Title: MIAS-Grid: An EU Provenance Case Study of Security Issues in a Provenance System

Authors: *Simon Miles, Victor Tan* University of Southampton
Daniele Turi, Katy Wolstencroft, Jun Zhao University of Manchester

Version: *1.0*

Date: *29th May 2006*

Status: *Public*

Summary

This document describes a scenario developed as part of the myGrid project and a sample provenance question within that scenario. The focus here is not to illustrate interesting examples of acquisition or utility of provenance; rather it seeks to examine the security issues that could arise in a provenance architecture and a way in which they could be addressed. internal controls that can be subjected to internal and external audit processes. This paper briefly describes how Provenance may provide a solution to financial auditing as an example of a common internal control process.

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

Members of the PROVENANCE consortium:

- IBM United Kingdom Limited United Kingdom
- University of Southampton United Kingdom
- University of Wales, Cardiff United Kingdom
- Deutsches Zentrum fur Luft- und Raumfahrt e.V. Germany
- Universitat Politecnica de Catalunya Spain
- Magyar Tudományos Akadémia Számítástechnikai és Automatizálási Kutató Intézet Hungary

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

1 Introduction

This document describes a scenario developed as part of the myGrid project and a sample provenance question within that scenario. The focus here is not to illustrate interesting examples of acquisition or utility of provenance, rather it seeks to examine the security issues that could arise in a provenance architecture and a way in which they could be addressed. The analysis should be a useful source for any project, including Provenance and myGrid, that requires the securing of a provenance infrastructure. We do not apply the Provenance methodology in it entirety as we choose to only focus on aspects of the sample provenance question that are particularly security relevant. Extensive application of the methodology is provided in the accompanying document: **myGrid: An EU Provenance Case Study**, where more interesting and in-depth provenance questions are explored and discussed.

2 Overview of Application

In the MIAS-Grid project, medical image analyses can be run as workflows. The images in the study are captured in particular format (DICOM), which supports the annotation of the image with metadata about the way that the image was acquired as well as clinical and personal data about the patient. These images, along with their accompanying metadata are subsequently stored in a designated store, which can be same as the store holding process documentation about the workflow itself.

The originating store is located within the hospital premises in which these images are acquired. Researchers interested in the images can make requests to appropriate hospital staff with access to these stores who subsequently transfer them to them. The primary security considerations in the application environment are:

1. Due to legal requirements on anonymity of patient data, metadata that can potentially identify a patient within the image will need to be anonymised.
2. There may be additional requirements restricting the way these images are manipulated in a workflow by researchers using them.

2.1 Physical Distribution

The deployment of people and resources in the scenario is as follows (Figure 1). The researchers utilizing the image data and the hospital staff managing access to it at the point of capture are all potentially physically separate and additionally, located in different security domains. Both researchers are assumed to have access to workflow enactors within their respective domains (not shown in diagram). Results from the workflow are stored in data stores within their domains as well. In addition, to answer the provenance question that we present later, researcher 2 may need to perform access across different security domains.

