Authenticity Management in Long Term Digital Preservation of Medical Records

Silvio Salza
Università degli studi di Roma "La Sapienza"
Dipartimento di Ingegneria informatica
via Ariosto 25, 00185 Roma, Italy
+39-06-77274-015
salza@dis.uniroma1.it

Maria Guercio
Università degli studi di Roma "La Sapienza"
Dipartimento di Storia dell'arte e spettacolo
piazzale Aldo Moro 5, 00185 Roma, Italy
+39-06-4967-002
maria.guercio@uniroma1.it

ABSTRACT
Managing authenticity is a crucial issue in the preservation of digital medical records, because of their legal value and of their relevance to the Scientific Community as experimental data. In order to assess the authenticity and the provenance of the records, one must be able to trace back, along the whole extent of their lifecycle since their creation, all the relevant events and transformations they have undergone and that may have affected their authenticity and provenance and collect the Preservation Description Information (PDI) as categorized by OAIS. This paper presents a model and a set of operational guidelines to collect and manage the authenticity evidence to properly document these transformations, that have been developed within the APARSEN project, a EU funded NoE, as an implementation of the InterPARES conceptual framework and of the CASPAR methodology. Moreover we discuss the implementation of the guidelines in a medical environment, the health care preservation repository in Vicenza Italy, where digital resources have a quite complex lifecycle including several changes of custody, aggregations and format migrations. The case study has proved the robustness of the methodology, which stands as a concrete proposal for a systematic and operational way to deal with the problem of authenticity management in complex environments.*

Categories and Subject Descriptors
H.3.2[**Information storage and retrieval**]: Information storage.

General Terms
Management, Documentation, Standardization, Legal Aspects.

Keywords
Authenticity, digital preservation, e-health, medical records.

1. INTRODUCTION
Authenticity plays a crucial role in the management and preservation of medical records. In most countries all the documentation related to the citizens' health, including of course digital files, has to be preserved for an indefinite period of time, some series potentially forever, and the continuing ability of assessing the authenticity and the provenance of the records is therefore an important issue both for the legal value of data, to properly allocate the responsibilities, and for the scientific community that considers the results of medical tests and medical reports as important experimental data.

The problem of managing the authenticity of digital resources in this as well as in other environments has been addressed, as an important part of its activities, by the APARSEN project [1], a Network of Excellence funded by the EU (2011-2014) with the goal of overcoming the fragmentation of the research and of the development in the digital preservation area by bringing together major European players. The research activity we present here is the prosecution of the investigation carried out within previous international projects, notably the conceptual framework defined by InterPARES [6] and the methodology proposed by CASPAR [5]. More specifically, the APARSEN proposal [2] has stressed the need to take into account the whole *digital resource lifecycle* to model the preservation process, as defined by *ERMS (Electronic Records Management Systems)* recommendations, and has defined *operational guidelines* to:

- conveniently trace (for future verification) all the events and transformations the digital resource has undergone since its creation that may have affected its authenticity and provenance;
- collect and preserve for each of these events and transformations the appropriate evidence that would allow, at a later time, to make the assessment and, more precisely;
- develop a model of the digital resource lifecycle, which identifies the main events that impact on authenticity and provenance and investigate in detail, for each of them, the evidence that has to be gathered in order to conveniently document the history of the digital resource.

The model and the guidelines that we have proposed have been successfully put to test on experimental environments provided by the APARSEN project partners. These case studies, which are documented in a project deliverable [3], provided important feedback and have proved on the field the substantial robustness of the proposal.

This paper relates about a case study in the medical environment, the repository of the health care system in Vicenza (Italy), a rather complex case since along the DR lifecycle there are several changes of custody that involve, beside the preservation repository, several keeping systems, some of them geographically distri-

* Work partially supported by the European Community under the Information Society Technologies (IST) program of the 7th FP for RTD - project APARSEN, ref. 269977.

buted in the district. Moreover there are several types of DRs (diagnostic images, medical reports etc.), each one with a distinct workflow.

The case is also interesting because the repository must comply both with the international standards and with the rather complex Italian legislation on the creation, keeping and preservation of electronic records, and with additional specific rules for the keeping of medical records, based on the widespread use of digital signatures and certified timestamps. The implementation has been strongly oriented toward standardized solutions based on XML schemas) and a common dictionary based on PREMIS.

The paper is organized as follows. In Section 2 we present the Vicenza health care preservation system that has been the object of our study, and we also provide some details on the procedures mandated by the Italian legislation on long term preservation of digital records. Section 3 and 4 are devoted to present the digital resource lifecycle model and the authenticity management policy, as well as the operational guidelines that we propose to implement the model in specific environments and to guide the process of designing an effective authenticity management policy. In Section 5 these guidelines are applied to model the Vicenza health care system, and this leads to the formalization of the authenticity management policy and to the definition of the Authenticity Protocols. Finally, concluding remarks are given in Section 6.

