healthcare information system. This information is necessarily disclosed to third parties and shared between the agents participating in the treatment of the patient. As long as the treatment and the data related to the treatment are distributed among the agents of the healthcare information system, privacy protection is focused on the protection of partial information pieces, but with the introduction of provenance into the system we re-integrate the different pieces. This is good to improve the medical processes, but it also introduces new privacy risks, because the agent with the provenance information knows much more about the patient than any other agent in the system. In this paper we are going to investigate the privacy aspects of introducing provenance into healthcare information systems and propose methods against the new types of risks.

The contents of this paper are as follows. In section 2 we will introduce the problem of heterogeneous, distributed EHCR's, how we can reconstruct the workflow of a patient data and the application scenario. In section 3 we will talk about the privacy issues to be solved. In section 4 we will discuss how privacy is handled in our EHCR application, and then in section 5 we will conclude with a summary of our approach and some references to related work.

2 Distributed and Heterogeneous EHCR Applications

Cooperation among people using electronic information and techniques is more and more common practice in every field including healthcare applications as well. However healthcare data is often distributed among several heterogeneous and autonomous information systems which are under the authorities of different healthcare

[1] The terms *privacy*, *confidentiality* and *security* are used in many ways when discussing the protection of personal medical information. The convention used in EU Regulations is to use privacy referring to the desire of an individual of limiting the distribution of his/her existing medical information. The term confidentiality will refer to the conditions under which personal medical information is shared and/or distributed in a controlled fashion. Security refers to the measures implemented by the organizations in order to protect the information in charge of them and the systems on which it is stored.

actors like general practitioners, hospitals, hospital departments, etc. Each actor has its own way of operation which means that they independently and autonomously define their workflows and data representation. This leads to fragmented and heterogeneous data resources and services forming islands of information. The data, containing the healthcare history of a single patient, and the corresponding workflow chunks are distributed among these islands of information. In order to provide better healthcare services, the treatment of the patient might require viewing these pieces of workflow and data as a whole. The integration of the information islands might bring about several advantages like consistent patient records and inter-organisational workflows.

The main reasons for the existence of the information islands are the diversity of healthcare activities, the diversity of developers and the legacy systems. Healthcare activities depend not only on the actual medical process, but also on the organisation carrying out the activity, the legal regulations, the cultural aspects, the preferences of health professionals' groups and many other factors. Therefore most healthcare actors developed their own information systems which can be modelled as autonomous agents. The developers of these information systems come from different sectors of the healthcare information technology market, often with products deeply specialised on certain problem or general products covering all aspects of hospital information systems without specialised functionality. The already existing systems contain high amount of data and development investment, which resist modification and evolution, therefore new developments build on legacy systems. These factors lead to heterogeneous and autonomous information systems under the control of diverse organisations.

As [3] points out, the ENV 13606 pre-standard [4] developed by CEN/TC251 (European Committee of Normalisation, Technical Committee 251) is vital for the exchange of clinical data between healthcare organisations and departments. Standardised electronic healthcare record (EHCR) architecture helps in structuring any medical data in a uniform way and presenting the multitude of different facts while preserving the meaning and context of the data.

Although the EHCR architecture defines how to exchange data, the linking of the workflow pieces which generated the data is not discussed in EHCR standards. The provenance architecture (discussed later in this paper) helps to document the way the data was created and link the workflow pieces together.

2.1 Electronic Healthcare Records and Case Antecedents

The EHCR architecture we are implementing provides a way to build electronic health records as well as a unified view of a patient's medical record. The architecture provides the structures to build a part of or the full patient's healthcare record drawn from any number of heterogeneous databases systems in order to exchange it with other healthcare information systems. It uses the ENV13606 pre-standard, which defines the messages, the retrievable objects, the healthcare agents and the distribution rules.

The pre-standard knows three types of messages: *request*, *provide* and *notification*. All of them contain the identification of the message, the issue date and time of the message, the EHCR *source/destination agents*,[2] urgency of the message, patient matching information (the subject of the message) and message receipt acknowledgement request. Besides these data, any message may contain EHCR *message related agents* which are important healthcare agent(s) other than EHCR source or destination. Message related agent can be for example an authorisation agent, or in our case the provenance system discussed later. The information in the message may include an identification of the nature of the enterprise environment and/or communicating community of the party sending the message (e.g. organ transplant management application). *Request* messages contain a reason for the request and *notification* messages contain the type and comment of the notification. EHCR data are sent in *provide* messages which also contain distribution rule directory to specify privacy protection rules.