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

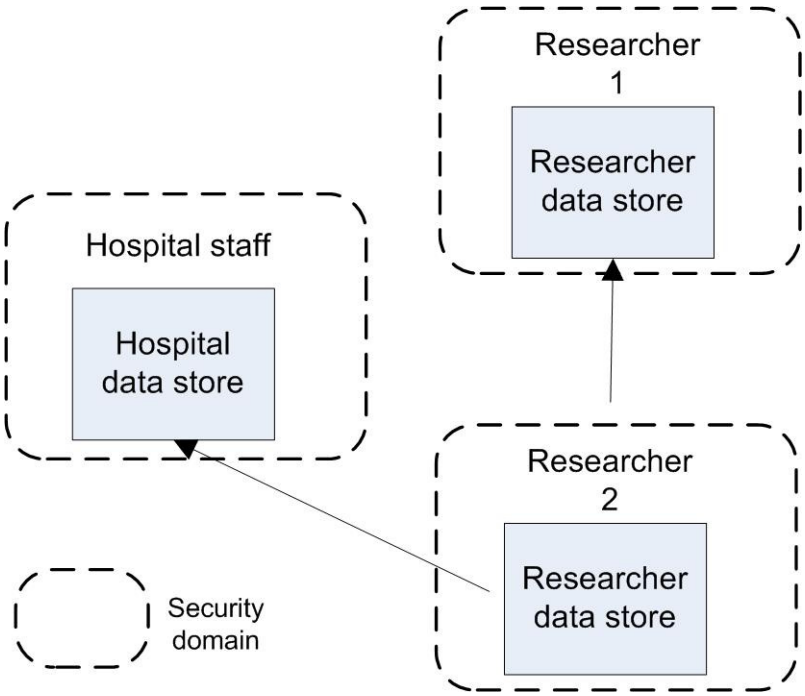


Figure 1: The high-level actors in the scenario and their distribution

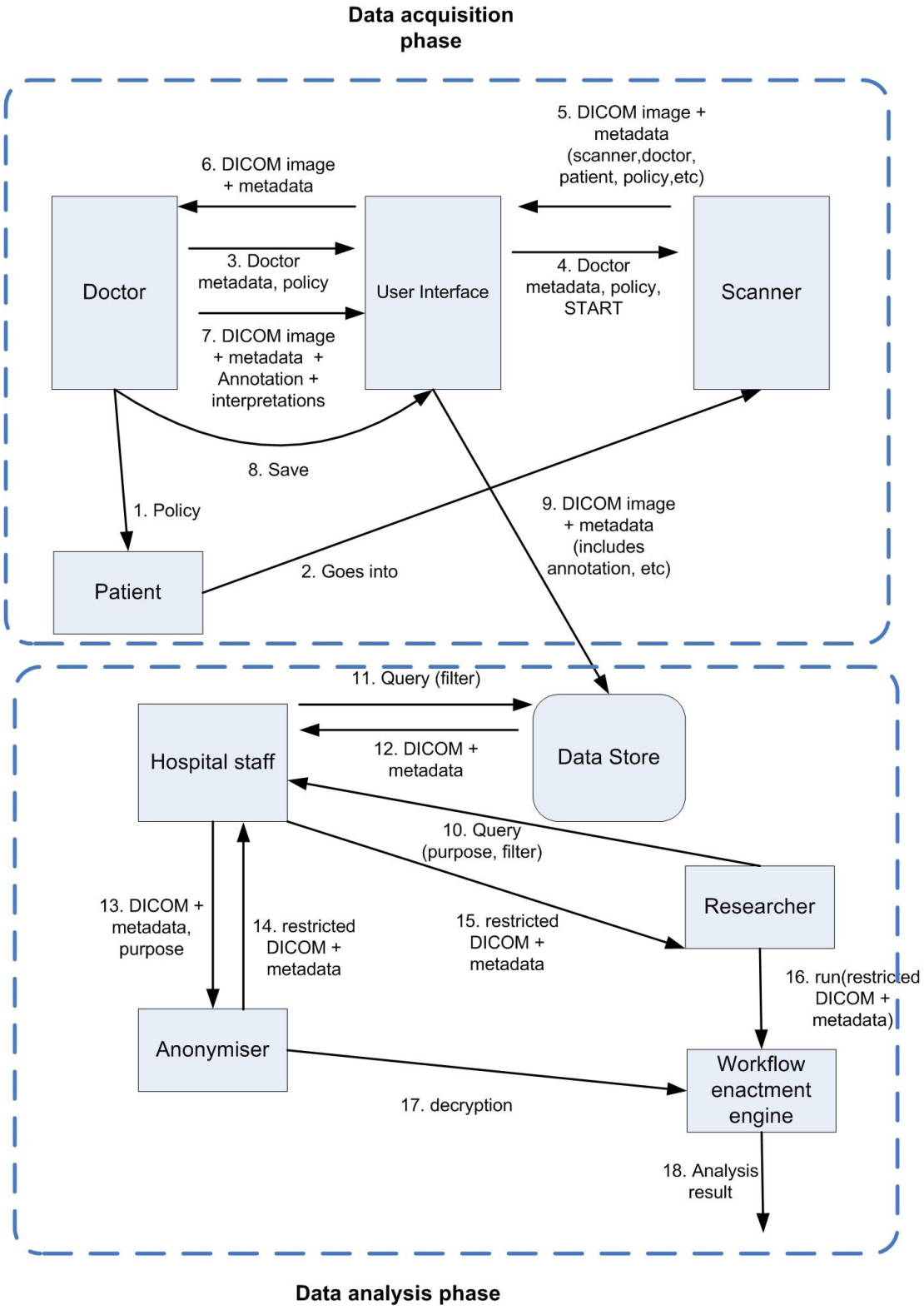
2.2 Workflow

The workflow is depicted below in Figure 2. At the time when a scan is performed, the patient is requested to sign a consent form indicating his / her consent to some policy restricting the way that his/her details and medical data are processed or accessed. This policy can also form part of the metadata included along with the medical image. All of this is detailed in sequence arrows 1 – 6 (Figure 2). Additionally, the doctor may choose to annotate the image with further interpretations of his or her own before saving it to a local data store in the hospital (sequence arrows 7 – 9). We can classify all of the interactions so far as the data acquisition process. Here, sensitive data (e.g. patient metadata) is not anonymised as it needs to be readily available to the doctor, however potentially process documentation of the interactions might need to be anonymised.

The remaining interactions (sequence arrows 10-18) comprise the data analysis phase which essentially involves a researcher making a request for image data that fulfills some criteria (as specified by a filter), and a hospital staff returning the requested data and its associated metadata along with the possibility for some security restrictions (e.g. encryption) applied on the requested data to ensure that the metadata is manipulated in some pre-agreed legally binding manner. The results returned could (sequence arrow 15) could have the image data (which is potentially very large) being replaced with references to the hospital data store in order to address scalability concerns, while its associated metadata could be passed on directly to the researcher. The researcher then passes the requested image results to a workflow enactment engine. The execution of the workflow may optionally require further interaction with

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

the anonymiser (sequence arrow 17) to ensure that security restrictions applied on the returned results (sequence arrow 15) are enforced appropriately. As an example, the anonymiser could pass back keys to decrypt relevant parts of the image data if the workflow engine is manipulating the image data in a pre-agreed valid manner.



The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

Figure 2: The actors, data and data flow making up the workflow

2.3 Provenance Questions