2. THE VICENZA REPOSITORY

2.1 The preservation infrastructure

The preservation infrastructure of the public health care system unit ULSS6 in Vicenza is based on the system Scryba, implemented and distributed by the Italian company MEDAS Srl, that has been designed according to the basic principles of the OAIS reference model and with additional specific features intended to make it compliant with the Italian regulations on long term digital preservation. Scryba is a modular system based on a set of functionalities that can be configured to meet the specific requirements that arise in different environments. Up to now it has been deployed as the core element of several digital preservation repositories in Italian hospitals.

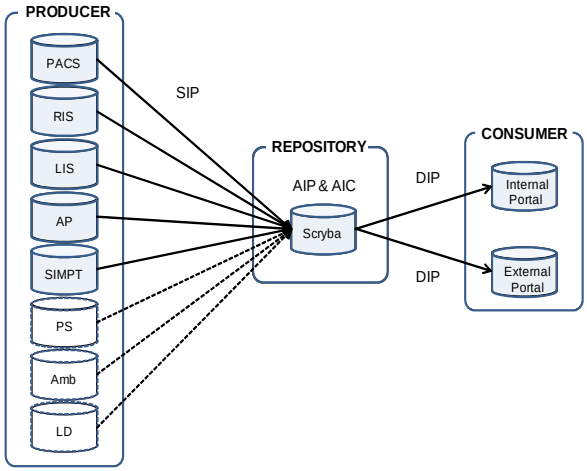

Figure 1. The preservation infrastructure.

In ULSS6-Vicenza the preservation infrastructure is interfaced with a variety of producers that deliver several different kinds of digital resources, mostly diagnostic images, test results and medical reports. The actual interface of the preservation system on the producers' side is towards a set of departmental systems that collect the digital resources for peripheral devices and satellite systems, such as digital imaging devices, workstation attended by physicians etc.

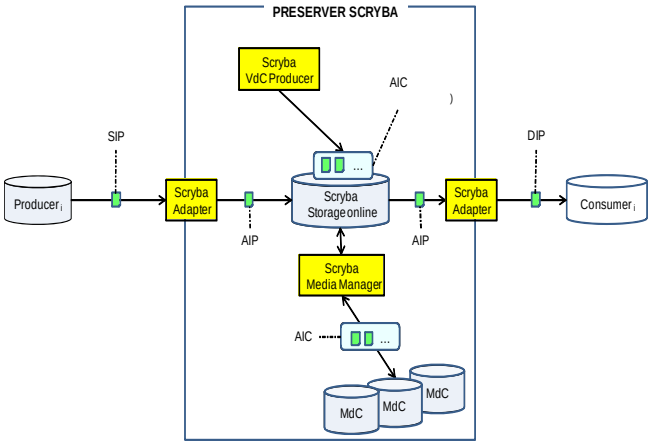

Figure 2. The Scryba preservation system.

The above mentioned departmental systems also act as short-term repositories and provide physicians and medical staff with immediate access to test results and reports. According to the Italian regulations, all medical records are delivered to the long-term preservation repository as soon as they are created and signed. Therefore, shortly after its creation and signature, each digital resource is preserved in two distinct copies, one in the departmental systems for consultation in the short period, and the other one in the *LTDP (Long Term digital Preservation)* repository as an official record.

The LTDP system can be accessed by consumers by means of two distinct interfaces:

- the internal portal which is used by physicians and medical staff, and allow authorized persons to get web access to the whole content of preserved digital resources;
- the external portal that provides citizens (or their authorized representatives) access to their own medical records.

Access to both interfaces requires strong authentication, according to the regulations on the privacy of medical records. An overview of the system is given in Figure 1 where the different kinds of producers are represented. Currently five different producers are supported, including diagnostic images in DICOM format (PACS) and medical reports of various kinds (RIS, LIS, AP). Support for additional producers is currently being implemented.

2.2 The Scryba preservation system

The Scryba system is based on the principles of the OAIS reference model and with additional specific features intended to make it compliant with the Italian regulations on long term digital preservation. The high level structure of the system is shown in Figure 2.

The system has a modular structure which is based on a core structure whose main functions are the management of the *AIPs (Archival Information Packages)*, the related transformations (aggregation, format migration) and their secure storage. Additional modules, called *adapters*, are deployed to manage the communication with the external world, i.e. the *producers* on one side and the *consumers* on the other side.

Adapters are implemented on a base structure that can be customized to meet the specific requirements of different producers and consumers. Scryba Adapters work in several ways (DICOM protocol, HL7 msg, IHE XDS.b profile, or specific host oriented web-services) to match all host communication protocols. The management of the AIPs and their secure storage are compliant with the OAIS reference model, but strongly influenced by some peculiarities of the Italian national regulations. According to these regulations, the preservation process is based on collecting the digital resources to be preserved in large batches, named *Preservation Volume (PV)*, which are the actual object of the preservation process and must undergo a well-defined formal procedure that includes digital signature, certified time stamping of the PV as well as periodical controls and possibly the generation of new copies on different storage medias.

The Italian regulations require also to produce a given number of *BCs (Backup Copies)* for every PV and to store them in different locations according to a predefined and formally stated schema.