Every message is about one and only one patient and his/her EHCR. An EHCR consists of record components. The simplest instance of an EHCR consists of an EHCR extract class containing a single text data item with the component role „Narrative Text". Main types of the record components are EHCR extract, folder, composition, headed section, cluster, link set item and the data item. The EHCR *extract* contains all the other record components in the message. A *folder* is a collection of record components. Contents of a folder data are collected by different people in different time and place, e.g. nursing notes, specialist departmental record, etc. We do not detail the other types of record components which are used to structure the EHCR according to time and place of care delivery, recording session, a common theme or healthcare process, etc. The smallest structural unit into which the content of the EHCR can be broken down without losing its meaning is the *data item*.

In the pre-standard, a *healthcare agent* is a healthcare person, a healthcare organization, a healthcare device (e.g. x ray machine, ECG machine), or a healthcare software component that performs a role in a healthcare activity. For instance, healthcare agents may be the sender / recipient of an EHCR message, the requester / provider of an EHCR, a person signing of a message or record entry, originator or author of a record entry. *Relationships* between two healthcare agents can be defined (e.g. employee / employer). The same healthcare agent can exist in different contexts (e.g. the same doctor working in different hospitals). A healthcare agent in context has a unique identifier, a reference to a healthcare agent, function (e.g. duty doctor, locum) and relationships to other healthcare agents. The *healthcare agents directory* may contain several healthcare agents (in context). Using this directory, the sender need only include the full details of any healthcare agents (in context) once.

With *distribution rules*, the provider of the EHCR (or somebody else) can define who, when, where, how and with what type of access can access a part of the EHCR, or add/invalidate distribution rules to that part of the EHCR. Besides this information, a distribution rule has the necessary data to be able to identify the author of the distribution rule. A distribution rule is attached to a message component with a distribution

[2] Please note here that by the word *agent* the ENV13606 pre-standard refers to an actor (a human, an organization, a software component) interacting in the system.

rule reference which contains information on who and when applied the rule to the message component, the interval of the validation of the rule, the country where the rule is valid and the reference to the healthcare agent in context who invalidated the rule within the period of time originally applied.

Current healthcare systems work by storing master copies data about individual medical interventions on a patient at the place where the interventions are carried out. Most commonly a single GP oversees a patient's medical history and thus integrates interventions not carried out under his/her own supervision post event. However there is no standard process for forwarding medical details which might form part of the record to a central registry or a master copy a particular patient's record. Information is retrieved from different healthcare providers on the basis of the patients Identity Number (ID). A healthcare provider A may only ask for record information from another provider B for a patient X if the patient X is physically being treated at A. Usually there is no central health authority database that could be relied upon to have a complete medical history.

In order to pull together the medical history of a patient we have essentially three options:

- Build a system mirroring the current one based on fragments of records in different places which can be pulled together to produce a unified view on demand (depending on the permissions of the viewer).
- Build a system of a more centralised nature with a master record which can be read and written to by authorised healthcare providers (in a controlled fashion) and possible cached at a particular healthcare provider.
- Build a hybrid system which stores fragments of data with providers but records high level events in a central master record.

In both cases the interchange protocol could be one of the new European prestandards. However it should be noted that current record systems are unlikely to change for quite some time.

2.2 Provenance in Service Oriented Architectures

As we could see in the previous section, there is need to collect the electronic trace of the medical history of patients. In order to support this we use the outputs from the EU Provenance project [2] which studies the provenance of electronic data in service oriented architectures. The aim of the Provenance project is to design, conceive and implement an industrial-strength open provenance architecture for distributed systems using Web Services or Grid technology, and to deploy and evaluate it in complex distributed applications, namely aerospace engineering and organ transplant management. The latter one is discussed in this paper. In the following we are going to describe provenance based on the Provenance Architecture document [5].