An example of a workflow run using these image data is to analyze the change of an anomaly in a patient (e.g. a cancerous tumour) over a period of time. This requires an initial process known as registration of the image data, where images belonging to a anatomical part of the patient that has been scanned over a period of time are aligned with each other using some image morphing technology. There may be various registration techniques used; it is important to ensure that the registration works in a way that doesn't distort the image so that, as an example, a tumour appears to be growing when it really isn't. A researcher intending to test out a new registration algorithm could ask a provenance question of this form:

PQ1: *What were the image inputs that were used in an existing registration algorithm ?*

2.4 Information Items

The information items required to answer provenance question 1 would be the image inputs that another researcher had run using a different registration algorithm in order to make appropriate comparisons with the workflow run using the new algorithm. This would be obtained via the process documentation of interaction 10 or 15 (Figure 1) and potentially, process documentation of other interactions in the data acquisition phase.

3 Security Issues

As discussed earlier, the two main security considerations for the application environment include anonymising patient metadata and changes in policies restricting the use of such metadata over time. In addition, as part of answering *PQ1*, researchers from a particular security domain may require access to images as well as process documentation stored or used by researchers from a different security domain due to physical distribution considerations. This would potentially require federation of identity across these multiple domains. Equivalently, researchers will also require appropriate access to the original images in the hospital store. In order to address the security issues of access control and federated identity, we describe a logical security architecture (Figure 3) for a provenance store developed as part of the Provenance project. Although our description refers to a provenance store, we will assume that this architecture is generic enough to be equally applicable to stores holding image data/metadata in the application environment

An actor (potentially a researcher) would access the provenance store via a prescribed set of provenance store interfaces (step 1), presenting the relevant security credentials (such as X509 certificates or username/password pairs) for authentication purposes. The verification of these credentials is undertaken by the identity validator (step 2),

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

which extracts the identity information from these credentials and produces an internal representation of the actor identity using information from a system administrator defined internal representation list (step 3). This internal identity is formatted into an appropriate access request and sent off to the authorisation engine (step 4).