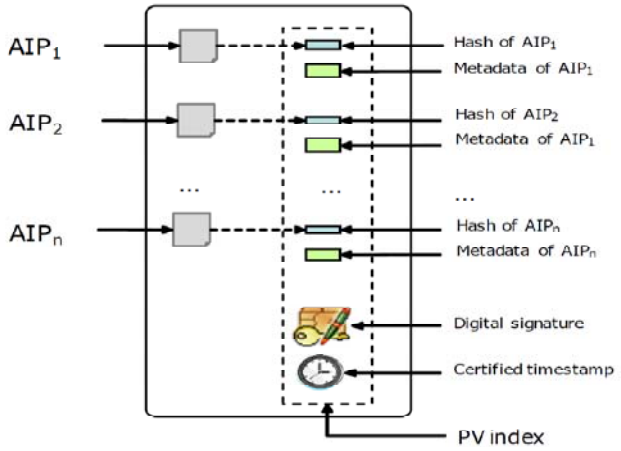

Figure 3. Structure of a Preservation Volume.

The structure of a preservation volume is shown in Figure 3. It contains all the aggregated digital resources plus an additional file, the *Preservation Volume index* (PV index), which is compliant to UNI SInCRO, a national metadata standard, digitally signed by person officially in charge of the preservation process (in Italian *Responsabile della Conservazione*) and marked with a temporal timestamp. The PV index is an XML file which contains:

- a hash file for each AIP in the PV;
- a set of metadata for each AIP in the PV;
- the digital signature;
- the certified timestamp.

In order to comply both with the OAIS model and the Italian regulations, the SIPs are ingested as soon as they are delivered to the Scryba system, and an AIP is generated for each SIP, i.e. for every individual study or medical report, and enters immediately the preservation process. On the other hand, a set of AIPs from each producer is periodically aggregated to generate an AIC (Archival Information Collections), an OAIS kind of Information Package that well corresponds to the PV (Preservation Volume) the Italian regulations ask for. In the Scryba system any given PV must contain digital resources of a single type and PVs are closed according to a double criteria:

- time: a PV must be closed before a maximum time since its opening elapses (currently 24 hours);
- size: a PV cannot exceed a maximum size. (currently 1GB).

We shall point out that the aggregation of several digital resources in a single preservation volume only depends on the national regulations and it is not performed to comply with OAIS.

3. THE AUTHENTICITY MODEL

3.1 The digital resource lifecycle

The main principle behind the authenticity management methodology that has been developed within the APARSEN project is that, in order to properly assess the authenticity of a Digital Resource (DR), we must be able to collect the information relevant for preservation and trace back, along the whole extent of its lifecycle since its creation, all the transformations the DR has undergone and that may have affected its authenticity and provenance. With specific reference to the transformations crucial for LTDP, for each of these transformations one needs to collect and preserve the appropriate evidence that would allow, at a later time, to make the assessment, and that we shall call therefore *authenticity evidence*.

Under quite general assumptions, we may consider the DR lifecycle as divided in two phases:

- **Pre-ingest phase**. This phase begins when the DR is delivered for the first time to a keeping system and goes on until the DR is submitted to a *Long Time Digital Preservation (LTDP) system*. During the pre-ingest phase, the DR may be transferred between several keeping systems and may undergo several transformations, and is finally transferred to the LTDP system.
- **LTDP phase**. This phase begins when the DR is ingested by a LTDP system and goes on as long as the DR is preserved. As for the pre-ingest, also during the LTDP phase the DR may undergo several transformations, notably format migrations, aggregations etc. Moreover it may get moved from a LTDP system to another one.

The pre-ingest phase has been introduced as a separate phase from the ingest to represent the part of the lifecycle that occurs before the delivery to the DR of a LTDP system. Collecting evidence for all the transformations the DR undergoes during this phase is of the utmost importance to assess its authenticity.

Each transformation a DR undergoes during its lifecycle is connected to an *event*, which occurs under the responsibility of one or more people, whom we shall call *agents*. A transformation may involve one or several DRs and one or several agents, and produces as a result a set of DRs, possibly new versions of the ones that were the object of the transformations.

Unfortunately, the variety of events that may occur during the DR lifecycle is very large and depends, at least in part, from the specific environment. Nevertheless, it is possible to consider at least a minimal *core set of events*, that includes the most important ones, as well as the ones which are likely to occur in most of the environments in which DRs are produced and managed. The core set, is briefly discussed in the following subsections, and may be considered as a preliminary step towards interoperability in the exchange of authenticity evidence among different keeping and preservation systems.

In our investigation we have considered a reasonable variety of environments, notably natural science data, health care data, social science data and administrative data repositories. As a result of our analysis, we have proposed the core set of events that we briefly outline. For a more complete description one should refer directly to the APARSEN project documentation [2].

3.2 Pre-ingest phase

During its stay in the keeping system the DR may undergo a series of transformations that may affect both its content and the descriptive information associated to it. For instance the DR may go through format migrations (even before it enters the LTDP custody), or it may get integrations of its content and/or of its metadata, or it may eventually be aggregated with other DRs to form a new DR. Moreover, before getting to LTDP, the DR may be transferred, one or several times, between different keeping systems.

The pre-ingest phase includes also the submission of the DR to the preservation repository. The content and the structure of the SIP (Submission Information Package) through which the DR is delivered must comply with a submission agreement established between the system where the DR was kept (i.e. the Producer in the OAIS reference model) and the LTDP system (the OAIS).

In the model, the core set for the pre-ingest phase comprises the following events:

- **CAPTURE**: the DR is delivered by its author to a keeping system;
- **INTEGRATE**: new information is added to a DR already stored in the keeping system;
- **AGGREGATE**: several DR, already stored in the keeping system, are aggregated to form a new DR;
- **DELETE**: a DR, stored in the keeping system is deleted, after its preservation time has expired, according to a stated policy;
- **MIGRATE**: one or several components of the DR are converted to a new format;
- **TRANSFER**: a DR is transferred between two keeping systems;
- **SUBMIT**: a DR is delivered by the keeping system where it is stored (producer) to a LTDP system.

3.3 LTDP phase

This phase begins when the DR is delivered to a LTDP system and goes on as long as the DR is preserved. During this phase, the DR may undergo several kinds of transformations, that range from format migrations to changes of physical support, to transfers between different preservation systems.

According to the OAIS reference model [4], many activities are carried out in connection with each of these events, but we restricted our attention to the sole aspects related to authenticity and provenance of the DR and to the information (authenticity evidence) that has to be gathered and preserved in the PDI (Preservation Description Information), and more specifically in the Provenance, Context and Fixity components.

Analyzing this phase many possibilities have to be considered, as for instance transfer between LTDP systems, which is quite likely to happen in the long run, and changes in the structure of the preserved DRs (integration, aggregation etc.), that routinely happen in the health care sector, since records must enter preservation as soon they are created and still there may be later the need to introduce corrections.

The resulting set of events is then:

- **LTDP-INGEST**: a DR delivered from a producer is ingested by the LTDP system and stored as an AIP.
- **LTDP-AGGREGATE**: one or several DRs stored in different AIPs, are aggregated in a single AIC;
- **LTDP-EXTRACT**: one or several DRs which are extracted from an AIC to form individual AIPs;
- **LTDP-INTEGRATE**: new information is added to a DR already stored in the LTDP system;
- **LTDP-MIGRATE**: one or several components of a DR are converted to a new format;
- **LTDP-DELETE**: one or several DR, preserved in the LTDP system and stored as part of an AIP are deleted, after their stated preservation time has expired;
- **LTDP-TRANSFER**: a DR stored in a LTDP system is transferred to another LTDP system.

3.4 Event templates

When giving the guidelines that should be followed to ensure interoperability on authenticity among keeping and LTDP systems, beside providing a precise definition of the event, the crucial point is to specify which controls should be performed, which evidence should be collected and how it should be structured.

In the model each event of the core set is represented according to a uniform schema, by providing an *event template*:

- the *agent*, i.e. the person(s) under whose responsibility the transformation occurs;
- the *input*, i.e. the preexisting DR(s) that are the object of the transformation, if any;
- the *output*, i.e. the new DR(s) that are the result of the transformation (possibly new versions of input DR(s));
- the *controls* that must be performed when the event occurs on the authenticity and provenance of the input DR(s) and to assess properties of the output DR(s) that are the results of the transformation connected to the event.
- the *Authenticity Evidence Record (AER)*, i.e. the information that must be gathered in connection with the event to support the tracking of its authenticity and provenance.

An event template is therefore a sort of checklist, enumerating all the controls that should be performed and all the authenticity evidence that should be gathered and preserved in order to guarantee an accurate management of the DR authenticity through its lifecycle.

Event templates have been defined in the model under very general assumptions, and therefore have been developed into very comprehensive checklists. That means that in a given specific environment only part of the controls may actually need to be performed and only part of the authenticity evidence that is listed in the AER may actually need to be gathered.

Therefore, the model and the templates should be considered as a very general and detailed reference, that needs accurate customization in each specific environment to get to the definition of an adequate authenticity management policy, a problem that will be addressed in Section 4.

3.5 Authenticity Evidence Records

A crucial part of the event template is the definition of the *Authenticity Evidence Records (AER)*. An AER is specified as a sequence of *Authenticity Evidence Items (AEIs)*, i.e. of the elementary items of information that should be gathered and preserved to document the authenticity and the provenance of the DR.

As the DR progresses along its lifecycle through a sequence of events, an incremental sequence of AERs, that we shall call *Authenticity Evidence History (AEH)*, is collected by the systems where the DR is kept or preserved, and strictly associated to it.

From a practical point of view, an authenticity evidence record is a structured set of information, according to our proposal an XML

file of predefined structure, which is strictly related to a given event. At any given stage of its lifecycle a DR brings with it, as part of its metadata, a (temporally) *ordered sequence* of such records, to document all the transformations the DR has undergone and to allow to assess its authenticity and provenance.

Authenticity evidence will follow the DR when it is transferred between different systems, and will accompany it along all its lifecycle. Thus, to ensure interoperability, it is necessary to standardize the way the authenticity evidence is collected and structured. To this purpose existing standards should be accurately considered, as for instance the Open Provenance Model (http://openprovenance.org).

4. THE OPERATIONAL GUIDELINES

Aim of this section is to present the procedure, i.e. the sequence of steps, that should be followed, when dealing with the problem of setting up or improving an LTDP repository in a given specific environment, to get to the definition of an adequate *authenticity management policy*, that is to formalize the rules according to which authenticity evidence should be collected, managed and preserved along the digital resource lifecycle.

4.1 Role of the Designated Community

The concept of *Designated Community (DC)* ("an identified group of potential Consumers who should be able to understand a particular set of information") is central to the OAIS reference model according to which "the primary goal of an OAIS is to preserve information for a designated community over an indefinite period of time". Therefore, as a first step, one should understand what authenticity means to the DC, that is:

- for which purpose and to which extent is the DC interested in being able to assess the authenticity and the provenance of the DRs that are preserved by the OAIS?
- what kind of evidence is considered by the DC as sufficient to make the assessment?

When dealing with an existing LTDP repository, that is analyzed to assess the adequacy of the current practices or to suggest improvements, the starting point should be understanding what kind of authenticity evidence is currently preserved and investigating if the DC actually deems it as sufficient for its purposes.

Altogether the result of this preliminary step is to set up a reference context in order to take appropriate decisions in the following steps of our procedure, i.e. when identifying the lifecycle events to be taken into account and the specific authenticity evidence to be gathered in connection with them.

4.2 Identifying the relevant lifecycle events

The next step is to analyze the workflow of the DRs that are to be preserved in the repository, from their creation on, to identify the lifecycle events that are relevant to the management of the authenticity. When, as it was in the Vicenza case, several DR types and several workflows are identified, the analysis is to be repeated for each workflow.

Once the relevant lifecycle events have been identified, they must be compared and fitted into the *core set events* that we have discussed in Sections 3.2 and 3.3 and that provide a reference and a template on the way authenticity evidence should be gathered and managed. According to our case study experience the core set that we have proposed has proved to be quite a robust choice, in the sense that all the relevant events we have identified could fit well in one of the core set events. However, it is still possible that in a given environment additional events may need to be considered that are specific to that environment.

Then, for each lifecycle event we have identified as relevant, the corresponding event templates should be considered to identify responsibilities and to understand which authenticity evidence should be gathered and which controls should be performed.

As we already pointed out, the templates are quite comprehensive. Therefore, it is often found that part of the authenticity evidence that the templates mandate to collect is not actually collected in the current practices. This does not necessarily mean that the current practices are inadequate: one should instead carefully consider every single missing item of evidence, taking into account the specific needs of the designated community and other details, as for instance the systems involved and their ownership.

For instance, criteria for deciding if an authenticity evidence item should *not necessarily* be recommended as part of the AER could be:

- the item is intended to document a control that is actually performed but not recorded in the AER by a system under the ownership of an organization which is trusted by the DC;
- the item is intended to prove that the integrity of the item has not been affected by the transfer between two systems that are under the ownership of the same organization which is trusted by the DC;
- the item relates to some provenance information which is of no interest to the designated community.

Anyway, besides a few general criteria as above, it is difficult, probably impossible, to give an exhaustive list of specific criteria for deciding whether a given authenticity evidence item should be recommended or not, mostly due to the variety of situations and the complexity of systems.

4.3 Defining the policy and the authenticity evidence records

As a result of the analysis performed in the previous step one should be able to reach, for any given authenticity evidence item in the template of a given lifecycle event, one of the following conclusions:

a) the evidence item is currently collected and preserved and must be part of the AER;
b) the evidence item is not currently collected and preserved, but this information is not necessary according to the definition of authenticity that is accepted by the DC;
c) the evidence item is not currently collected and preserved, but it is not possible to prove that this information is not necessary, and it must therefore become part of the AER.

In all three cases the conclusions should be explicitly and clearly documented. In case c) an improvement of the current practices should be recommended and the information to be collected should be clearly specified, along with the procedure to collect it.

The result of all the above actions is the definition of the *authenticity management policy* that should be adopted by a given LTDP repository to comply with the guidelines we propose and satisfy the needs of its DC. This is made up of the following components:

i. a general statement about the meaning of authenticity to the DC, accompanied by a clear delimitation of the DC and by the

explanation of how the opinion of the DC was actually gathered;

ii. the specification of the lifecycle, and more precisely of the events in the lifecycle that have been identified as relevant to the management of authenticity;

iii. for every relevant event in the lifecycle the definition of the controls corresponding to that event that must be performed and of the AER. together with the specification of the procedures that should be followed to collect it.