The concept of *provenance* is already well known in fine art where it refers to the trusted, documented history of some work of art. Given that documented history, the object attains an authority that allows scholars to understand and appreciate its importance and context relative to other works of art. Objects that do not have a trusted, proven history may be treated with some scepticism by those that study and view them. This concept of provenance may also be applied to data and information generated within a computer system, especially when the information is subject to regulatory control over an extended period of time. The EU project defined *provenance* concept as: “the provenance of a piece of data is the process that led to the data”. Provenance enables users to trace how a particular result has been achieved by identifying the individual and aggregated services that produced a particular output.

The aim of the Provenance project is to conceive a computer-based representation of provenance that allows users to perform useful analysis and reasoning. The provenance of a piece of data will be represented in a computer system by some suitable documentation of the process that produced the data. This documentation can be complete or partial (for instance, when the computation has not terminated yet); it can be accurate or inaccurate; it can present conflicting or consensual views of the actors involved; it can be detailed or not. The Provenance project assumes that provenance is investigated in open, large-scale systems typically designed using a service-oriented approach [6]. Services are regarded as components that take inputs and produce outputs. Such services are brought together to solve a given problem typically via a workflow that specifies their composition. In this abstract view, interactions with services (seen as *actors*) take place using messages that are constructed in accordance with service interface specifications.

Actors may have internal states that change during the course of execution. An actor’s state is not directly observable by other actors; to be seen by another actor, the state (or part of it) has to be communicated within a message sent by its owner actor. The technology-independent approach of the Provenance project to service-oriented architectures (SOAs) has formal foundations in the π-calculus [7] and asynchronous distributed systems [8]. According to this view, messages are the only mechanism used to transfer information between actors. The π-calculus is of interest in this context because of its approach to defining events that are internal to actors as hidden communications. This view also allows us to formally define mappings with agent-mediated services and to use the Provenance project results for Multiagent Systems.

The provenance of a data item is represented in a computer system by a set of *p-assertions* made by the actors involved in the process that created it. A p-assertion is a specific piece of information documenting some step of the process made by an actor and pertains to the process. There are two kinds of p-assertions that capture an explicit description of the flow of data in a process: *interaction p-assertions* and *relationship p-assertions*. An interaction p-assertion is an assertion of the contents of a message by an actor that has sent or received that message. A relationship p-assertion is an assertion about an interaction, made by an actor that describes how the actor obtained output data or the whole message sent in that interaction by applying some function to input data or messages from other interactions. In addition, there is the *actor state p-assertion* which is an assertion made by an actor about its internal state in the context of a specific interaction.

The long-term facility for storing the provenance representation of data items is the *provenance store*. The provenance store is used to manage and provide controlled access to the provenance representation of a specific data element. The provenance lifecycle is composed of four different phases. First, actors create p-assertions that are aimed at representing their involvement in a computation. After their creation, p-assertions are stored in a provenance store, with the intent they can be used to reconstitute the provenance of some data. After a data item has been computed, users or applications can query the provenance store. At the most basic level, the result of the query is the set of p-assertions pertaining to the process that produced the data. More advanced query facilities may return a representation derived from p-assertions that is of interest to the user. Finally the provenance store and its contents can be managed (subscription management, content relocation, etc).

The Provenance project develops an architecture, tools and a reference implementation to support this provenance life-cycle.

2.3 Organ Transplant Management Application

The developments of the Provenance project will be demonstrated on the Organ Transplant (OTM) application described in this section. The OTM application is an excellent case study of both provenance and the privacy issues of provenance.

Treatment of patients through the transplantation of organs or tissue is one of the most complex medical processes currently carried out. This complexity arises not only from the difficulty of the surgery itself but also from the fact that it is a distributed problem involving several locations (donating hospital, potential recipient hospitals, test laboratories and organ transplant authorities, see figure 1), a wide range of associated processes, rules and decision making. Depending on the country where a transplant is being carried out, procedures and the level of electronic automation of information / decision making may vary significantly. However, it is recognized worldwide that ICT solutions which increase the speed and accuracy of decision making could have a very significant positive impact on patient care outcomes.

In [9, 10] we presented CARREL, an Agent-Mediated Electronic Institution for the distribution of organs and tissues for transplantation purposes. One of the aims of the CARREL system was to help speeding up the allocation process of solid organs for transplantation to improve graft survival rates. The policy that we implemented followed the Spanish guidelines for organ and tissue procurement and Spanish regulations for allocation, as Spain is world leader in the area, followed as a model by other countries. In most of the official organ allocation organizations, the process is composed of three phases:

- Each hospital informs the related organ transplant authority (OTA) about patients that have been added to or removed from the waiting list of that hospital, or to be added to or removed from the national-wide Maximum Urgency Level Waiting List.
- When a donor appears, the hospital informs the OTA of all the organs suitable for donation in the form of offers sent to the OTA, which then assigns the organ(s).
- The organ(s) are extracted from the donor, checked, brought to the recipient hospital(s) and then implanted in the recipient(s). All these steps of the process involve quite some decision making and quality assessment, and all this should be recorded properly for future potential audits of the process.

Several prototypes of the CARREL system have been developed using JADE [11]. Although medical practitioners positively evaluated the prototypes, system administrators proved to be very reluctant to manage agent platforms for critical medical applications, and prototypes didn't go through.

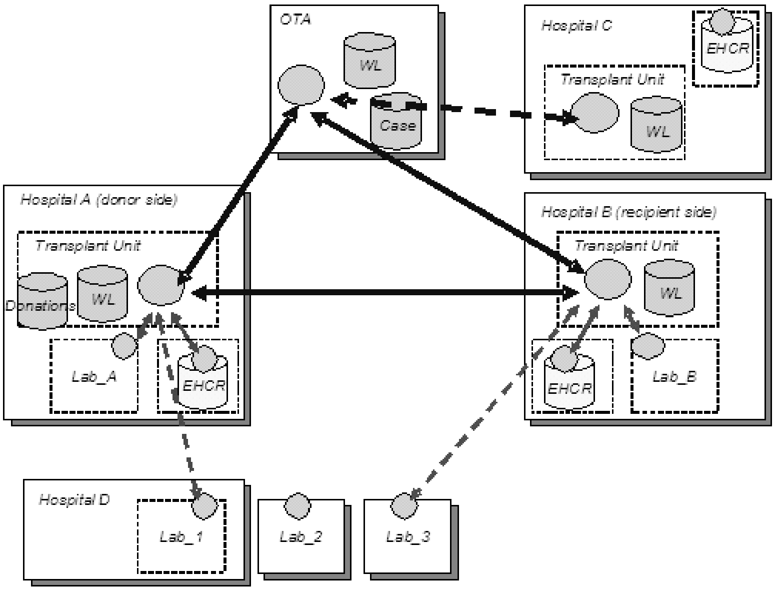

Fig. 1. The OTM Application. Solid boxes denote the different administrative domains and dashed boxes denote units that are involved during a transplantation management scenario. Each of these interact with each other through Web Service interfaces (circles) using Agent Communication Language. Some of the involved data stores are: the patient records, stores for the transplant units and the Organ Transplant Authority (OTA) recipient waiting lists (WL). Hospitals that are the origin of a donation also keep records of the donations performed, while hospitals that are recipients of the donation may include such information in the recipient's patient record. The OTA has its own records of each donation, stored case by case.

In [12] we proposed a connection between Agent Communication Languages and Web Service Inter-Communication. This allows us to implement agent systems by means of web services which can interact following the same FIPA protocols [13]. With this approach we are developing a new prototype, the Organ Transplant Management (OTM) Application which uses standard web service technology and it is able to interact with the provenance stores in order to keep track of the distributed execution of the allocation process for audit purposes.

3 Privacy Issues

In healthcare applications enforceable privacy rules are extremely important. Individuals share a lot of sensitive, personal information with their doctors like physical conditions, personal habits, sexual practices, mental state, medications, family history, etc. Full disclosure is necessary for proper diagnosis and treatment. Patient information is then shared with many people, including doctors, hospitals, pharmacies, employers, relatives, schools, researchers, insurance companies, pharmaceutical companies, public health officials, and even the press and marketers. Many of these disclosures are necessary to treat patients, process claims, measure outcomes, and fight disease, therefore privacy protection should not focused on nondisclosure, but on controlled and irreversible disclosure, which mainly means the protection of the identity of the patient.

When we extend healthcare systems with provenance in order to provide better services for patients, then we face new privacy issues in addition to those already handled in the healthcare information system. In the following we are going to summarize the privacy issues in healthcare record management and the new challenges in the provenance extension.

3.1 Privacy in Healthcare Record Management

Protection of individual's health-related data has been a continued concern of the medical body from the very beginning of the medical practice, as reflected in the famous Hippocratic oath. It obliges the physician to conserve the secret of as many information he or she will obtain from the patient during the treatment. There exist considerable efforts to put in practice a body of policies which ensure the protection of medical data in a scenario of massive use of computers in the health sector.

The European Union has always manifested a special concern about the protection of their citizen's personal data. In 1995 there were already several countries with suitable legislations about processing personal data. However, these legislations did not allow the exchange of personal information between states. European Parliament created the 95/46/CE Directive [14] with the purpose of homogenizing legal cover on data protection, in order to warrant an appropriate protection level on each transfer inside the European Union. Then in 1997 the Recommendation R(97)5 [15] about medical data was drafted. This recommendation basically arose for two reasons: 1) it was noticed that the advances on medical science strongly depend on the availability of medical data about individuals, and 2) an increase of the use of automated information systems for processing medical data was detected, not only for medical assistance and research, hospital management and public health, but also outside the health sector (e.g., insurances), which was a reason of concerns. The R(97)5 permits the collection and processing of medical data for preventive, diagnostic or therapeutic purposes related to the affected person or a relative in his/her same genetic line; by reasons of public health; to establish, exercise or defend a legal complaint. People have the right to know, check and cancel the medical data that organizations have about themselves. Some countries such as Spain have also defined extra regulations defining guidelines about the adequate organizational and technical measures that must be taken along the following aspects:

- *Separation of data*: as a general rule, the design of data structures, procedures, and allowed selective accesses must be such that it allows the separation of a) identifiers and data related to person identity, b) administrative data, c) medical data, and d) genetic data. Such separation must ensure that no unauthorized person can connect the identity of the patient with his medical or genetic data.
- *Usage control*: data must be protected against any kind of unauthorized processing, including the unauthorized alteration and communication of such data, introducing identification and authentication mechanisms for the persons and institutions with authorization.
- *Memory and telematic transmissions*: unauthorized inputs, queries, modifications or deletions of the data while they are stored in the computer memory of the information system, as well as while the data are sent through the network from a computer to another, must be avoided.
- *Facility access and data media control*: no unauthorized person must be neither able to access to facilities where personal data are stored or processed nor read, copy, alter or take away the data media.
- *Data loss protection*: all organizational and technical measures must be taken in order to protect the data against accidental or illegal destruction, and against accidental loss.
- *Access and data input logging*: the system must guarantee that it will be possible to establish and verify a posteriori when and who accessed the system, and which information has been entered.

3.2 Privacy and Provenance

The above mentioned organizational and technical measures help to protect the privacy of the patients in usual healthcare information systems. In these systems the patient and the medical data are stored in EHCR management systems and transmitted between these systems. The separation of data and the different kinds of access control techniques protect the identity of the patient. The anonymity of medical data allows controlled and irreversible disclosure for different purposes mentioned earlier. In these systems the data is completely under the control of the agents comprising the distributed system and data sharing is controlled by the agents.

When we want to increase trust in data and to increase the quality of medical services in distributed medical applications by introducing provenance concepts, we introduce new privacy risks as well. We introduce an additional agent type into the system: the provenance agent. In order to be able to trace how a particular result has been arrived at by identifying the individual and aggregated services that produced a particular output, the agents of the system must entrust information to the provenance store. This way healthcare agents give up part of the control over the data and the autonomy of the healthcare information is shared with the provenance store which is then able to link data and the workflow pieces that generated the data.

One of the problems of healthcare information systems is that there are information islands. While the healthcare data exchange standards help the information exchange between these islands, the provenance system helps the integration of the islands which raises additional privacy risks.

As mentioned before, the purpose of using a provenance system in the OTM application is to be able to trace back each of the allocation processes that happened whenever an audit is needed to verify, e.g., the chain of decisions made for each donation, or the compliance of an allocation with respect to the related regulations. However there may be a conflict between provenance and privacy. While for provenance we need as much information as possible about the whole process (*who* did *what* and *when*) to be able to trace back all that has happened, for privacy we need to restrict as much as possible the information available in order to avoid identification of patients and practitioners by unauthorised users.