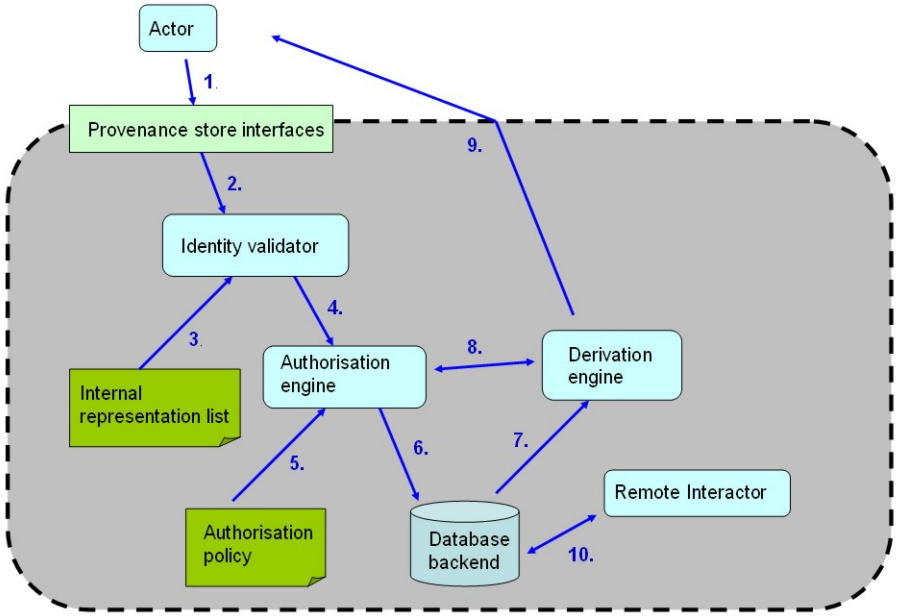


Figure 3: Security architecture for a provenance store

The authorisation engine performs the actual access control to the underlying backend store holding the process documentation. It determines whether the access request is permissible on the basis of authorisation statements in the authorisation policy (step 5); this can involve, for example, associating roles/identities with operations that can be performed and the subset of process documentation they can be performed on. If the operation is permitted, the results are retrieved and sent to the derivation engine (step 6 and 7). The task of this component is to perform any other necessary transformations on results to be returned to the actor, and may potentially involve filtering of some sort on the basis of authorisation statements in the authorisation policy (step 8). The final set of results are then returned to the actor (step 9). Occasionally, the set of initial results returned from the backend store (step 10) may indicate that further interaction is required with external parties to obtain a complete result set satisfying the original access request. In such a situation, the remote interactor component engages in the required interaction with applications in other security domains (in particular, other provenance stores) in order to obtain the required results pertaining to the current access request. This is particularly relevant for the case when identity is federated, which we discuss next.

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

Federation of identity will require that identity information be shared between different security domains. A common way of accomplishing this is to employ a trusted third party to issue credentials that are subsequently acceptable for authentication purposes by a group of different security domains that have agreed in advance to federate their identity information. Credentials issued in this manner will be in a format understood by the identity validator in all these different domains, and may contain relevant security assertions by the trusted third party about the actor that the credential is issued to.

A possible interaction sequence is shown in Figure 4. Here, actors intending to access any provenance store will first authenticate themselves to the security token service (the trusted third party), which in issues them with credentials containing the required assertions about their access rights. These tokens can then be subsequently presented to provenance stores in different domains that share a mutual trust in the security token service. In the event that a provenance store needs access to another provenance store in a different domain in order to return a set of process documentation results that satisfy an actor's access request, it can elect to use the token initially presented by the actor in order to authenticate to this remote store.