4.4 Formalizing authenticity protocols

The next step in implementing the authenticity management policy is the formal definition of the controls that must be performed in connection with each event and of the procedures that must be followed to collect the AER. To this purpose, we propose an implementation strategy which is based on the concept of *Authenticity Protocol (AP)* that has been defined within the CASPAR project [5] as the specification of the procedure that must be followed to assess the authenticity of specific type of DR.

In our methodology the AP becomes the procedure that must be followed in connection with a given lifecycle event to perform the controls and to collect the AER as specified by the authenticity management policy. Accordingly, the execution of the AP corresponding to a give lifecycle event generates the AER that the authenticity management policy mandates to collect in correspondence to that event.

In the formal definition an AP is characterized by:

- *DR type*: the type of digital resource
- *Event type*: the lifecycle event to which the AP corresponds
- *Agent*: the person under whose responsibility the protocol is executed
- *AER*: the AER that is generated by the execution of the AP
- *AS sequence*: the sequence of authenticity steps (AS) that must be performed

In turn, every AS in the AP consists in a set of elementary actions meant to perform a specific control and/or to collect one or more authenticity evidence items, and is characterized by:

- *Controls*: the set of controls that must be performed
- *Input*: the items from the content of the processed DR and its AEH on which the AS operates
- *Output*: the set of authenticity evidence items generated by the execution of the AS
- *Actions*: a set of additional actions that are (possibly) performed as a result of the controls

Defining the APs is therefore a long and repetitive process, though a rather systematic one once the procedure is established.

5. THE VICENZA CASE STUDY

5.1 Modeling the DR lifecycle

As part of the APARSEN project activities, the Vicenza health care system preservation repository, that we have discussed in Section 2, has been selected as one of the test environments for the implementation of the authenticity model that we have presented in the previous sections. In this case study the APARSEN authenticity management guidelines have been applied to their full extent, i.e. from the preliminary analysis to the formal definition of the authenticity management policy, that is to the specification of the APs. Referring to the guidelines has provided valuable help, both in pointing out any weakness in the current practices and in providing a reasonable way to fix the problems.

In this section we shall discuss in some detail the management of medical reports, one of the several DR types which are managed by the Scryba preservation system. Further details can be found in the project documentation [3].

Medical reports are written by physicians to interpret and comment studies of diagnostic images, to which they are connected through the accession number. Reports are written using a specific *Radiology Information System (RIS)* application which is run on local systems, and are digitally signed by the physicians who write them. The digital signature process, which is directly managed by the RIS application, follows the Italian regulations and is based on the digital certificate of the physician which is held in his own smart-card or in a *HSM (High Security Module)* device for remote signature. As soon as they are completed reports are stored in a central archive managed by a centralized RIS.

According to the Italian regulations, digitally signed reports are in pkcs#7 format, a cryptographic envelope that contains:

- the report;
- the digital certificate of the physician;
- a hash file of the report encrypted with the private key of the physician.

The above information is of crucial importance to assess the authenticity and provenance of the report.

Reports are submitted by the RIS system to the preservation system almost as soon as they are completed (an upload procedure is run every 5 minutes). A SIP is generated for every single report, which is made up of two components:

- the pkcs#7 (i.e. report + certificate + signature);
- a XML metadata file.

Metadata include:

- DICOM identifier of the study to which the report refers
- Version ID (several versions of the report may be submitted and must be treated as different documents)
- Patient ID
- Patient Name;
- Patient birth date
- Patient gender
- Date of the exam

As soon as a SIP is accepted by the repository, a unique identifier (ID-DOC-Scryba) is assigned to the digital resource and a confirmation message is sent to the RIS. Then a set of controls are performed during the ingestion process:

- *Unicity check*: a check is performed to check in the repository database that the given report with the same version number and the same hash is not already in the repository.
- *Provenance check*: the digital certificate contained in the pkcs#7 file is checked against the information downloaded from the certification authority (original certificate and revocation list). This check guarantees the identity of the physician who has signed the report, and hence its provenance.
- *Fixity check*: the digital signature is decrypted and the resulting hash is compared against the hash of the report component of the pkcs#7 file. This check guarantees the integrity of the report.

Moreover a certified timestamp of the report is generated. This guarantees the existence and the content of the report at the time the timestamp is generated. In Italy the timestamp has a legal validity of 20 years.

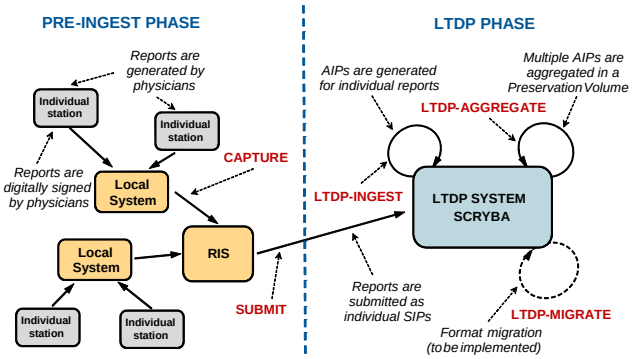

Figure 4. RIS lifecycle model.

The RIS workflow lifecycle can be conveniently modeled according to the APARSEN guidelines, and all events which are relevant to the management of authenticity, namely changes of custody and transformations of the digital resources, prove to fit well in the core set events. The resulting lifecycle model is shown in Figure 4. In the picture the two lifecycle phases, the *pre-ingestion phase* and *LTDP phase* are clearly identified, as well as the five events that we consider relevant for the management of authenticity: CAPTURE, SUBMIT, LTDP-INGEST, LTDP-AGGREGATE and LTDP-MIGRATE.

5.2 Defining the policy

The next step is, according to the guidelines, for each lifecycle event, to compare the controls and the authenticity evidence recommended by the event templates with the current practices in the repository (see Section 4.2). This analysis has pointed out that some of the of the controls are currently missing and that some of the authenticity evidence is not gathered. It is therefore necessary to carefully investigate if there is a solid justification for this.

It actually turns out that the lack of part of the authenticity evidence items that are recommended by the templates is the result of the following assumptions by the repository management, which are in turn based on a general notion of trust:

- all transfers among systems are carried on private lines that are under the ownership of a single administration (the Vicenza Public health care system), and are managed with adequate security provisions;
- access to the systems is given only to registered users, and a proper rights management policy is enforced;
- reports, after they are generated, get to the preservation repository in a very short time, therefore threats to their integrity can be considered as negligible.

These assumptions are indeed quite reasonable, and altogether we may rate the current practices in handling this event as acceptable, as long as one makes clear that:

- no controls are performed and no evidence is documented when the DRs are transferred between systems in the pre-ingestion phase;
- the integrity of data and metadata strictly depends on trusting the whole infrastructure under the ownership of the Vicenza Public health care system.

These issues and the related threats should be carefully discussed with the Designated Community, who should clearly confirm its understanding and its consensus. A preliminary analysis shows that the main (and perhaps the only) concern of the DC is the compliancy with the national regulations on LTDP, which actually can be proved.

Nevertheless one should consider that the DRs we are dealing with may become evidence in court cases about forgery or loss of data, and therefore it may be necessary to prove that their integrity has been maintained in a more substantial way. It can be argued that substantial evidence in proving the integrity could come from system logs and from the rights management policies, but this raises the further question of how long this information is maintained and how it is preserved.

Therefore we would like to suggest that some additional authenticity evidence should be preserved, for instance, for every transfer of the digital resource, a record of the time of the transfer and the identification of the source and destination system administrators.

5.3 Implementing authenticity protocols

To implement the authenticity management policy it is necessary to define the authenticity protocols for all the events in the lifecycle model. In this section we give, as an example, the authenticity protocol for the event INGEST. According to our methodology (see Section 4.4) the protocol consists in the specification of all controls and actions that must be performed during the ingestion to check the authenticity and the provenance of the DR and to generate the Authenticity Evidence Record (AER), which comprises the following *Authenticity Evidence Items (AEIs)*:

- *AEI-1. Event type*: ingest
- *AEI-2. Original identifier:* identifier from the report metadata.
- *AEI-3. New identifier in the LTDP system:* ID-DOC generated by Scryba
- *AEI-4. Context information:* DICOM identifier of the study to which the report refers.
- *AEI-5. Date and time the ingestion has been completed:* from the certified timestamp
- *AEI-6. Identification and authentication data of the LTDP system administrator:* generated by Scryba
- *AEI-7. Assessment on the authenticity and provenance:* outcome of controls on the digital signature
- *AEI-8. Digest of the AIP*: from the certified timestamp.

As discussed in Section 4.4, the protocol consists in a general specification (DR type, event type, agent etc.) and in a sequence of AS, each meant to perform a specific control and/or to collect one or more authenticity evidence items:

- *DR type*: RIS - Digitally signed medical reports
- *Event type*: LTDP-INGEST
- *Agent*: administrator of the Scryba system
- *AER*: as defined above
- *AS sequence*: steps from **AS-1** to **AS-12**

The individual authenticity steps are detailed as follows:

STEP AS-1 - CHECK PROVENANCE

- **AS-1.1**: get the digital signature certificate from the pkcs#7 file
- **AS-1.2**: get the original digital certificate from the Certification Authority
- **AS-1.3**: check the certificate in the pkcs#7 file against the original certificate
- **AS-1.4**: check the expiration date in the digital certificate against the current date
- **AS-1.5**: get the revocation list from the Certification Authority and check it

– **AS-1.6**: if any of the checks in **AS-1.3**, **AS-1.4** and **AS-1.5** fails then abort ingestion

STEP AS-2 - CHECK INTEGRITY
– **AS-2.1**: generate the hash file of the report component in the pkcs#7
– **AS-2.2**: decrypt the digital signature in the pkcs#7 file by using the public key
– **AS-2.3**: compare the two hash files generated in steps AS-2-1 and AS-2.2
– **AS-2.4**: if the check in **AS-2.3** fails then abort ingestion

STEP AS-3 - CHECK CONTEXT
– **AS-3.1**: extract the identifier of the study to which the report refers from **AER RIS-CAPTURE**
– **AS-3.2**: check the Scryba DB to verify that a study exists with identifier generated in step **AS-3.1**
– **AS-3.3**: if the check in **AS-3.2** fails then abort ingestion

STEP AS-4 - GENERATE INTERNAL IDENTIFIER
– **AS-4.1**: generate an internal unique identifier that identifies the DR in the repository

STEP AS-5 - GENERATE TIMESTAMP
– **AS-5.1**: generate a hash file of the content information of the AIP
– **AS-5.2**: send the hash file generated in **AS-5.1** to the Certification Authority to get a certified timestamp;

STEP AS-6 - GENERATE AEI: Original Identifier
– **AS-6.1**: generate AEI-2. *Original identifier* which is given the value extracted in **AS-4.1**.

STEP AS-7 - GENERATE AEI: Internal Identifier
– **AS-7.1**: generate an internal unique identifier for the DR in the Scryba system
– **AS-7.2**: generate AEI-3. *New identifier in the LTDP system* which is given the value generated in **AS-7.1**

STEP AS-8 - GENERATE AEI: Context Information
– **AS-8.1**: generate AEI-4. *Context information* which is given the value extracted in **AS-3.1**.

STEP AS-9 - GENERATE AEI: Date And Time
– **AS-9.1**: extract date and time from the certified timestamp
– **AS-9.2**: generate AEI-5. *Date and time the ingestion has been completed* which is given the value extracted in **AS-9.1**.

STEP AS-10 - GENERATE AEI: Administrator Data
– **AS-10.1**: generate AEI-6. *Administrator data* with the Scryba system administrator data

STEP AS-11 - GENERATE AEI: Assessment on Authenticity and Provenance
– **AS-11.1**: generate AEI-7. *Assessment on authenticity and provenance* which documents the outcome of the checks performed in **AS-1** to **AS-4**

STEP AS-12 - GENERATE AEI: DIGEST OF THE AIP
– **AS-12.1**: generate AEI-8, *Digest of the AIP* which is given the value of the hash file generated in **AS-6.1**.

6. CONCLUDING REMARKS

In this paper we have presented the model we propose for the management of the authenticity of the digital resources through their lifecycle, including the LTDP phase, and the operational guidelines for its deployment and the definition of the authenticity management policy in a specific environment. Moreover we have reported a case study, a repository of medical records, in which the methodology has been successfully tested.

The case study has been a quite interesting and fruitful experience, both for our team, which was concerned with the testing of the methodology and for the management of the repository which was interested in assessing the current practices and in devising possible improvements. The specific environment was indeed well suited for the purpose in several ways:

- the designated community shows a clear interest (and a strong commitment) in the problem of properly managing authenticity and provenance of DRs;
- the repository manages a variety of DRs and with quite a reasonable lifecycle complexity (changes of custody and transformations of the DRs), ;
- the repository has to comply with the quite demanding and detailed Italian rules on LTDP and the keeping of medical records, which mandate authentication of the records through digital signatures and certified time stamping, and consequently provide crucial evidence on the integrity and provenance of the records.

The model has proved to be robust enough and allowed to conveniently accommodate all the transformations and the changes of custody in the workflow. On the other hand, the templates provided by the model for the authenticity evidence records have been a comprehensive checklist to verify which authenticity evidence was actually gathered in the current practices of the repository, and to understand what information was missing and which improvements should be possibly suggested.

Another positive outcome of the case study was to confirm the flexibility of the approach that we propose, that is the ability to guide the definition of an authenticity management policy tailored to the needs of the specific environment. This is indeed a crucial issue, since different communities may have different needs and may attach to the concept of authenticity a different meaning and a different value. The balance between cost and effectiveness may therefore have quite different points of equilibrium.

7. REFERENCES

[1] APARSEN Project – Alliance Permanent Access to the Records of Science in European Network (2011-2014), http://www.alliancepermanentaccess.org/

[2] APARSEN Project: D24.1. Report on Authenticity and Plan for Interoperable Authenticity Evaluation System (2012) http://aparsen.digitalpreservation.eu/pub/Main/ApanDeliverables/APARSEN-DEL-D24_1-01-2_3.pdf

[3] APARSEN Project: D24.2. Implementation and testing of an Authenticity Protocol on a Specific Domain. (2012) http://aparsen.digitalpreservation.eu/pub/Main/ApanDeliverables/APARSEN-DEL-D24_2-01-2_2.pdf

[4] CCSDS: Reference Model for an Archival Information System – OAIS. Draft Recommended Standard, 650.0-P-1.1 (Pink Book), Issue 1.1 (2009), http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/CCSDSAgency.aspx

[5] Factor M., Guercio M., et al.: Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage. TaPP '09 (2009) http://www.usenix.org/event/tapp09/tech/full_papers/factor/factor.pdf

[6] InterPARES Projet: Requirements for Assessing and Maintaining the Authenticity of Electronic Records (2002), http://www.interpares.org/book/interpares_book_k_app02.pdf