The use of distributed provenance stores to register all relevant information in a distributed medical information system poses two main risks:

- *cross-link risk*: the risk that unauthorised users are able to link some piece of medical data with an identifiable person by cross-linking information from different sources.
- *event trail risk*: the risk to be able to identify a person by connecting the events and actions related to that person (e.g., the hospitals he has visited in different countries).

In the following we will discuss these privacy issues of the application of provenance.

4 Protecting Privacy in the OTM Application of the Provenance Project

Comparing the two main risks identified above, the cross-link risk is more considerable than the event trail risk. In order to identify a person by exploiting the event trail risk, information not available in the healthcare information system (the places where he lived) has to be matched with the information in the healthcare information system. This is still a risk, but it requires more effort and information to exploit, than the cross-link risk which can be exploited using information available only in the healthcare information system. Because of these reasons currently we focus on the cross-link risk.

We introduced two techniques to reduce the cross-link risk: a) we do not put medical data in the provenance store that can be easily used to identify the patient, and b) we anonymise the patient data. These techniques are discussed in the following.

4.1 Medical Data and the Provenance Store

Storing medical information in the provenance store poses two problems: reduced access control to the replicated information, and cross-linking of the information.

The provenance store is out of the access control of the healthcare information system, it is a 3rd party service from the point of view of the clients. In healthcare information systems clients can specify the different distribution rules when they exchange medical data. The provenance store is a general system and at this time, there is no specific way to tell to the provenance store what data are accessible in what circumstances, i.e. who, when, how (read/write) can access the data when p-assertions are made. Although the healthcare information system can put these access control information into the provenance store, but there are no built-in mechanisms to enforce these access control rules, therefore they are very easy to breach.

The cross-linking of information is much easier in the provenance store (which has a set of tools to cross-link information to build execution traces), then in the distributed healthcare information system (where data in the distributed system can be accessed and collected only with well specified access control and with declared aims). This raises the questions: Can we put medical data into p-assertions which will be stored within provenance stores? Can we put person identifiers into p-assertions? Is it enough if we anonymise patients in p-assertions? How can we safely anonymise the patient? If we do not store medical information in the provenance store, then how can we retrieve the provenance of medical data?

When mapping the provenance architecture to the OTM application, we decided not to store sensitive medical data in the provenance store, but only references to such data. In addition public identifiers of patients are not stored in the provenance store, only anonymised identifiers generated from the public identifier are used. This way the provenance store contains only the linkage and the skeleton of the provenance of the medical data, and the healthcare data can be laid on the skeleton by retrieving it from the healthcare information system when needed. The retrieval is done by an EHCR system which is completely under the control of EHCR access rules. With this approach we keep the same privacy degree of medical data as in the original system. Moreover we also minimalise the amount of transferred data.

4.2 Anonym Identity in the Provenance Store

One might think that if we do not store medical information about patients in the provenance store, then there is no need to anonymise the patients and we can use real patient identifiers, because no medical information can be inferred on the patient. However this is not the case. Even the fact that the patient was treated, can be sensitive information (it may increase the event-trail risk mentioned in section 3.2). Moreover the reference to the place where the medical data of the treatment was carried out may contain sensitive information. The type of institution can reveal the type of medical intervention. For example if the institution is specialised on heart diseases, then the reference to this institution reveals that the patient was treated with heart problems. Therefore at least the patient identity has to be anonymised. This is similar to the method of EHCR systems and yields similar privacy degree.

The anonymisation process has to satisfy several requirements. If two sets of p-assertions are related to the same patient, then there should be a way to link anonymised patient identifiers referring to the same patient in the different sets of p-assertions. The anonymisation procedure should be irreversible: nobody should be able to tell the real identity of the patient by knowing the anonymised identifier. As a consequence of this, no component in the system should store the real identifier of the patient and its anonymised identifier together.

In the OTM application the EHCRS systems applies case identifiers (identifiers created at run-time) as tracers to make connections between sets of p-assertions. The case identifier is anonymous, because it does not contain the identity of the patient.

When we want to connect different cases, we have two choices: a) we store two or more case identifiers in any of the p-assertions and define relationships between them, or b) we use a global anonymous tracer of the patient and connect each case identifier to that anonymous tracer in p-assertions. The first solution corresponds to the current practice where the doctor knows (for example from the patient) that the current case is related to a previous one. In this case the doctor connects the cases with a p-assertion. The second solution helps to connect those cases as well, that are not explicitly known.

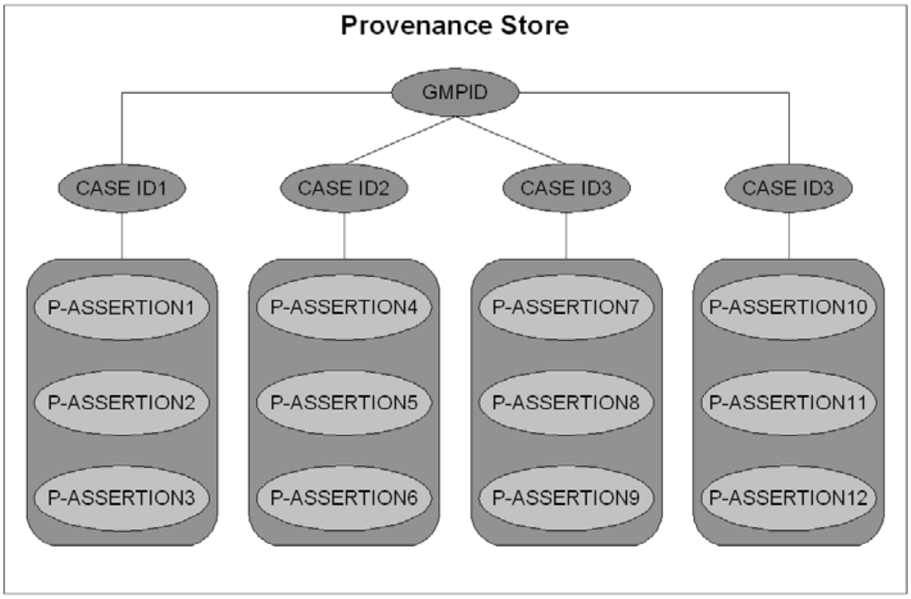

Fig. 2. Linking of p-assertions in the provenance store with Global Medical Patient Identifier (GMPID). The p-assertions related to the same case are linked together with anonymous case identifiers. Cases related to the same patient are linked together with the anonymous GMPID.

In the second solution (shown in figure 2) we send the case identifier and the public identifier of the patient to an authorisation agent who is responsible for all authorisations in our systems. Currently the authorisation is based on username and password, but it can be made more sophisticated later. The authorisation agent generates from the public identifier of the patient (such as national insurance number) an anonymous tracer, called Global Medical Patient Identifier (GMPID[3]) and makes a p-assertion in the provenance store connecting the case identifier with the GMPID. This way all case identifiers, which are about the same patient, will be connected to the GMPID of the patient in the provenance store. When we query the p-assertions that relate to the same patient, we can use any of the case identifiers, because case identifiers related to the same patient are linked together. This way the authorisation agent connects the different identity domains together.

When we want to retrieve the medical history of the patient, then we ask from the provenance store where and when the patient was treated, and then we have to query all the other information directly from the hospitals where the patient was treated because they are not stored in the provenance store.

4 Discussion

The novel concepts and techniques under development by the Provenance project may increase the quality of medical services by providing a unified view of the medical history of patients. The organ transplant management application of the Provenance project demonstrates the application of provenance in healthcare information systems. This demonstration application raises new privacy issues which we investigated in this paper.

We have identified two privacy risks when a provenance system is introduced in healthcare applications. We reasoned that the most critical of these risks is the cross-link risk. We proposed methods to eliminate this risk.

Currently we are working on the implementation of the organ transplant management application with provenance extension and we are implementing the privacy protection methods outlined in this paper. Elimination of the event trail risk needs further investigations.

Because the Provenance project is the first project to investigate provenance in service oriented architectures, the application of provenance to healthcare information systems is novel and the privacy issues investigated and methods proposed in this paper are novel as well.

5 Acknowledgements

This work has been funded mainly by the IST-2002-511085 Provenance project. Javier Vázquez-Salceda's work has been also partially funded by the "Ramón y Cajal" program of the Spanish Ministry of Education and Science.

[3] We do not detail here how the GMPID is generated. We assume that the algorithm satisfies the general anonymisation rules: it is not public and not reversible. The patient identifier and the GMPID are never stored together, and they are present at the same time temporarily only in the authorisation agent when the GMPID is generated.

All the authors would like to thank the Provenance project partners for their inputs to this work. The partners of the Provenance project are IBM United Kingdom Limited, University of Southampton, University of Wales, Cardiff, Deutsches Zentrum fur Luft- und Raumfahrt s.V, Universitat Politècnica de Catalunya, MTA SZTAKI. Some of the concepts, especially those related to the provenance architecture are conceived by other project partners, and some of the results of the authors in the paper originate from interactions between the consortium members.

References

1. J.L. Nealon and A. Moreno, editors. Applications of Software Agent Technology in the Health Care Domain. Whitestein Series in Software Agent Technologies. Birkhäuser Verlag, Basel, 2003.
2. The EU Provenance Project Enabling and Supporting Provenance in Grids for Complex Problems (IST 511085), http://www.gridprovenance.org/, Project partners are: IBM United Kingdom Limited, University of Southampton, University of Wales, Cardiff, Deutsches Zentrum fur Luft- und Raumfahrt s.V, Universitat Politecnica de Catalunya, MTA SZTAKI
3. Maldonado Segura, J.A.; Robles Viejo, M.; Cano Cerviño, C. Integration of distributed healthcare information systems: application of CEN/TC251 ENV13606. 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Proceedings of the 23rd Annual International Conference of the IEEE 2001(4), pp: 3731 -3734. ISBN: 0-7803-7213-1. IEEE Catalog Number: 01CH37272C, Istambul, 2001.
4. CEN/TC251 WG I.: Health Informatics-Electronic Healthcare Record Communication- Part 1: Extended architecture and domain model, Final Draft prENV13606-1 (1999).
5. Groth, P., Miles, S., Tan, V., Moreau, L.: Architecture for Provenance Systems, version 0.4, Technical Report, ECS, University of Southampton, 2005, http://eprints.ecs.soton.ac.uk/11310/
6. Munindar P. Singh and Michael N. Huhns. Service-Oriented Computing: Semantics, Processes, Agents. John Wiley & Sons, Ltd., 2005.
7. Robin Milner. Communicating and mobile systems: the π-calculus. Cambridge University Press, 1999.
8. Nancy Lynch. Distributed Algorithms. Morgan Kaufmann Publishers, December 1995.
9. J. Vázquez-Salceda, U. Cortés, J. Padget, A. López-Navidad, F. Caballero. "The organ allocation process: a natural extension of the Carrel Agent Mediated Electronic Institution". AI Communications vol. 16 num. 3, pp. 153-165. IOS Press, 2003
10. J. Vázquez-Salceda, J.A. Padget, U. Cortés, A. López-Navidad, F. Caballero. "Formalizing an Electronic Institution for the distribution of Human Tissues". Artificial Intelligence in Medicine vol. 27 issue 3 , pp. 233-258. Elsevier, March 2003.
11. Bellifemine, F., Poggi, A., Rimassa, G.: "JADE - A FIPA-compliant agent framework", In Proc. of the Fourth International Conference and Exhibition on the Practical Application of Intelligent Agents and Multi-Agents (PAAM'99), London, UK, (1999) pp. 97-108., The Practical Application Company Ltd 1999.
12. Steven Willmott, Félix Oscar Fernández Peña, Carlos Merida Campos, Ion Constantinescu, Jonathan Dale, and David Cabanillas. Adapting Agent Communication Languages for Semantic Web Service Inter-Communication. The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), Compiègne, France, September 2005, pp. 405 - 408, ISBN:0-7695-2415-X, IEEE Computer Society, 2005

13. The Foundation for Intelligent Phisical Agents. FIPA Specifications, 2000. http://www.fipa.org/repository/fipa2000.html.

14. Directive 95/46/CE of the European Parliament and of the Council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and of the free movement of such data, October 1995.

15. Recommendation No. R(97)5 of the Committee of Ministers to Member States on the Protection of Medical Data. Council of Europe, 13 February 1997.