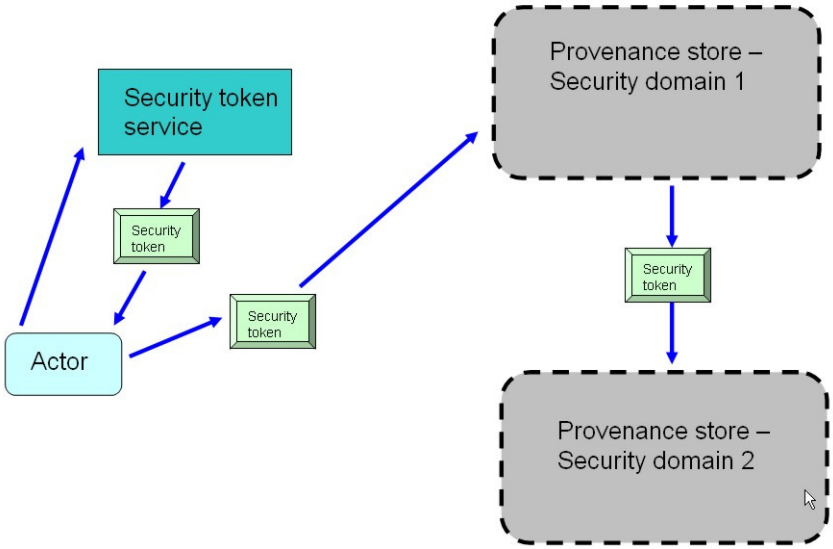


Figure 4: Federation of identity

3.1 myGrid security architecture

The myGrid project has developed a security architecture to safeguard access to the intermediate data as well as metadata generated in a myGrid workflow run (both of which are currently stored in the same logical backend). Figure 5 demonstrates the architecture, and its components fulfil all of the functionality required in the logical security architecture that we described in Figure 3. We briefly describe its functionality below.

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

Access is achieved through a portal or client GUI that authenticates the user via lsid credentials. This is mapped to user specific attributes in the SAML format after consultation with an internal user directory and an authorisation authority. The policy enforcement point then decides whether or not to permit the request in consultation with the policy decision point. The latter component uses a policy outlined in the XACML format to express the relevant authorisation statements. If the request is permitted, then the requested data / metadata is then returned to the requesting user.

The user directory corresponds to the internal representation list of the provenance store security architecture, while the authorisation authority maps to the identity validator. The policy enforcement and policy decision point enforces functionality encapsulated within the authorisation engine of the provenance store security architecture, while the XACML policy becomes effectively the authorisation policy consulted by the authorisation engine. If a provenance store were to be plugged into myGrid, the XACML policy could ostensibly replace the authorisation policy of the provenance store so that access control could be enforced in an identical manner to both the myGrid store and the provenance store

For the case of multiple security domains, the authorisation authority could be externalized as an independent component that would perform the functionality equivalent to a trusted third party in the logical security architecture of the provenance system. A user would initially authenticate to the authorisation authority in order to obtain a signed assertion containing, in this example, SAML attributes that it could then present to all myGrid stores in different security domains for which is wishes to request data / metadata from.

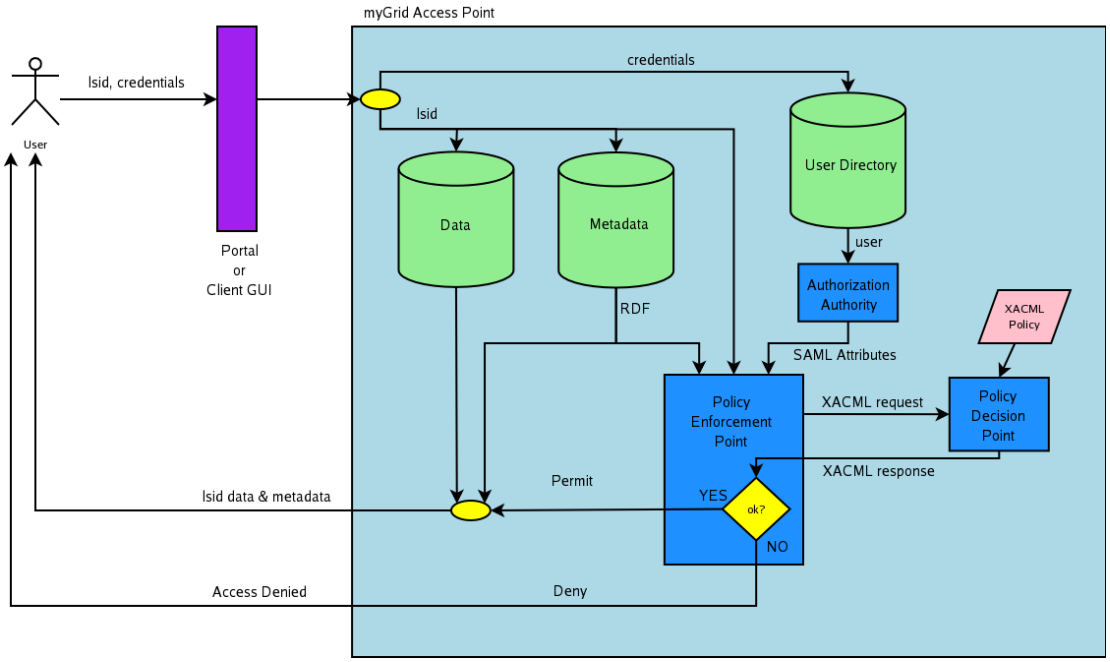


Figure 5: myGrid security architecture

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme