Project no. 269977

APARSEN
Alliance for Permanent Access to the Records of Science Network

Instrument: Network of Excellence

Thematic Priority: ICT 6-4.1 – Digital Libraries and Digital Preservation

D24.2 IMPLEMENTATION AND TESTING OF AN AUTHENTICITY PROTOCOL ON A SPECIFIC DOMAIN

Document identifier:	**APARSEN-REP-D24_2-01-2_3**
Due Date:	2012-02-29
Submission Date:	2012-04-30
Work package:	WP24
Partners:	CINI, CERN, UKDA
WP Lead Partner:	CINI
Document status	FINAL
URN	urn:nbn:de:101-20140516169

Abstract:

In the original definition given in CASPAR, *Authenticity Protocols (APs)* are the procedures to be followed in order to assess the authenticity of specific type of *Digital Resource (DR)*.

The CASPAR definition is quite general and does not make reference to a specific authenticity management model. As part of the activities of APARSEN WP24 we have formalized an authenticity management model, which is based on the principle of performing controls and collecting authenticity evidence in connection to specific events of the DR lifecycle. This allows to trace back all the transformations the DR has undergone since its creation and that may have affected its authenticity. The model is complemented by a set of *operational guidelines* that allow to set up an *Authenticity Management Policy*, i.e. to identify the relevant transformations in the lifecycle and to specify which controls should be performed and which authenticity evidence should be collected in connection with these transformations.

To formalize the policy we have indeed resorted to CASPAR's AP definition, and we have adapted and extended to integrate it in our authenticity management model. In our methodology the AP therefore becomes the procedure that is to 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 original content of this deliverable, which was aimed at "implementing and testing an authenticity protocol on a specific domain", has been adapted and extended to encompass the whole scope of the authenticity evidence management guidelines.

The current aim of the deliverable has therefore become to test the model and the guidelines at the operational level when dealing with the concrete problem of setting up or improving a LTDP repository in a given specific environment, to get to the definition of an adequate authenticity management policy. Moreover, instead of concentrating on a single environment, we have decided to extend the analysis to multiple test environments provided by APARSEN partners.

Shifting to a practical ground and facing the actual problems that arise in the management of a repository has indeed been an important move to fill the gap that still divides the mostly theoretical results of the scientific community from the actual practices carried on in most repositories, and to reduce the fragmentation among the different approaches that prevents interoperability. And the case studies have proved the validity of this approach. On the one hand they have proved to be easily applied and well understood in all the test cases, and on the other hand the simple and yet rigorous concepts introduced by the model may provide a common ground for the management of authenticity evidence and for exchanging it among different systems.

In at least one of the case studies, the guidelines have been applied to their full extent, i.e. from the preliminary analysis, to the identification of the relevant lifecycle events, to the detailed specification of the authenticity evidence to be collected, to the formal definition of the authenticity management policy, that is to the specification of the AP. In all cases, 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.

Delivery Type	REPORT
Author(s)	Silvio Salza, Mariella Guercio (CINI); Hervé L'Hours, Matthew Woollard (UKDA); Suenje Dallmeier-Tiessen, Patricia Herterich (CERN)
Approval	David Giaretta/ Simon Lambert
Summary	
Keyword List	Authenticity management, authenticity evidence, authenticity protocol, preservation, data repository, case study, Health Care, Social Sciences, High Energy Physics, experimental data
Availability	☒ PUBLIC

Document Status Sheet

Issue	Date	Comment	Author
0.1	2012/03/15	Preliminary draft	Silvio Salza (CINI)
1.0	2012/04/11	Complete draft ready for internal revision	Silvio Salza (CINI)
2.0	2012/04/23	Complete revised draft that incorporates comments from internal reviewers and two new sections (now sections 5 and 6) to discuss integration within the project and with other projects, as well the transfer of the results into practical environments. Section 2.7 (concluding remarks for section 2) has been added as well.	Silvio Salza (CINI)
2.1	2012/04/28	Minor format corrections	David Giaretta
2.2	2012/04/28	Minor amendments	Silvio Salza (CINI)
2.3	2012/08/16	Further minor format changes after acceptance by EC	Simon Lambert

Project information

Project acronym:	**APARSEN**
Project full title:	**Alliance for Permanent Access to the Records of Science Network**
Proposal/Contract no.:	**269977**

Project Co-ordinator: Simon Lambert/David Giaretta

Address:	STFC, Rutherford Appleton Laboratory Chilton, Didcot, Oxon OX11 0QX, UK
Phone:	+44 1235 446235
Fax:	+44 1235 446362
Mobile:	+44 (0) 7770326304
E-mail:	simon.lambert@stfc.ac.uk / david.giaretta@stfc.ac.uk

CONTENT

1 INTRODUCTION

In the original definition of CASPAR [3], an *Authenticity Protocol (AP)* is the procedure that must be followed in order to assess the authenticity of specific type of Digital Resource (DR). More precisely, an AP is an ordered sequence of interrelated steps, each one of which we will refer to as an *Authenticity Step (AS)*. Each AS is performed by an *actor*, which can act either in an automatic or in a manual way. The execution of an AP generates an *Authenticity Protocol Report (APR,)* that documents that the sequence of ASs has been executed and collects all the values associated with the data elements analyzed in every AS, and possibly the outcome of the execution.

As part of the activities of APARSEN WP24 we have instead presented an authenticity management model, which is documented in detail in the companion deliverable D24.1 *Report on authenticity and plan for interoperable authenticity evaluation system* [1], and is based on the general principle of performing controls and collecting authenticity evidence along the whole DR lifecycle, in order to be able to trace back all the *transformations* the DR has undergone since its creation and that may have affected its authenticity. Each transformation 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.

The model (see [1] sect. 4) concentrates on the definition of 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 should be considered as a sort of common basis on which different keeping and preservation systems may agree, thus achieving at least a basic degree of interoperability in the exchange and management of authenticity evidence.

To each event from core set the model associates an *event template*, that is the specification of the controls that should be performed and of the set of information that must be gathered in connection with the event to support the tracking of authenticity and provenance, which is called the *Authenticity Evidence Record (AER)*. As a 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. This evidence will follow the DR when it is transferred between different systems, and will accompany it throughout its lifecycle.

The next step has been to move to the operational level and to define the procedure that should be followed, when dealing with the practical problem of setting up or improving a LTDP repository in a given specific environment, to define 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. To this purpose we have developed a set of operational guidelines to deal with the problem in a systematic way, that is a sequence of steps that go from understanding the meaning of authenticity for the designated community, to the identification of the relevant lifecycle events, to the definition of the policy, that is the formal specification of the controls that have to be performed and of the authenticity evidence that should be gathered in connection with the relevant lifecycle events (see [1] sect. 5).

To formalize the policy we have resorted to CASPAR's AP definition, but we have adapted and extended the definition. In our methodology an AP becomes the specific procedure to 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, and will operate on the authenticity evidence collected so far, that is on the AEH, to produce a further AER.

As a result of this, the original content of this deliverable, which was aimed at "implementing and testing an authenticity protocol on a specific domain", has been adapted and extended as well to encompass the whole scope of the authenticity evidence management guidelines that we have devised in [1]. The current aim of the deliverable has therefore become to test the model and the guidelines at

operational level when dealing with the concrete problem of setting up or improving a LTDP repository in a given specific environment, to get to the definition of an adequate authenticity management policy.

Moreover, instead of limiting our evaluation to a single environment, we have decided to extend the analysis to multiple test environments provided by APARSEN partners. The case studies are presented in the following sections.

The repository of the health care system in Vicenza (Italy) is discussed in section 2. It is a rather complex case since along the DR lifecycle there are several changes of custody that involve, beside the LTDP repository, several keeping systems, some of them geographically distributed in the district. Moreover there are several types of DRs (diagnostic images, medical reports etc.), each one with a distinct workflow. This case is also interesting because the repository must comply 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. These rules have introduced a good deal of complexity in the workflow, but have proven, on the other hand, to have a positive effect with regard to the authenticity management, since, for instance, the widespread use of digital signatures and certified timestamp provides valuable authenticity evidence.

The guidelines have been tested in the Vicenza case study (at least for two of the workflows) to their full extent, i.e. from the preliminary analysis, to the identification of the relevant lifecycle events, to the detailed specification of the AERs. Moreover in one case the process has been carried out to the formal definition of the authenticity management policy, down to the specification of the authenticity protocol. In all cases, referring to the guidelines has provided valuable help, both in pointing out some weakness in the current practices and in providing a reasonable way to fix the problem.

A second case study is presented in section 3 and deals with the social science and humanities repository at the UK Data Archive at the University of Essex. This Government and academic heterogeneous repository holds over 5,000 data collections relating primarily to society and the economy and offers extensive supporting services. The Archive has existed since 1967 and has hosted the Economic and Social Data Service for a decade and more recently it provides secure access to sensitive data through the Secure Data Service and is engaged in a number of data management initiatives.

The Archive ingests extremely heterogeneous collections with limited influence over actions in pre-ingest keeping systems (limiting their detailed responses to the SUBMIT and INGEST events) as their processes evolve in a far less legislatively defined environment that in the Vicenza case study. Like the CERN case study (described below) they exist within a highly connected, long-standing relationship with their depositors (across the governmental and academic sectors) but unlike the other case studies their workflow involves extensive curation and enrichment (mainly for standardization and context) during AIP creation relying on complex manual processing by specialized teams.

Another element of interest is that this institution has already passed through a formal auditing process and the repository management is based on detailed and well devised procedures. In this case too, referring to the guidelines has proved interesting both in analysing potential issues and in suggesting ways to solve them.

The last case study is discussed in section 4 and is devoted to the scientific experimental data management at CERN, and more generally to the High Energy Physics community, which manages an immense data flow, and has to face the considerable complexity and diversity of the research data output. As for the designated community consumers of shared data and preserved data, to date, vastly overlap with those who produced it and the assessment of the authenticity of the data is apparently still heavily based on the notion of trust. However it is clearly realized that the implementation of a solid strategy for data preservation, appropriate tools and a culture of sharing, would allow a growth in the community, as those who are not part of the original research that has been conducted could use the data, too.

Finally, in section 5 we describe how this work is related with the other work packages and tasks of APARSEN, in section 6 we discuss the integration of the activity in WP 24 with other projects and how the results of the RTD activity could be actually translated into practice, and in section 7 we give our concluding remarks.

2 HEALTH-CARE DATA REPOSITORY IN VICENZA

2.1 MANAGING AND PRESERVING HEALTH CARE DATA IN ITALY

2.1.1 Regulations on Long Term Digital Preservation in Italy

The regulations about long term digital preservation in Italy is very complex as it is based on a series of integrated general and technical rules dedicated both to the creation, management, keeping and preservation of electronic records and to the specific e-health data and records production and retention. The main difficulties are those related to the choices the Italian legislators have made since 1997 about implementing the digital signature as the crucial (initially the only) mechanism to ensure the integrity and authenticity of the digital objects.

As a consequence, the validity over time of records digitally signed and required to be preserved for long-term has implied the approval of complex and dedicated regulations, very detailed, frequently modified (in 1994, in 1998, in 2001, 2004) but never provided with standardized sets of representation information and metadata in order to build a preservation environment, interoperable over space and over time. In particular, an initial high level of rigidity of this *corpus* of rules made their concrete and effective implementation impossible for the period 1994-2004. The second phase (2004-2010), characterized by a more simplified legal framework, has increased the number of applications mainly in the e-health field, but has not been able to ensure efficient and qualified solutions and interoperable mechanisms because of the absence of standardized specifications and common sets of metadata. This lack has resulted in a low level of uniformity of the data creation and keeping even in the same producer departments in the pre-ingestion phase and weak controls on the digital repositories and on the crucial events, even those identified by the legislator as relevant in order to provide identity and integrity of the preserved digital objects.

In 2009 a few e-service providers (specifically those involved in the e-health sector) seriously considered the risks involved in this situation and asked for the definition of a uniform framework, to be approved at least as a national standard. The provider MEDAS, whose digital repository model is described and evaluated here because it was implemented by the Azienda sanitaria di Vicenza, promoted and strongly supported this effort. In 2010 the National body for standardization (UNI-Ente nazionale italiano di unificazione) has accepted the need to develop such a standard under the responsibility of the Italian Standard Committee on Record and Archives. In 2011 the new standard was approved and denominated *UNI 1386:2010 SInCRO - Supporto all'Interoperabilità nella Conservazione e nel Recupero degli Oggetti digitali*. Its goal refers explicitly to the need to support the interoperability of digital objects provided with digital signatures and time stamps for their long-term preservation and distribution [7].

The working group for the formal definition of these new technical rules included the main public institutions involved in the ICT legislation at national and at regional level (CNIPA – Centro nazionale per l'informatica nella pubblica amministrazione; Direction general of Archives, Regions charged for creating trusted digital repositories like Regione Emilia Romagna), the Ministry for Health, the main national providers charged for the e-infrastructure of digital signatures and e-archiving (Telecom, Postecom, Infocamere) and important Italian software companies in the field (SIAV, Technint, Medas). The final standard is not focused only on the medical data and records but is intended to define a common template of metadata and basic workflows OAIS compliant for exchanging sets of any kind of digital objects to be preserved for medium or long-term.

The main 'political' result has been the inclusion of the rules as crucial part of the new regulation for medium and long-term digital preservation prepared by the Italian government in 2011. The regulations will be operative in the second half 2012 as application directive of the legislative decree n. 235/2010 called Codice dell'amministrazione digitale – CAD (Code of digital administration).

A large part of the general principles and methodological and juridical dispositions present in the CAD requires guidelines specified in detailed regulations. Specifically, all the questions related to the digital preservation (interoperability of digital signatures, the persistency of the legal value of the records/digital objects, the schemas for metadata and representation information of information contents, the criteria to assess and monitor the preservers and the preservation processes quality are prescribed and described in detail by these technical rules. The rules issued in the previous phase (2004-2010) have been changed and improved according to the new standards (UNISInCRO) and in order to be compliant with the OAIS model.

In summary the regulations in place are based on the lifecycle principles and concern elements, events and responsibilities relevant for the management of identity and protection of integrity both in the keeping/archiving of the digital objects in the pre-ingestion and their acquisition and handling in the preservation repository:

2.1.1.1 Main rules according to the national regulation in place for 2004-2011

1. The electronic records required fixed content and form; they cannot include macro-instruction or executable codes; the following formats are not prescribed but only suggested: PDF/A, TXT, ODF, XML.
2. Digital signature and time stamp are the mechanisms in place to ensure the control of provenance and fixity; in case of public administrations (but not necessarily for e-health records created by public hospitals), provenance and context are basically ensured by a registry system which implies obligatory elements (unique identifier including a reference code for each organizational unit, names of the creator, sender/addressee, the date of the records creation, the hash of any digital object, the subject and the classification code aimed at establishing a functional context.
3. Hash is required and has to be associated to the digital object when preserved in the repository.
4. For preservation reasons the hashes of digital objects/records could be aggregated in one file called *Preservation Index file (file Indice della Conservazione)* which must include for reference the codes related to each record/digital object managed for preservation (a sort of index whose obligatory elements were not identified).
5. Agents responsible for the preservation are appointed both within the producer and within the repository: they have to apply their digital signature with related time stamp to each *volume* to ensure fixity.
6. The media to preserve the *volumes* must be adequate (optical or magneto-optical disks are suggested).
7. The rules identify specific responsibilities (the *responsible for the preservation*) even if they do not list skills and competences, but only the tasks which include: the creation of a database of digital objects, the respect of security and defined procedures for audit trail; the general capacity to ensure the readability, intelligibility, preservation and accessibility for the digital resources.
8. The tasks for the preservation can be delegated but not with reference to its legal responsibility which cannot be transferred according to the national juridical system. The regulations do not define or suggest any formal procedures for this delegation of responsibilities.

2.1.1.2 Main rules according to the national regulation updated in 2012

The new rules for preserving private and public digital objects/records are intended to define qualified and standardized parameters able to certify the processes and assess preservers and services according to an international perspective. The main changes include

1. A common glossary integrates the definitions present in the main legislation and in the international standards and ensures accuracy and completeness.
2. The OAIS architecture is generally adopted for preservation services with specific reference to the definition of any transfer of digital objects/records: the information packages for submission, archiving and distribution are identified as the obliged form for handling the transmission; only the archival information package is strictly detailed even if not in the form of an XML schema.
3. The chain of responsibilities for the preservation is clearly identified and supported by scenarios and organizational models.
4. The main procedures are defined in detail in the form of a *manual of preservation*, whose chapters are.
5. The national standard for interoperability of digital volumes approved by UNI (UNISInCRO) is recognized as a reference guideline.

2.1.2 Regulations and prescriptions on Health care data in Italy

Administrative records are not discussed in this deliverable because they are not subject to different legislation in health care environment. A dedicated legislation applies to the clinical records and data and their aggregation in case of patient's hospitalization. The regulations in this specific sector have to integrate the national rules and technical guidelines. The main issue concerns the fact that the dossier and files related to the patient can and have to imply the capacity of sharing health data among various bodies and professionals, have to be updated, and must be handled carefully with reference to their integrity, authenticity and completeness. At the same time the guarantees for privacy and controlled access have to be ensured at the highest level of quality. These regulations are here identified under the following sections: national rules for the retention, privacy and access regulations, technical rules for digitization processes.

2.1.2.1 Rules for the retention of health care records

The legislation in place at national level for the retention of health care and clinical records and data concerns both private and public sector. For the records created by public organizations the Direction General of Archives has been always in charge with the authority of controlling the creators' preservation plans and approving the final decisions for disposal. A new law has been passed in 2004 (Code for cultural heritage n. 41/2004) with specific reference to the scientific and cultural value of the records. The Code makes explicit the control of the National Archives not only on the historical archives for any kind of public agencies (as previously recognized by a long series of regulations) but also on the current records created in the public environment, including records and data of public value produced for health care both in traditional and in digital form. There is a specific rule concerning patient records and files: the circular of the Ministry for Health Care approved in 1986 (19 December 1986 n. 61). According to this circular, the patient records and the medical reports require long-term preservation – that in Italy this means an unlimited term of preservation in archival dedicated repositories – because of their public nature and their historical value. The same decree stated that the radiological images could be destroyed after a 20 years term. In 1997 a ministerial decree (dm 14 February 1997) reduced this second term to a 10 years period and included in this retention category also the records and data created by the outpatients departments. The decree describes how the images have to be acquired, archived and accessed and how they have to be preserved when not delivered to the patients.

2.1.2.2 Privacy and access regulations

The main specific regulation, based on the European guidelines 00323/07/EN WP131 "*Working Document on the processing of personal data relating to health in electronic health records (EHR)*", is

the legislative decree 230/1997 which identifies the obligations for health care public and private agencies to make available their clinical records. Obligations and parameters have been partially modified by the new general privacy legislation approved in 2003 (dlgs 196/2003). Special controls have been also defined by the legislative decree 37/2010

The main changes with practical consequences in terms of digital data and records creation and keeping have been stated by the guidelines of the *Garante per la protezione della privacy* specifically dedicated to the creation of the e-health file and report: "*Guidelines for electronic files and health report*" (16 July 2009) and "*Guidelines for health reports on line*" (19 November 2009).

The patients case history records created and preserved in the hospitals and their related clinical documentation produced by the outpatients departments can be accessed by the internal medical staff and hospital administrators for clinical and legal reasons, while special permission and an undertaking to follow the respect of privacy legislation are always required both in case of internal or external access for scientific or statistical activities and for requests from the patient's medical doctor or from another hospital.

Authenticated copies of patients case history records are provided by the hospital director, but the request must follow special requirements according to the quality of its author (the patient, his/her delegate) and the nature of the transmission channel (by fax, by mail, by certified e-mail). Also the delivery of these records is subject to specific rules to ensure the integrity and a controlled access in the course of the delivery procedure. These limits have specific consequences in the digital environment and require tools and mechanisms able to verify the digital identity of the sender and of the addressee.

2.1.2.3 Technical rules for the digitization of e-health care data and records

The national legislation on digitization and preservation of records has to be applied also for e-health care records. The Ministry for Health Care approved in 2007 guidelines for handling digital resources for health-care to direct the digitization initiatives in the field of diagnostic images [4].

The guidelines are dedicated to determine the juridical nature of the records to be produced and kept in the health care environment in order to identify the level of authentication required by the national legislation when a digitization process is applied. The suggestions provided by the technical guidelines are not very detailed. They include:

- the definition of the clinical records:
  - the health care reports have juridical value and must be signed by the medical doctor,
  - the radiographic images or the tests/screening data have no official nature as they are just data able to support the medical diagnostic reporting: they do not require a signature,
- the instructions for applying the national rules described at 1.1,
- suggestions for transforming the traditional health care report into a structured health care report inclusive of both text and images in the form of a structured bit stream according to a standardized protocol like DICOM or as an hypertext as suggested by the CDA HL7.

2.1.3 General procedures for digitization of radiographic images and medical diagnostic reports

As previously described the main clinical documentation here examined concerns:

- the medical diagnostic reports
- the radiographic images

The Guidelines approved in 2007 defined the main procedures for the digitization process of both the reports and the images.

2.1.3.1 Digitization of radio-graphical images.

With reference to the radiological diagnostics, they make a distinction between the health care report and the images even if in the digital environment the image is created – according to the standard DICOM – as a as a set of iconographic, anagraphic and numeric data, able to be easily transformed into a structured report.

Specifically this structured report has the form of an hypertext which includes metadata related to:

- patient information,
- annotations on patient case history,
- clinical questions,
- description of the techniques related to the radiologic investigation,
- description of the radiologic report,
- final diagnosis,
- images automatically created by the modality and identified by the radiologist because of their relevance,
- any other type of analysis, reconstruction of reformatting made by the radiologist for producing the diagnosis, including the changes of visualization values: all these changes have to be tracked in the digital archives in the form of different "representation status".

The structured report is subject to the same rules of the other textual reports. Any digital images have to follow a protocol which implies the association with other information, like patient information, methods of acquisition and visualization.

The digital evidence created in the form of DICOM are transferred (according to the requirements established by the requirements agreed by the producer and the preserver) to the PACS (Picture Archiving and Communication System) where they are subject to the archiving procedures (through their association with a unique identifier created by the modality itself). The archived data are sent to a long-term preservation system which can be another one than PACS. In any case as soon as archived (that is at the inclusion of the images in the PACS) it is impossible to change any data. The guidelines suggest the images are submitted to the preservation system as soon as possible to avoid alterations difficult to track if special measures are not in place.

2.1.3.2 The digitization of medical diagnostic report

According to the national regulations on electronic records (see 1.1) and to the definition of health care reports as records, the national guidelines and the present practices imply the digital signature of all the health care reports, including those related to the medical diagnostic. As soon as digitally signed the report is associated with a unique reference code and archived waiting for submission to a long-term preservation repository. The responsible for preservation is required to define the terms for the preservation and the requirements necessary to ensure integrity and authenticity to the reports.

2.2 THE HEALTH CARE SYSTEM IN VICENZA

2.2.1 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 that have been discussed in section 2.1.1. 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.

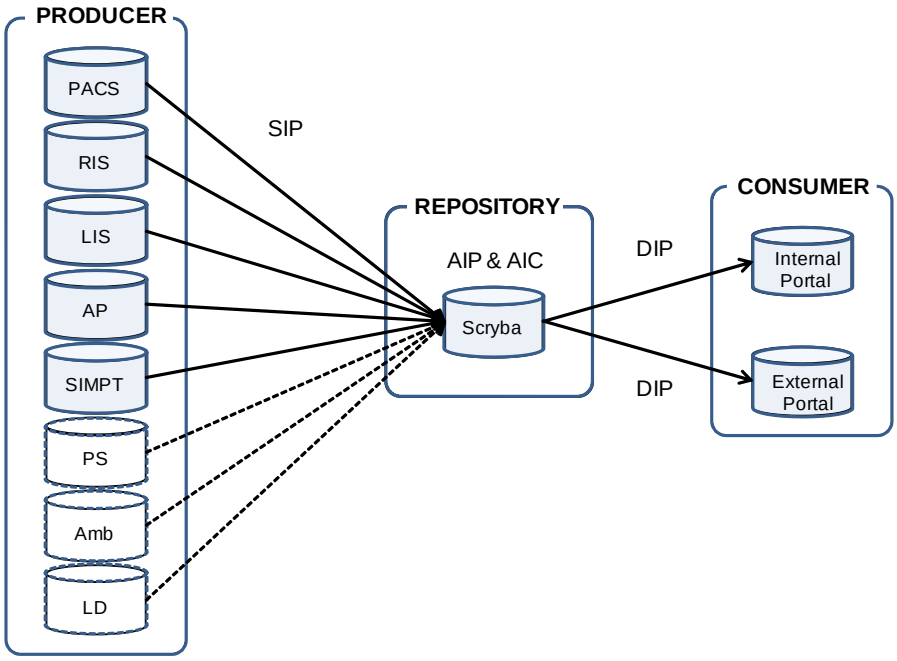


Figure 2.1 - The preservation infrastructure

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 digital imaging devices, workstation attended by physicians etc.

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 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 3.1 were the different kinds of producers are represented. Currently five different producers are supported and have been operational since 2007, but support for three additional producers is being implemented.

The currently supported producers are:

- *PACS (Picture Archiving and Communication System)*

  This system relies on a well known digital imaging technology which provides storage of, and convenient access to, images from multiple *modalities* (source machine types) as X-ray, computed tomography (CT) and magnetic resonance imaging (MRI). Digital resources from

this producer are *studies*, i.e. collections of diagnostic images, and are stored in DICOM (Digital Imaging and Communications in Medicine) format, an acknowledged standard for this kind of data objects. In addition to a set of images a DICOM file contains a large number (hundreds) and a large variety of metadata, some of which are relevant for the preservation process, and more specifically for the management of the authenticity. Studies are correlated to medical reports which are managed by the RIS system (discussed below). Each report is univocally linked to a given study by means of the *DICOM accession number*, an unique identifier which is automatically assigned to each study by the modality.

- *RIS (Radiology Information System)*

  This system manages the medical reports written by physicians, using a specific RIS application, to interpret and comment the studies of diagnostic images. Each report includes two components: a typed text in pdf format and a set of metadata in XML format. Each report is correlated to a specific digital image study through the *DICOM accession number*, a specific DICOM tag. Due to the relevance of their content and to the related responsibilities, all the reports are digitally signed by the physicians who write them. The digital signature is a crucial element in assessing the integrity and the provenance of the report and, according to the Italian regulations, has the same legal value as a written signature. In next future the report format will be change to CDA2 HL7.

- *LIS (Laboratory Information System)*

  This system is similar to RIS, but based on a different software system, and manages reports from medical laboratories. Each report includes two components: a typed text in pdf format and a set of metadata in XML format. As in the case of RIS, due to the relevance of their content and to the related responsibilities, all the reports are digitally signed. The reports which pass a series of sophisticated automated controls are signed by means of a special automated device, the HSM (Hardware Security Module), and bear the digital signature of the head of the laboratory; the other ones are digitally signed by individual physician. The use of HSM devices is explicitly contemplated by the Italian regulations, and the cryptographic process is exactly the same, therefore using either procedure makes no difference as far as the validity of the signature, and its value as authenticity evidence is concerned. In next future the report format will be change to CDA2 HL7.

- *AP (Anatomical Pathology)*

  This system is similar to RIS (and LIS) and manages the reports from the anatomical pathology laboratories. As in the two previous cases reports include text and metadata and are digitally signed. In next future the report format will be change to CDA2 HL7.

- *SIMT (Servizio Immunoematologia e Medicina Trasfusionale)*

  This system is similar to the previous ones (RIS, LIS and AP) and manages the reports from the immunology laboratories. As in the previous cases reports include text and metadata and are digitally signed.

2.2.2 The Scryba preservation system

As already noted 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 2.2.

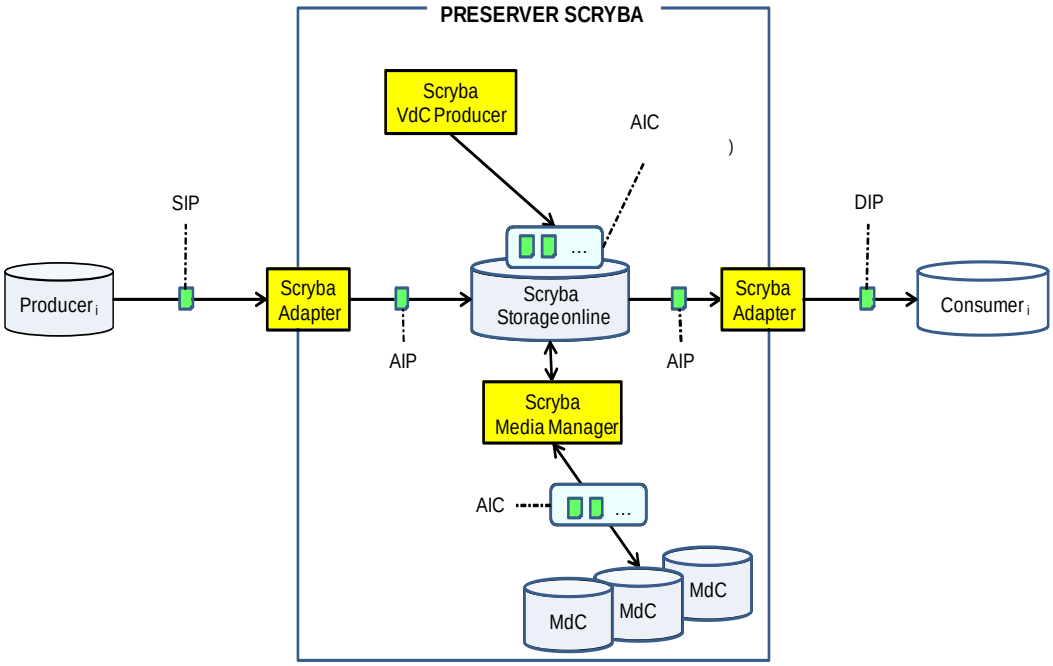


Figure 2.2 - The Scryba preservation system

The system has a modular structure which is based on a core structure whose main function 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. Adapters on the producer side are configured according to the specifications of the SIP (Submission Information Package) to perform the ingestion, with all the related controls and transformations which are needed to build the AIP out of the information delivered in the SIP. Similarly, adapters on the consumer side are configured to accept and process requests from customers and to build DIPs (Distribution Information Packages) according to the specifications.

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 PV (Preservation Volume), 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.

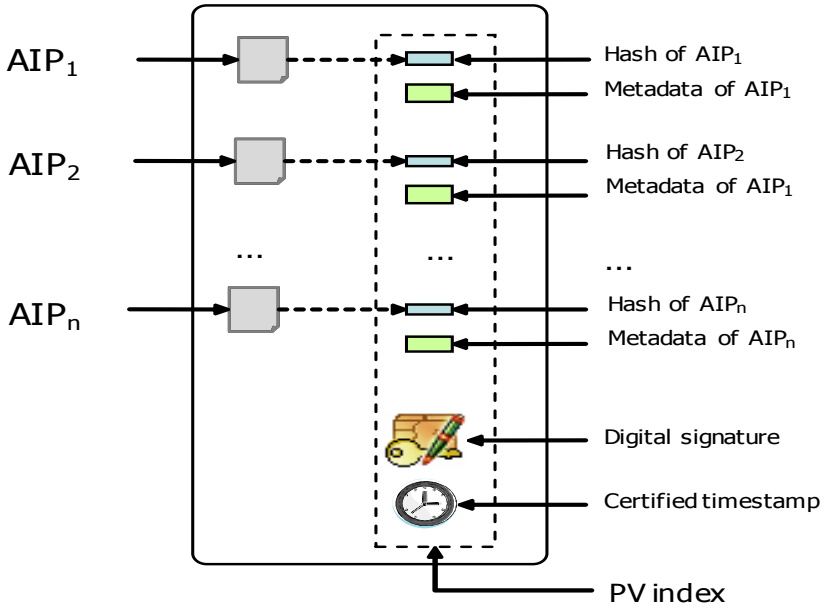


Figure 2.3 - Structure of a Preservation Volume

The structure of a preservation volume is shown in figure 2.3. It contains all the aggregated digital resources plus an additional file, the *Preservation Volume index* (PV index), which is UNI SInCRO compliant (see sect. 2.1.1 and [7]) and 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:

- an 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;
- size: a PV cannot exceed a *maximum size*.

Currently the maximum time is set to 24 hours and the maximum size to 1 GB for all types of digital resources.

We shall point out that there is actually no need to aggregate several digital resources in a single preservation volume in order to comply with OAIS. This is only done to conform with national regulations, which are in this case somehow odd, at least in our opinion.

2.3 SAMPLE WORKFLOWS

In our study we decided to concentrate on two of the workflows that have been presented in section 2.3.2: the radiological images submitted by the PACS and the medical reports submitted by the RIS. Considering all the workflows would not add very much to the discussion, since the digital object submitted by the LIS, AP and SIMT systems are too medical reports and have pretty much the same structure than the ones submitted by RIS.

2.3.1 Radiological image studies submitted by PACS

2.3.1.1 DICOM studies

As discussed above (see 2.2.2), digital resources in this workflow are studies, i.e. set of radiological images, The studies are generated by several different types of devices, technically called modalities, and initially archived in local systems that are geographical distributed in several locations. Studies are digital resources with a quite complex internal structure, organized according to the DICOM standard.

A DICOM study is a set of images, each stored in a DICOM file. Both the image and the related metadata are contained in a single file which is organized as a sequence of elements, one of which contains the pixel data. That means that a file of a chest X-Ray image, for example, actually contains the patient ID within the file, so that the image can never be separated from this information by mistake. Multiple images in a single study are organized in series have three levels of unique identifiers (UID): study identifier, series identifier and image identifier. Most importantly, the study has also a special attribute, *the DICOM accession number*, an unique identifier that allows to correlate further information, namely medical reports, to the study.

Altogether, DICOM (ISO 12052:2006) is a complete standard for handling, storing, printing, and transmitting information in medical imaging. It enables the integration of scanners, servers, workstations, printers, and network hardware from multiple manufacturers into a picture archiving and communication system (PACS).

2.3.1.2 Management of DICOM studies in local and central PACS

An overall picture of the PACS workflow is represented in figure 2.4. Studies are initially stored in archives managed by local PACS. These archives are situated in the local facilities of Vicenza Public health care system, and provide access to local users to all the images that have been generated at a given local facilities.

Studies are then sent, through an automatic procedure, to the central PACS where an archive of all studies is managed, which provide access to (authorized) users from all locations. These are anyway additional copies of the studies, since the original files are kept in the local archives. The communication procedure between the PACS system is part of the DICOM standard.

Next, according to Italian regulations we have discussed in sect. 2.1, the studies are at once submitted to the Scryba system, i.e. to the LTDP repository. The transmission follows the DICOM protocol. Furthermore the consistency between the repository and the central PACS archive is maintained by periodically running an application that scans the central archive to check for studies that have not yet sent to the repository and vice versa. The application is run on a daily, weekly and monthly basis. Therefore, since studies are moved from local to the central PACS within one day, a study generally gets to preservation at most 24 hours after it has been generated by a modality.

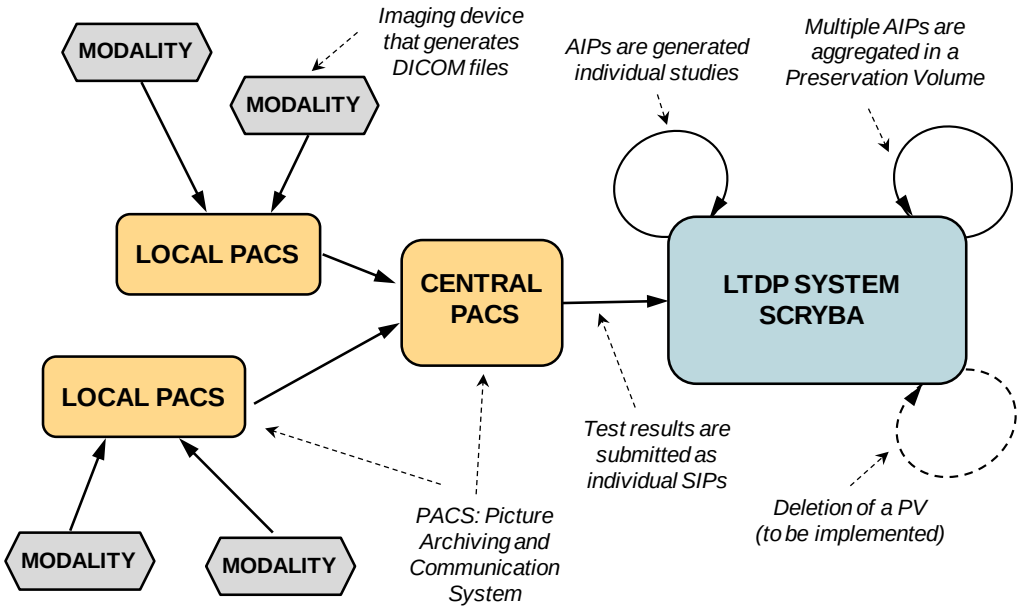


Figure 2.4 - The PACS workflow

Studies are submitted individually to the Scryba system, which means that a SIP (Submission Information Package) is prepared for each study, which contains the DICOM files (i.e. the images) and, in a separate XML file, the most important metadata, as for instance:

- DICOM accession number
- Patient ID
- Patient Name
- Patient birth date
- Patient gender
- Date of the exam

Further metadata are implicit, i.e. enclosed in DICOM files, and may be extracted during the ingestion.

The transmission of the studies at all levels, i.e. from the modality to the local PACS, from the local to the central PACS and from the central PACS to the Scryba system is performed on proprietary networks. Furthermore system authentication is based on digital certificates. Therefore, since all systems, including the LTDP repository Scryba, and communication infrastructures are within the Public health care administration, no specific controls are performed when digital resources are exchanged, and the assessment of the authenticity and provenance is mostly based on the notion of trust. However, as we shall discuss later, individual systems are run under different responsibilities, and therefore, in case of data loss or forgery, it may be difficult to locate specific responsibilities. This issue will be discussed in a later section.

2.3.1.3 Long term preservation of DICOM studies

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 submitting PACS.

SIPs are then ingested by Scryba and an AIP is generated for every SIP, that is for every individual study. A minimal set of controls are performed, mostly to check that no study is already in the repository with the same ID and the same set of images (that should be the same version of the same study), no further controls are performed on this kind of digital resource since, as we will see later when discussing the RIS workflow, controls are mostly connected with digital signatures, and a digital signature is not currently required for radiological studies.

The PDI is composed by the explicit metadata in the SIP, by further metadata extracted from the DICOM files and by information documenting the submission and ingestion processes.

According to the Italian regulations (see sect. 2.1.1) several studies are aggregated in a PV, which will be the final object of the preservation process (see sect. 2.2.3). This corresponds in the OAIS model to aggregating several AIPs in a single AIC (Archival Information Collection).

The Italian law requires the preservation radiological images for ten years. Once this time has elapsed the studies can be disposed. Since each PV contains digital resources of the same type and stays open at most for one day (see sect. 2.2.3), when the preservation time elapses, all the studies contained in a given PV expire at the same time and the whole volume can be deleted.

Actually the repository has been operational for only a few years, and hence the deletion procedure has not yet been implemented.

2.3.2 Medical reports submitted by RIS (Radiology Information System)

2.3.2.1 Digitally signed medical reports

An overall picture of the RIS workflow is represented in figure 2.5. 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 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 digital the 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.

The structure of the report is based on HL7-CDA2 (Health Level 7-Clinical Document Architecture), an XML-based markup standard intended to specify the encoding, structure and semantics of clinical documents for exchange. As with the DICOM format, a great variety of metadata are inside the document.

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

- the report;
- the digital certificate of 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 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 file with 'explicit' metadata (further metadata can be found inside the HL7-CDA2 file).

Explicit metadata include:

- DICOM accession number 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)

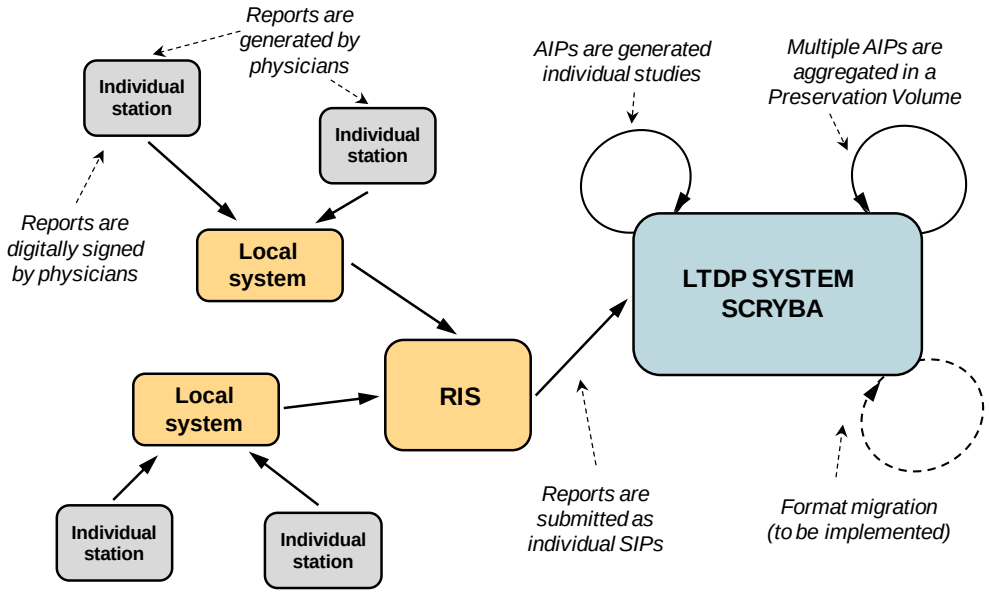


Figure 2.5 - The RIS workflow

- Patient ID
- Patient Name;
- Patient birth date
- Patient gender
- Date of the exam

2.3.2.2 Preservation of digitally signed medical reports

Similarly to the PACS workflow, 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. But different controls are performed in this case during the ingestion process, since reports are digitally signed.

- *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 HL7-CDA2 file. This check guarantees the integrity of the report.

Moreover a certified timestamp of the report is generated. The timestamp guarantees the existence and the content of the report at the time it was generated. In Ital it has a legal validity of 20 years.

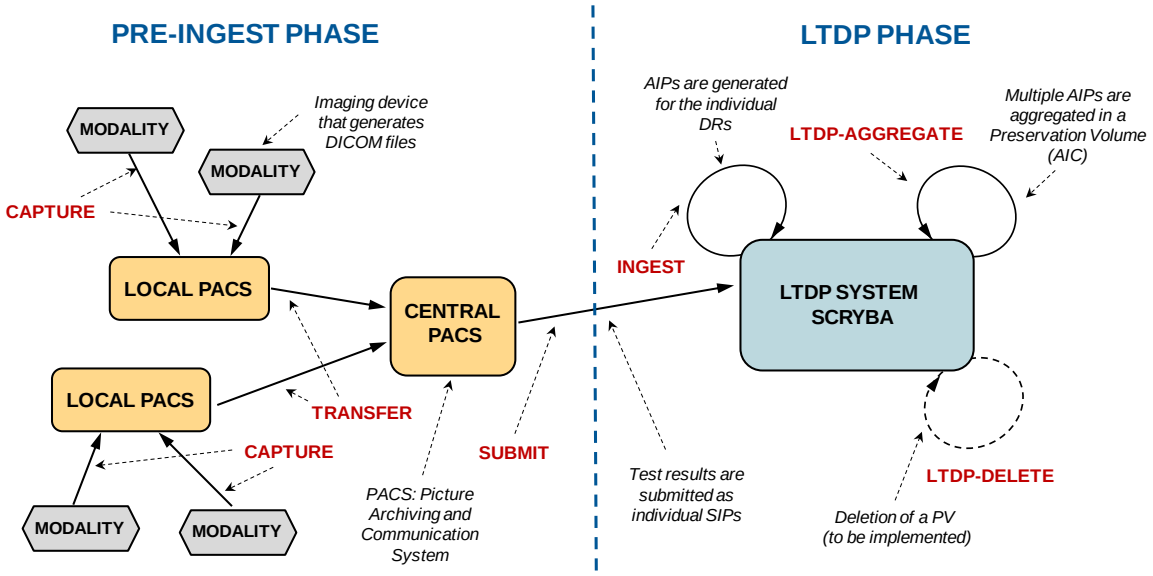


Figure 2.6 - The PACS workflow lifecycle

As for the PACS workflow an AIP is generated for every SIP, that is for every report, and multiple AIP are aggregated in a single PV (i.e. in single AIC)

On the other hand the Italian law does not currently set any time limit to the preservation of medical reports (though the studies to which they refer may be discarded after 10 years), hence no deletion procedure is envisaged for this kind of digital resource.

2.4 MODELING THE DIGITAL RESOURCE LIFECYCLE

Aim of this section is to check the validity of the model and the guidelines that have been proposed within the APARSEN project for the management of the authenticity evidence and that have been presented in project deliverable D24.1 [1] and briefly outlined in section 1.

To this purpose we will analyse the PACS and RIS workflows, that we have discussed in sect. 2.3.1 and 2.3.2, to understand which events of their lifecycles are relevant to the management of the authenticity, and to check how we can properly fit event the templates that have been suggested in [1] to model these events.

More precisely we aim at the following:

a) understand if all the relevant events in these specific lifecycles fit in one of the core set events that has been proposed in the model, or if additional events, not contemplated by the model, have to be considered;

b) which of the controls recommended in the guidelines for a proper management of the authenticity are part of the current practices;

c) which part of the authenticity evidence recommended by the guidelines is actually gathered.

At this stage we shall restrict our assessment to the adequacy of the model and the guidelines and to pointing out which authenticity evidence and which controls are missing. Conclusions will be drawn in sect. 2.5, which is devoted to the assessment of the current practices, to the proposal of improvements and to the definition of the authenticity management policy.

2.4.1 PACS workflow

The PACS workflow lifecycle is modelled in figure 2.6, according to the guidelines given in [1]. In the picture we have clearly identified the two phases of the lifecycle, the *pre-ingestion phase* and LTDP phase and six events that we consider relevant for the management of authenticity.

We shall discuss these events in the following subsections, in which represent them according to the templates in the model.

2.4.1.1 PACS-CAPTURE

- **Description: t**he DICOM study is generated by a modality and is delivered to a local PACS.
- **Agents**:
  - *author*: the person that was managing the modality when the study has been generated.
  - *keeping system administrator*: the administrator of the local PACS.
- **Input**: none
- **Output**: the DICOM study
- **Authenticity evidence record**:
  - *Identity of the DR*: DICOM study UID (part of DICOM tag).
  - *Date and time the DR has been created by the author*: creation date and creation time (part of the DICOM tag).
  - *Date and time the DR has been delivered*: the time the DICOM study has been delivered to the local PACS.
  - *Identification and authentication data of author(s):* part of the DICOM metadata.
  - *Identification data of the keeping system*: NO
  - *Identification data of the keeping system administrator* : NO
  - *Digest of the of the DR produced by the author*: NO
  - *Assessment by the keeping system administrator on the delivery of the DR and the subsequent controls*:
    - *Assessment on the identification and authentication of the author*: NO
    - *Assessment on the integrity check*: NO
  - *Context information associated to the DR*: NO
  - *Digest of the of the DR produced by the keeping system administrator*: NO

The event fits pretty well in the **CAPTURE** template: the agents, the action and the output are clearly identified. On the other hand only part of the authenticity evidence suggested by the template is actually gathered and preserved in the captured digital resource. Moreover some of evidence is gathered as part of the DICOM metadata, but is not explicitly represented as an authenticity evidence record, in the way the model suggests. Even if no digest is produced, communication DICOM parameter and DICOM tag are used to guarantee the authenticity.

2.4.1.2 PACS-TRANSFER

- **Description**: the DICOM study is delivered by the local PACS to the central PACS.
- **Agents**:
  - *owner*: the local Health Care system.
  - *origin system administrator*: the administrator of the local PACS.
  - *destination system administrator*: the administrator of the central PACS.
- **Input**: the DICOM study in the local PACS
- **Output**:
  - *DR'*: the original DICOM study plus additional metadata documenting the successful transfer.
  - *DR''*: the original DICOM study plus the authenticity evidence record (AER).
- **Authenticity evidence record**:

  Origin system
  - *Event type*: transfer-out

- *Identification data of the origin keeping system:* NO
- *Identification data of the destination keeping system:* NO
- *Date and time the DR has been sent*: NO
- *Identification and authentication data of the owner*: NO
- *Identification and authentication data of the origin system administrator:* NO
- *Evidence that the DR has been received and accepted by the destination system*: NO
- *Digest of the DR authenticated by the origin keeping system administrator*: NO

Destination system

- *Event type*: transfer-in
- *Identification data of the origin keeping system*: NO
- *Identification data of the destination keeping system:* NO
- *Date and time the DR has been received from the origin system:* NO
- *Identification and authentication data of the destination system administrator:* NO
- *Assessment by the destination keeping system administrator on the delivery of the DR by the origin keeping system and on the subsequent controls:*
  - *Identification and authentication of the origin keeping system:* NO
  - *Trustworthiness of the channel used of the data channel used for the transfer:* NO
  - *Integrity check on the digest produced by the origin system administrator:* NO
- *Digest of the DR produced and authenticated (digitally signed) by the destination keeping system administrator:* NO

The event fits pretty well in the TRANSFER template: the agents (but the owner), the input and the output are clearly identified. However, in the current practice no memory is kept in the origin system that the transfer has been performed successfully (and when), and, more importantly, no authenticity record is kept in either system.

One may possibly argue that the evidence that the guidelines suggest to gather is indeed very detailed and that in this particular case most of it is not necessary, since the transmission is performed on a secure channel and both system belong to the same administration. The argument is quite reasonable and in fact we suggest in the guidelines that simplification may be introduced in specific cases and taking into account the needs of the designated community. However some improvements may be necessary, as we shall discuss in section 2.5

2.4.1.3 PACS-SUBMIT

- **Description**: a SIP containing a DICOM study and a metadata file is delivered by the central PACS to the Scryba system.
- **Agents**:
  - *owner*: it is not clear who is the owner of the DICOM study.
  - *origin system administrator*: the administrator of the local PACS.
  - *LTDP system administrator*: the administrator of Scryba.
- **Input**: any DR in the central PACS
- **Output**:
  - *DR'*: the SIP plus the keeping system's authenticity evidence record (AER).
  - *DR''*: the SIP plus the LTDP system's authenticity evidence record (AER).
- **Authenticity evidence record**:

  Keeping system

  - *Event type:* submit-out
  - *Identification data of the keeping system:* NO
  - *Identification data of the LTDP system:* NO
  - *Date and time the DR has been prepared for submission:* NO

- *Identification and authentication data of the owner:* NO
- *Identification and authentication data of the keeping system administrator:* NO
- *The evidence that the DR has been received and accepted by the LTDP system:* NO
- *Digest of the DR produced and authenticated (digitally signed) by the keeping system administrator:* NO

LTDP system

- *Event type:* submit-in
- *Identification data of the keeping system:* NO
- *Identification data of the LTDP system:* NO
- *Date and time the DR has been received from the origin system:* NO
- *Identification and authentication data of the LTDP system administrator:* NO
- *Assessment by the LTDP system administrator on the delivery of the DR:*
  - *Identification and authentication of the keeping system:* NO
  - *Trustworthiness of the data channel used for the transfer:* NO
  - *Integrity on the digest produced by the keeping system administrator:* NO
- *Digest of the of the DR authenticated by the LTDP system administrator:* NO

The event fits pretty well in the SUBMIT template: the agents (but the owner) the input and the output are clearly identified. However, as in the previous case, the suggested authenticity evidence is not gathered, and we may substantially repeat the same considerations.

2.4.1.4 PACS-INGEST

- **Description: a** SIP containing a DICOM study and a metadata file is ingested and the corresponding AIP is generated.
- **Agents**:
  - *LTDP system administrator*: the Scryba system administrator.
- **Input**: the SIP, submitted by the central PACS.
- **Output**: the AIP containing the DICOM study.
- **Authenticity evidence record**:
  - *Event type*: ingest
  - *Original identifier of the submitted DR:* the DICOM accession number.
  - *New identifier of the DR in the LTDP system, if given:* the ID-DOC-Scryba
  - *Context information*: NO
  - *Date and time the DR has been accepted by the LTDP system:* NO
  - *Date and time the ingestion has been completed:* NO
  - *Identification data of the LTDP system:* NO
  - *Identification and authentication data of the LTDP system administrator:* NO
  - *Assessment by the LTDP system administrator on the ingestion of the DR and the subsequent controls:*
    - *Assessment on format migrations, if any, including a statement that the intellectual content of the DR has not changed, specifying the criteria that have been adopted to perform the assessment:* Not Applicable
    - *Assessment on the authenticity and provenance evidence in the submitted DR:* NO
  - *Digest of the of the AIP produced by the ingestion process*: NO

The event fits pretty well in the INGEST template: the agents, the action and the input and the output are clearly identified. However, only part of the suggested authenticity evidence is gathered, but, being part of the DICOM metadata, it is stored in the Content Information and not in the PDI (Preservation Description Information) as it should be.

2.4.1.5 PACS-AGGREGATE

- **Description:** several AIPs, each containing a single DICOM study, are aggregated in a Preservation Volume PV (AIC according to the OAIS terminology).
- **Agents**:
  - *LTDP system administrator*: the Scryba system administrator.
- **Input**: the set of AIPs which are aggregated in a single PV.
- **Output**: the newly created PV.
- **Authenticity evidence record**:
  - *Event type:* aggregate.
  - *Date and time the aggregation has taken place:* date and time of the certified timestamp.
  - *Identification data of the LTDP system:* NO
  - *Identification and authentication data of the LTDP system administrator:* the person who has digitally signed the PV (certificate in the PV index).
  - *Description of the criteria according to which the aggregation was performed:* NO (but are clearly stated in the management rules of the deposit).
  - *Identity of the aggregated AIPs:* DICOM accession numbers and ID-DOC-Scryba recorded in the PV index.
  - *Identity of the AIC:* ????
  - *Digest of the AIC generated (or modified) by the aggregation, authenticated by the system administrator:* the digital signature of the PV.

The event fits pretty well in the AGGREGATE template: the agent, the action and the input and the output are clearly identified. Moreover almost all the suggested evidence is gathered and stored in the PV index.

2.4.1.6 PACS-DELETE

As discussed above this transformation has not yet been implemented. We will give a few hints about its implementation in sect. 2.5.1.

2.4.2 RIS workflow

The RIS workflow lifecycle is modelled in figure 2.7, according to the guidelines given in [1]. In the picture we have clearly identified the two phases of the lifecycle, the *pre-ingestion phase* and *LTDP phase* and five events that we consider relevant for the management of authenticity.

We shall discuss these events in the following subsections, in which represent them according to the templates in the model.

2.4.2.1 RIS-CAPTURE

- **Description:** a report is written and digitally signed by a physician and is delivered to the RIS.
- **Agents**:
  - *author*: the physician who has written and digitally signed the report.
  - *keeping system administrator*: the administrator of the RIS.
- **Input**: none

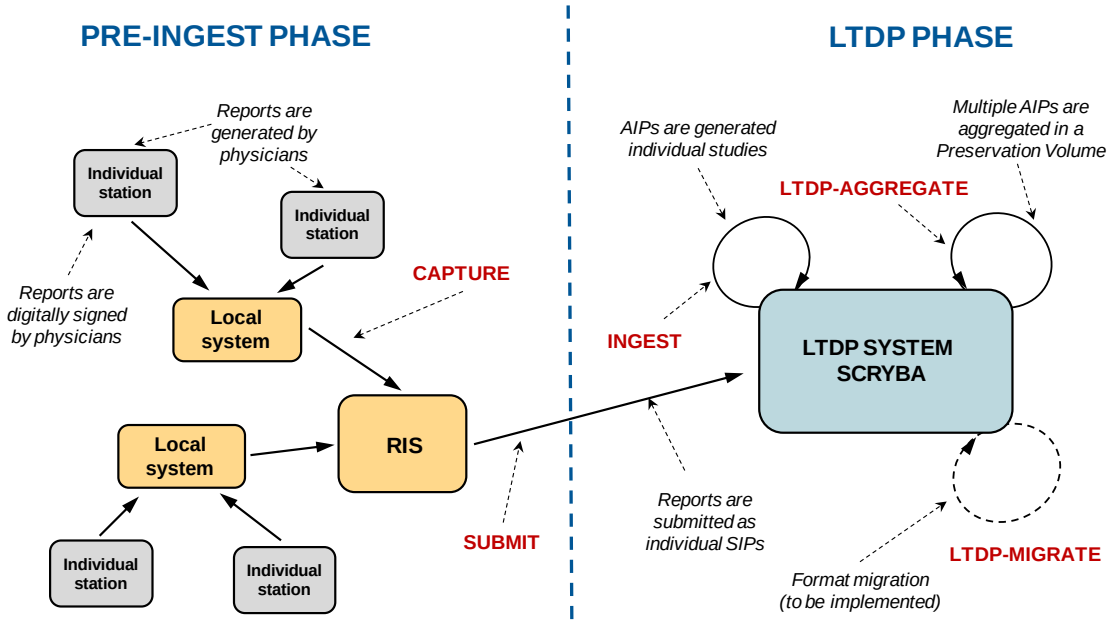


Figure 2.7 - The RIS workflow lifecycle

- **Output**: the digitally signed report, as managed by the RIS.
- **Authenticity evidence record**:
  - *Identity of the DR*: part of the HL7-CDA2 metadata.
  - *Date and time the DR has been created by the author*: creation date (part of the nel formato HL7-CDA2 metadata).
  - *Date and time the DR has been delivered*: the time the report has been delivered to the RIS.
  - *Identification and authentication data of author(s):* the person who has digitally signed the report.
  - *Identification data of the keeping system*: NO
  - *Identification data of the keeping system administrator*: NO
  - *Digest of the of the DR produced by the author*: the digest in the digital signature.
  - *Assessment by the keeping system administrator on the delivery of the DR and the subsequent controls*:
    - *Assessment on the identification and authentication of the author*: NO
    - *Assessment on the integrity check*: NO
  - *Context information associated to the DR*: the DICOM accession number of the study to which the report refers.
  - *Digest of the of the DR produced by the keeping system administrator*: NO

The event fits pretty well in the **CAPTURE** template: the agents, the action performed the input and the output are clearly identified. Moreover most of the authenticity evidence suggested by the guidelines is gathered, although not explicitly structured as an authenticity evidence record.

If we compare with the same event in the PACS lifecycle, the situation is considerably better. This is mostly due to the fact that reports are digitally signed, and the information in the signature is of crucial value for the fixity and provenance components.

2.4.2.2 RIS-SUBMIT

- **Description**: a SIP containing a digitally signed report in pkcs#7 format and a metadata file is delivered by the RIS to the Scryba system.
- **Agents**:
  - *owner*: the physician who has digitally signed the report.
  - *origin system administrator*: the administrator of the RIS.
  - *LTDP system administrator*: the administrator of Scryba.
- **Input**: any report in the RIS.
- **Output**:
  - *DR'*: the SIP plus the keeping system's authenticity evidence record (AER).
  - *DR''*: the SIP plus the LTDP system's authenticity evidence record (AER).
- **Authenticity evidence record**:

  Keeping system
  - *Event type:* submit-out
  - *Identification data of the keeping system:* NO
  - *Identification data of the LTDP system:* NO
  - *Date and time the DR has been prepared for submission:* NO
  - *Identification and authentication data of the owner:* NO
  - *Identification and authentication data of the keeping system administrator:* NO
  - *The evidence that the DR has been received and accepted by the LTDP system:* NO
  - *Digest of the DR produced and authenticated (digitally signed) by the keeping system administrator:* NO

  LTDP system
  - *Event type:* submit-in
  - *Identification data of the keeping system:* NO
  - *Identification data of the LTDP system:* NO
  - *Date and time the DR has been received from the origin system:* NO
  - *Identification and authentication data of the LTDP system administrator:* NO
  - *Assessment by the LTDP system administrator on the delivery of the DR:*
    - *Identification and authentication of the keeping system:* NO
    - *Trustworthiness of the data channel used for the transfer:* NO
    - *Integrity on the digest produced by the keeping system administrator:* NO
  - *Digest of the of the DR authenticated by the LTDP system administrator:* NO

The event fits pretty well in the SUBMIT template: the agents the input and the output are clearly identified. However, as in the case discussed in sect. 2.4.1.3, most of the suggested authenticity evidence is not gathered, and we may substantially repeat the same considerations.

2.4.2.3 RIS-INGEST

- **Description: a** SIP containing a digitally signed report and a metadata file is ingested and the corresponding AIP is generated.
- **Agents**:
  - *LTDP system administrator*: the Scryba system administrator.
- **Input**: the SIP, submitted by the RIS.
- **Output**: the AIP containing the digitally signed report.
- **Authenticity evidence record**:
  - *Event type*: ingest
  - *Original identifier of the submitted DR:* part of the HL7-CDA2 metadata.

- *New identifier of the DR in the LTDP system, if given:* the ID-DOC-Scryba
- *Context information*: the DICOM accession number of the study to which the report refers.
- *Date and time the DR has been accepted by the LTDP system:* from Scryba log
- *Date and time the ingestion has been completed:* date and time of the certified timestamp (see sect. 2.3.2.2).
- *Identification data of the LTDP system:* NO
- *Identification and authentication data of the LTDP system administrator:* NO
- *Assessment by the LTDP system administrator on the ingestion of the DR and the subsequent controls:*
  - *Assessment on format migrations, if any, including a statement that the intellectual content of the DR has not changed, specifying the criteria that have been adopted to perform the assessment:* Not Applicable
  - *Assessment on the authenticity and provenance evidence in the submitted DR:* the controls are performed while checking the digital signature and the validity of the certificate (see sect. 2.3.2.2)
- *Digest of the of the AIP produced by the ingestion process*: the digest in certified timestamp.

The event fits pretty well in the INGEST template: the agents, the action and the input and the output are clearly identified. Moreover a substantial part of the authenticity evidence suggested by the guidelines in gathered, although it is not clearly organized in the AER.

2.4.2.4 RIS-AGGREGATE

- **Description:** several AIPs, each containing a single report, are aggregated in a Preservation Volume PV (AIC according to the OAIS terminology).
- **Agents**:
  - *LTDP system administrator*: the Scryba system administrator.
- **Input**: the set of AIPs which are aggregated in a single PV.
- **Output**: the newly created PV.
- **Authenticity evidence record**:
  - *Event type:* aggregate.
  - *Date and time the aggregation has taken place:* date and time of the certified timestamp
  - *Identification data of the LTDP system:* NO
  - *Identification and authentication data of the LTDP system administrator:* the person who has digitally signed the PV (certificate in the PV index).
  - *Description of the criteria according to which the aggregation was performed:* NO (but are clearly stated in the management rules of the deposit).
  - *Identity of the aggregated AIPs:* DICOM accession numbers and ID-DOC-Scryba recorded in the PV index.
  - *Identity of the AIC:* unique Scryba identifier
  - *Digest of the AIC generated (or modified) by the aggregation, authenticated by the system administrator:* the digital signature of the PV.

The event fits pretty well in the AGGREGATE template: the agent, the action and the input and the output are clearly identified. Moreover almost all the suggested evidence is gathered and stored in the PV index.

2.4.2.5 RIS-MIGRATE

As discussed above this transformation has not yet been implemented. We will give a few hints about its implementation in sect. 2.5.3.

2.5 DEFINING THE AUTHENTICITY MANAGEMENT POLICY

2.5.1 General remarks

The discussion in sect. 2.4 proves that the model proposed in the guidelines fits quite well to both the PACS and the RIS workflows.

- All the events in both lifecycles which are relevant to the management of authenticity, namely changes of custody and transformations of the digital resources, fit well in one of the core set events. There has been no need to define new events.
- For each event the agent, the action performed the input and the output can be clearly identified.
- The templates in the model provide an effective way to represent the transformations undergone by the digital resources.

However, the lack of part of the authenticity evidence items that are recommended by the templates is likely to be the result of the following assumptions, which in turn are based on a general notion of trust.

- DICOM studies are going to be managed through all their lifecycle, both in the pre-ingest and the LTDP phase, by systems (local and central PACSs and Scryba) that are under the ownership of a single administration (the Vicenza Public health care system).
- All transfers among systems are carried on private lines that are managed by the same administration with reasonable security provisions.
- Access to the systems is given only to registered users, a proper rights management policy is enforced. Moreover access to DICOM studies is given in read mode only.
- DICOM studies, after they are generated, get to the preservation repository in a very short time, therefore threats to their integrity can be considered as minimal.

Part of these assumptions is reasonable, but some of them is questionable, and we shall suggest some of improvements later in this section, but altogether we may still rate the current practices in handling this event as acceptable, as long as one makes clear that:

a) the initial provenance information is gathered from the modality as DICOM file metadata;
b) no fixity evidence (e.g. authenticated hash) of the original digital resource is gathered;
c) no controls are performed and no evidence is documented when the DRs are transferred between keeping systems in the pre-ingestion phase;
d) at least for the PACS workflow, the integrity of data and metadata (including the initial provenance and authenticity information!) 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 (diagnostic images and medical reports) 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 (especially during the pre-ingest part of the workflow) in a more substantial way than relying on a general notion of trust in the infrastructure.

It can be argued that substantial evidence in proving the integrity could come from system logs and from a proper documentation of the rights management policies and their implementation. This is reasonable but raises the further question of how long this information is maintained and how it is preserved.

Therefore we would like suggest that some additional authenticity evidence should be preserved to allow proving at least that the integrity of the DRs has been maintained and to allow tracing their ownership along their lifecycle. More precisely, in our opinion, the following additional authenticity evidence should be gathered and preserved:

- an authenticated hash of the digital resource at the time of the capture;
- for every subsequent transfer of the digital resource, a record of the time of the transfer and the identification of the source and destination system administrators.

A last remark is on *where* the authenticity evidence is stored, since, as we have remarked in the previous sections, some sometime the authenticity evidence is part of the content information (e.g. part of the DICOM file) instead of being explicitly managed as metadata. Once the substantial issue of collecting the information has solved, this may be considered as a secondary issue at implementation level, but we should nevertheless recommend that all the authenticity evidence should be clearly outlined and stored as metadata, in a dedicated structure.

2.5.2 PACS workflow

According to the previous remarks we propose that the following authenticity evidence records should be gathered.

2.5.2.1 AER for PACS-CAPTURE

- *Event type*: capture
- *Identity*: the DICOM accession number.
- *Date and time created*: creation date (extracted from DICOM metadata).
- *Date and time delivered*: generated by the local PACS.
- *Identification and authentication data of author(s):* extracted the DICOM metadata.
- *Identification data of the keeping system administrator*: generated by the local PACS
- *Digest*: generated by the local PACS (possibly authenticated by the administrator)

2.5.2.2 AER for PACS-TRANSFER

Origin system (local PACS)

- *Event type*: transfer-out
- *Date and time sent*: generated by the local PACS
- *Identification and authentication data of the system administrator:* generated by the local PACS.

Destination system (central PACS)

- *Event type*: transfer-in
- *Date and time received:* generated by the central PACS.
- *Identification and authentication data of the system administrator:* generated by the central PACS.

2.5.2.3 AER for PACS-SUBMIT

Origin system (central PACS)

- *Event type*: submit-out
- *Date and time sent*: generated by the central PACS
- *Identification and authentication data of the system administrator:* generated by the central PACS.

Destination system (Scryba)

- *Event type*: transfer-in
- *Date and time received:* generated by Scryba.
- *Identification and authentication data of the system administrator:* generated by Scryba.

2.5.2.4 AER for PACS-INGEST

- *Event type*: ingest
- *Original identifier:* DICOM accession number.
- *New identifier in the LTDP system:* ID-DOC generated by Scryba
- *Date and time the ingestion has been completed:* generated by Scryba
- *Identification and authentication data of the LTDP system administrator:* generated by Scryba

2.5.2.5 AER for PACS-AGGREGATE

- *Event type:* aggregate.
- *Date and time of the aggregation:* from the certified timestamp.
- *Identification and authentication data of the administrator:* from the digital signature of the PV
- *Identity of the aggregated AIPs:* DICOM accession numbers and ID-DOCs from the PV index.
- *Digest of the AIC authenticated by the system administrator:* the digital signature of the PV.

2.5.3 RIS workflow

According to the previous remarks we propose that the following authenticity evidence records should be gathered.

2.5.3.1 AER for RIS-CAPTURE

- *Event type*: capture
- *Identity*: report identifier and version identifier from the HL7-CDA2 metadata.
- *Version*: from HL7-CDA2 metadata
- *Date and time created*: from the HL7-CDA2 metadata.
- *Date and time delivered*: generated by the RIS.
- *Identification and authentication data of author(s):* the certificate of the digital signature.
- *Identification data of the keeping system administrator*: generated by the RIS
- *Context information*: the DICOM accession number of the study to which the report refers
- *Digest*: the digital signature

2.5.3.2 AER for RIS-SUBMIT

Origin system (RIS)

- *Event type*: submit-out
- *Date and time sent*: generated by the central RIS
- *Identification and authentication data of the system administrator:* generated by RIS.

Destination system (Scryba)

- *Event type*: transfer-in
- *Date and time received:* generated by Scryba.
- *Identification and authentication data of the system administrator:* generated by Scryba.

2.5.3.3 AER for RIS-INGEST

- *Event type*: ingest
- *Original identifier:* identifier from the HL7-CDA2 metadata.
- *New identifier in the LTDP system:* ID-DOC generated by Scryba
- *Context information:* the DICOM accession number of the study to which the report refers.
- *Date and time the ingestion has been completed:* from the certified timestamp
- *Identification and authentication data of the LTDP system administrator:* generated by Scryba
- *Assessment on the authenticity and provenance:* outcome of controls on the digital signature
- *Digest of the of the AIP*: from the certified timestamp.

2.5.3.4 AER for RIS-AGGREGATE

- *Event type:* aggregate.
- *Date and time of the aggregation:* from the certified timestamp.
- *Identification and authentication data of the administrator:* from the digital signature of the PV
- *Identity of the aggregated AIPs:* DICOM accession numbers and ID-DOCs from the PV index.
- *Digest of the AIC authenticated by the system administrator:* the digital signature of the PV.

2.6 IMPLEMENTATION OF THE AUTHENTICITY PROTOCOLS

2.6.1 Implementing the CASPAR proposal at operational level

According to the guidelines defined in D24.1 [1], the next step in implementing the authenticity management policy we have defined in the previous subsections 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. As discussed in [1] we propose an implementation strategy which is based on the concept of *Authenticity Protocol* that has been introduced within the CASPAR project, and that we have conveniently extended and adapted to fit our needs. For the reader's convenience, and to avoid confusion, we briefly summarize in this section the main points of our extended definition of authenticity protocol.

According to the definition in CASPAR [3] an *Authenticity Protocol (AP)* is the definition of the procedure that must be followed in order to assess the authenticity of specific type of DR. More precisely, an AP is an ordered sequence of interrelated steps, each one of which we will refer to as an *Authenticity Step (AS)*. Each AS is performed by an *actor*, which can act either in an automatic or in a manual way. The execution of an AP generates an *Authenticity Protocol Report (APR,)* that documents that the sequence ASs has been executed and collects all the values associated with the data elements analyzed in every AS, and possibly the outcome of the execution.

We have therefore resorted to the CASPAR definition and adapted it to our purposes in order to formalize the process of performing controls and collecting authenticity evidence in connection with the lifecycle events in the way specified by the authenticity management policy. More precisely, in our case, an AP becomes the procedure that is to 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, instead of an Authenticity Protocol Report 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. Moreover each AP will operate on the authenticity evidence collected so far, that is on the Authenticity Evidence History, which is the sequence of all the AERs gathered for the previous lifecycle events.

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 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. We therefore need only give a complete example for a single event in the following section.

2.6.2 Authenticity protocol for RIS-INGEST

In this section we give the example of a specific event of the RIS workflow: RIS-INGEST.

As discussed in the previous section, 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.

Therefore, to understand the development of the protocol one should consider:

- the general description of the RIS workflow (see sect. 2.3.2);
- the description of the RIS-INGEST event (see sect. 2.4.2.3)
- the specification of the AER for the RIS-INGEST event (see sect. 2.5.3.3)

According to this information we may easily draw the protocol:

- *DR type*: RIS workflow digitally signed medical reports
- *Event type*: RIS-INGEST
- *Agent*: the administrator of the Scryba system (*Responsabile della conservazione*)
- *AER*: the as defined in section 2.5.3.3
- *AS sequence*: steps from **AS-1** to **AS-12**

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 to verify the certificate has not been revoked

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 file by using the hash algorithm specified in the digital certificate in the pkcs#7 file

AS-2.2: decrypt the digital signature in the pkcs#7 file by using the public key in the digital certificate to generate the hash file that has been signed

AS-2.3: compare the two hash files generated in steps AS-2-1 and AS-2.2 to verify they are identical

AS-2.4: if the check in **AS-2.3** fails then abort ingestion

Step AS-3 - Check context

AS-3.1: extract the DICOM accession number of the study to which the report refers from **AER RIS-CAPTURE**

AS-3.2: check the Scryba DB to verify that a report exists with the DICOM accession number generated in step **AS-3.1**

AS-3.3: if the check in **AS-3.2** fails then abort ingestion

Step AS-4 - Check unicity

AS-4.1: extract the report identifier and the version identifier (from the HL7-CDA2 metadata) from the **AER RIS-CAPTURE**

AS-4.2: check unicity of the report in the Scryba database: the report is considered unique and accepted if no report exist in the repository with the same value of the three parameters:

- Report ID (from **AER RIS-CAPTURE**)
- Version ID (from **AER RIS-CAPTURE**)
- Hash (from **AS-2.1**)

AS-4.3: if the check in **AS-4.2** fails then abort ingestion

Step AS-5 - Generate internal identifier

AS-5.1: generate an internal unique identifier that identifies the DR in the repository

Step AS-6 - Generate timestamp

AS-6.1: generate an hash file of the content information of the AIP (that is of the pkcs#7 file containing the signed report)

AS-6.2: send the hash file generated in **AS-6.1** to the Certification Authority to get a certified timestamp; the timestamp certifies the DR existed at the time of the certification with the content corresponding to the hash file, and is preserved as part of the PDI.

Step AS-7 - Generate AEI: Original identifier

AS-7.1: generate AEI *Original identifier* in the AER INGEST which is given the value extracted in **AS-4.1**.

Step AS-8 - Generate AEI: Internal identifier

AS-8.1: generate an internal unique identifier for the DR in the Scryba system

AS-8.2: generate AEI *New identifier in the LTDP system* in the AER INGEST which is given the value generated in **AS-8.1**

Step AS-9 - Generate AEI: Context information

AS-9.1: generate AEI *Context information* in the AER INGEST which is given the value extracted in **AS-3.1**.

Step AS-10 - Generate AEI: Date and time

AS-10.1: extract date and time from the certified timestamp

AS-10.2: generate AEI *Date and time the ingestion has been completed* in the AER INGEST which is given the value extracted in **AS-10.1**.

Step AS-11 - Generate AEI: Administrators data

AS-11.1: generate AEI *Identification and authentication data of the LTDP system administrator* with the Scryba system administrator data

Step AS-12 - Generate AEI: Assessment on authenticity and provenance

AS-12.1: generate AEI *Assessment on authenticity and provenance* which documents the outcome of the checks performed in **AS-1** to **AS-4**

Step AS-13 - Generate AEI: Digest of the AIP

AS-13.1: generate AEI *Digest of the AIP* which is given the value of the hash file generated in **AS-6.1**.

2.7 CONCLUDING REMARKS

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

- according to the Italian legislation, in the health care sector all the documentation (including of course digital files) has to be preserved for an indefinite period of time, potentially forever;
- the designated community has a clear interest (and a strong commitment) in the problem of properly managing authenticity and provenance of DRs since when dealing with the results of medical tests and medical reports, authenticity becomes a crucial issue both for the legal value of the records, to properly allocate responsibilities, and to the scientific community as important experimental data;
- the repository manages a variety of DRs (currently five different types), each organized on a different workflow, and the two workflow that we have analysed in more detail have quite a reasonable lifecycle complexity, i.e. there are several changes of custody and transformations of the DRs, thus allowing a thorough test of our model and guidelines;
- 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 a good ground for gathering authenticity evidence.

The model and the definition of the core set of events have proved to be robust enough, since they allowed us to easily accommodate all the transformations and the changes of custody in the workflow. On the other hand, the templates for the authenticity evidence records that we have given in the definition of the events (see D24.1 sect. 4.2 and 4.3) have provided a comprehensive checklist to verify which authenticity evidence was actually gathered in the current practices of the repository, and to understand which was missing and should be possibly taken into account.

Altogether, in the first part of the case study (see sect.2.4), where the current practices have been analysed, the methodology has proved to be a quite effective tool to assess the policies and the procedures used in the repository for the management of authenticity and provenance. However, it is important to give this assessment the correct interpretation, and to understand that the mere fact that part of the authenticity evidence in the AER templates is missing in the current practices is not in itself a problem. The assessment should just be used to point out that one needs to carefully check if, in the specific environment we are dealing with, the gathering of that evidence is actually necessary, and, of course, to justify why it may be disregarded.

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

In the case of the Vicenza health care system, the main point is that the DRs are managed throughout all their lifecycle, both in the pre-ingest and the LTDP phase, by systems (local and central PACSs and Scryba) that are under the ownership of a single administration (the Vicenza Public health care system). Moreover a strict rights management policy is enforced and the DRs after they are created get to the preservation repository in a very short time. Threats to their integrity should be therefore considered as minimal.

According to these considerations, a reasonable authenticity management policy has been defined (see sect. 2.5 and 2.6), which properly fits the needs of the designated community and allows to a proper definition of the AERs, and to formalize, for each event in the lifecycle, the authenticity protocol, i.e. the specification of the controls and the actions that are performed in connection with the event. The

policy suggests a few more authenticity evidence items that should be produced and gathered, and, of course, a more systematic organization of the authenticity evidence. All the suggested improvements are indeed feasible and are currently taken into careful consideration by the management of the repository.

3 SOCIAL SCIENCE DATA REPOSITORY AT THE UK DATA ARCHIVE

The UK Data Archive provides a case study of current practice and future plans in the context of the authenticity and provenance model proposed under the work package.

With the exception of caveats over the need for clear rights, information transfer related to the key lifecycle events described the model provides an accurate baseline with appropriate prioritisation. The UKDA agrees with the work package's evaluation of current practice in the production environments of LTDP system as somewhat distant from the current debates on best draft practice in academic circles. The case study outlines the Archive's strategy of concentrating on clear governance, records management and information security model to prepare for the availability of agreed best practice and clarifies that this needs to be supported by the availability of common standards and standard-compliant tools. The UKDA considers these to be appropriate areas of debate and action for a network of excellence such as that under development by the APARSEN project and evaluates the logical sequence of next steps for the network and the repositories as:

- Agreement on clear best practice in the areas of authenticity and provenance
- Conversion of those conceptual best practices into an accurate structured data model/schema
- Guidelines on technical and procedural implementation for:
  - The development of supporting data/metadata capture and management tools
  - The deployment of such tools throughout the DR lifecycle.

3.1 THE UK DATA ARCHIVE: STRUCTURE AND CONTEXT

Founded in 1967, the UK Data Archive (http://www.data-archive.ac.uk) based at the University of Essex is a specialist repository dealing mainly with materials for social science and humanities research. It holds over 5,000 data collections relating to society and the economy. The majority of the material is quantitative, but there are also qualitative and audio-visual collections.

The UKDA has hosted the Economic and Social Data Service for a decade and more recently it provides secure access to sensitive data through the Secure Data Service. The UKDA is engaged in a number of data management initiatives.

The UKDA's activities as a centre of expertise in data acquisition, preservation, dissemination and promotion are largely funded by the ESRC, the JISC and the University of Essex.

Links to the public sites of some of our key services providing continuous access to data from the academic, public, and commercial sectors are listed below.

- Economic and Social Data Service (ESDS) (http://esds.ac.uk)
- Secure Data Service (http://securedata.data-archive.ac.uk/)
- Census.ac.uk
- History Data Service (http://hds.essex.ac.uk/)

3.2 AUTHENTICITY, PROVENANCE AND THE DESIGNATED COMMUNITY

3.2.1 The Producers

The UK Data Archive deals mainly with Government and academic data producers, though in theory anyone may offer materials for deposit with the UK Data Archive. All offers are evaluated for their relevance to research and teaching and their relationship to the UKDA thematic scope (in line with the UKDA *Collections Development Policy*). An agreement is held with the ESRC whereby its grant holders are required to offer their data outputs for deposit with the Archive. Agreements are held with many of the other major funding bodies for the social science disciplines. Extensive links with government departments and with the Office for National Statistics (ONS) have been established resulting in the acquisition of many datasets sponsored or collected by the ONS.

All these individuals and institutions are Producers in the OAIS sense of a Producer/Archive interface but for business process management at the UKDA further agent designations are required to clarify the roles in more detail including depositor, rights holder and grant holder.

3.2.2 The Consumers

The UK Data Archive's Preservation Policy offers the following definition:

> *The UK Data Archive's designated user community is made up of social science and related data users within HE and FE in the UK, though best efforts are made for all users. All users are expected to have a basic understanding of social science methods and techniques relevant to the data collections being accessed. Additional requirements will be expected In the case of users of the Secure Data Service.*

It is explicitly understood that there is a significant overlap between registered users of our services who download data collections and of current and future data producers via original deposits and secondary analysis of the materials in our current collection.

The key audiences for the UK Data Archive range from school children to professors, commercial researchers and members of the public.

3.2.3 Relevance of authenticity to the designated community

The authenticity of digital collection is inherent in the UKDA's approach to long term digital preservation:

> *"The archival organisation is also responsible, in selecting material for archiving, for ensuring the reliability and logical integrity of the data collection. For paper-based documents, the reliability and integrity of elements of a data collection are usually assured by a simple audit trail; for digital material these requirements cannot be guaranteed in such a straightforward fashion, because some significant properties of a data collection may have to be altered in order to ensure a level of software independence and guarantee long-term usability.*
>
> *Any strategy for the long-term preservation of any electronic information must address the issue of software dependence. For most electronic information it is generally possible to eliminate software dependence by sacrificing structure, but the end products of these transformations are not authentic versions of the original. In these cases the authenticity needs to be re-established through the documentation of the actions taken and validation that the substantive content has not been altered.*
>
> *Thus the primary goal of the Archive's preservation policy is to ensure the long-term accessibility of electronic information while ensuring the highest level of authenticity of any formats disseminated. In effect this means that all the inherent qualities of the electronic information upon which their authenticity depends are preserved."*
>
> (Edited extract from the UK Data Archive Preservation Policy)

Though there is a consistent and general understanding of the importance of 'preservation' the increased use of the term 'continued access' in its place goes some way to explaining the perception of the role of preservation among stakeholders in the digital objects' lifecycle; the focus of the designated community is generally on their confidence in the institution to provide continued access to resources with data of sufficient quality for replication of results with sufficient documentation for accurate interpretation of the content information.

A recurring issue in dealing with the designated community and by extension with producers and funding bodies is that many aspects of the preservation process are not high-profile issues among these stakeholders. Furthermore the explicitly OAIS understanding of these terms is specialised to the Archival community rather than to the designated community so a direct comparison of interpretations by all stakeholders is challenging. The concept of authenticity in the limited interpretation of 'data

integrity' is gaining ground but much of this is actually driven by initiatives from the Archive in supporting the community in developing pre-deposit best practice in data management and sharing (http://www.data-archive.ac.uk/create-manage). Even when these best practices become part of the general community understanding this does not necessarily translate into best practice Producer/Archive interfaces if this involves the development of new skills (including fixity checking and encryption) by depositors.

Beyond data integrity in the strict computational sense and the issue of assessing data for its fitness for purpose through the replication of results there is little understanding of the concept of 'Authenticity' in the strict archival sense and no obvious analogous terms are perceptible in our extensive customer service interactions with our Designated Community.

The definition of authenticity provided in the UK Data Archive's Preservation Policy is adapted from the ISO 15489-1:2001.

> *A record's authenticity can be said to rely on three significant provable properties. The three properties are: that the record is what it purports to be, that it was created by whomsoever it is purported to have been created by; and that it was created at the time when it is purported to have been created.*

In the context of the response above it is clear that explicit evidence of authenticity is primarily an internal archival goal rather than an issue prevalent in our interactions with the Designated Community. The focus of our users is on data which has sufficient supporting documentation to be interpretable in the first instance and sufficient change log information over time to identify whether analyses of currently available releases are sufficient to replicate results of analyses run against earlier releases (as currently historical series of releases are not available as DIPs).

The term 'trust' has a specific meaning in our sphere, specifically since the signing of a memorandum of understanding for a European Framework for Trusted Digital Repositories between three groups which are working on standards for Trusted Digital Repositories being David Giaretta in capacity as Chair of the CCSDS(Consultative Committee for Space Data Systems)/ISO Repository Audit and Certification Working Group (RAC), Henk Harmsen in his capacity of Chair of the Data Seal of Approval (DSA) Board and Christian Keitel in his capacity as Chair of the DIN Working Group "Trusted Archives - Certification" and since that time the ISO16363 on Trusted Digital Repositories has received formal certification. As an Archive which has self-audited against ISO16363 as part of the APARSEN efforts we will avoid the term trust and instead settle on the concept of 'confidence' between the various stakeholders we serve. Trust implications are examined further in part two of this use case.

Much of the material demanded by and supplied to users is intended to support the interpretation of data collections, the validity of that content and of our underlying processes is delivered at an institutional level through long-standing relationships, close cooperation and an open approach to the sharing of procedural standards. The importance of the attachment of the specifics of processing details to individual data collections has long been understood by the UK Data Archive but is not explicitly supported by the Data Documentation Initiative version 2.1 which to date is the Archives primary system standard. With this in mind detailed documentation is provided primarily in semi-structured prose forms as 'Read' files which accompany each data download. Where a significant property might be impacted by an ingest event (or to align with the latest revision of OAIS, when a Transformational Information Property is actively retained after a transformation) these are documented but a significant proportion of preservation actions taken on the data collection are not deemed relevant to interpretation and are instead stored as internal-only Notes files. These events take place but are not recorded as explicit discrete events in terms of structured inputs, outputs and results.

As data standards such as DDI3 are implemented we would expect to have greater control over the presentation of granular details about processing without too great an impact on processing time. As this implementation continues we will endeavour to share additional evidence more widely (for best

practice reasons rather than to meet any perceived demand) but this depends on our being able to deliver the information in a way which does not obscure the contextual information prioritised by the designated community. For the majority of our designated community the Archive is a recognised or required place of deposit for materials and has been performing related services for such a long period that a high level of confidence in those services is inherent in the community.

3.3 TYPOLOGY OF THE PRESERVED DIGITAL RESOURCES

UK Data Archive business processes divide data into Government and Academic:

- 'Government data' are those data collections that are produced directly by government departments or by private sector companies or commercial companies on behalf of those departments. They are often regular or major series deposits. Also derived variables or additional data for such data collections, which may be produced by academics, fall within this group.
- 'Academic data' are data collections produced by academic researchers. All researchers funded by the ESRC or British Academy are contractually required to offer their data for deposit. Other researchers may equally offer data to the Archive. ESRC data collections are acquired on behalf of the Economic and Social Data Service.

The UKDA's *Collections Development Policy* notes that data is acquired "for four central purposes:

- secondary use and analysis for research;
- teaching and learning use;
- replication and validation of research;
- archival preservation;"

Within the thematic scope of our collections the materials are qualitative and quantitative data (see section 3.5 for more details on the impact of these two types of data on ingest processing), increasingly in mixed methodology data collections.

3.4 THE SUBMISSION TO THE REPOSITORY

3.4.1 Authenticity evidence delivered by the producers

The keeping systems implemented by producers/depositors are outside the remit of the UK Data Archive in terms of both access and authority. As full lifecycle metadata in the form of DDI3 and associated standards is implemented at the Archive we may have further opportunities to provide support. The level of evidence required for deposit will also necessarily change as depositors and the designated user community seek to archive and access administrative data from government transactional databases.

The UK Data Archive deals directly with human agents for all acquisition negotiations. Custodial transfer and custody history are not available in any machine-actionable or controllable way and no formal data standard is adopted across our depositors sufficient for the normalization of these processes. Contact is with a designated depositor (person/institution) and full history of the content information is not available.

Most of the provenance metadata (explicit or implied) is recorded during the selection process through data deposit forms and licences rather than formally collected by Producers and incorporated into a SIP.

The Archive is keen to close this gap in the interface not only for reasons of best practice but also because such a development would greatly reduce ingest processing time.

The Deposit Licence Agreement or other deposit contracts (managed via the *Rights and Access Strategy*) govern the rules under which data collections are deposited. Deposits of data must be accompanied by a Data Collection Deposit Form and Deposit Licence. Deposit Forms provide an

opportunity to deliver standard metadata alongside the SIP but minimum standard metadata is always assured during ingest.

Acquisitions Procedures are fully documented and the standard Data Deposit Form requests details such as:

- Data and Documentation file formats
- Data Collection description data including File name, file format and content information for data, supporting documentation and any other materials provided
- A list of any hard copies sent accompanying electronic submission
- Weighting used in any quantitative analysis alongside further details explaining weighting choices provides provenance information related to changes made since the original data collection
- Software details of any compressions software used
- Full details of key contact (including department/section/institution) including
  - Depositors
  - Data Creators
  - Award Holder
  - Data Collectors and
  - the details and role of "any other person or organisation responsible in any significant way for these data"

The collection of detailed information about key individuals and their roles in the digital resources lifecycle provide us with some provenance information and implied evidence related to the custodial history prior to deposit. This is further augmented by questions relating to funding sources including grant funding identifiers where applicable.

It is clear that beyond weighting (which is covered in more detail in the appendix) only limited knowledge and information about transformations which have taken place earlier in the lifecycle is available.

3.4.2 Identification and authentication of the producers

Depositors in effect self-identify, the UKDA noted in its self-audit against ISO16363 in response to metric 4.1.4 (The repository shall have mechanisms to appropriately verify the depositors of all materials) noted:

> *"The process of pre-ingest negotiation will tend to ensure that Archive staff can identify the depositor by validating their institution and/or funding body. In most cases there is an on-going relationship with data owners. The Archive is working to increase the strength of this aspect of procedures in the future but has never experienced an incident where a depositor has been misidentified."*

Internally developed software is used to manage depositors' details and a controlled vocabulary of their associated organisations. Formal checks are undertaken to verify new (or changed) institutions in the system and unless a prior relationship exists offers of data are fully reviewed (with standard procedures and sample data) by the UKDA's Acquisitions Review Committee (in line with agreed terms of reference) prior to an offer being made to a depositor.

During this detailed pre-ingest phase relations with depositors are handled by dedicated managers for Academic and Government data and the UK Data Archive provides a legal framework within which materials may be accessioned into the Archive.

The UKDA is aware that more procedures for capturing authenticity and provenance at this stage are desirable and implementation will become more practical as best practice standards are developed

within the archival community. This will be examined during our migration to more lifecycle-oriented data standards and expanded upon in section two of this case study.

3.4.3 Secure transfer

Submission of SIP to the UK Data Archive is restricted to more secure methods over time as Producers adopt best practice approaches but in most case the Archive has no authoritative control over these. Acquisitions Procedures applicable to all non-sensitive deposits state that:

> *"the data and documentation may be sent by via the University of Essex ZendTo Service; by mail on CD, DVD or memory stick; or via secure electronic transmission. If data files contain sensitive or personal information, they should be encrypted before submitting"*

For reasons of the legal admissibility of signatures Licence Agreements are currently transmitted via the postal service.

The Secure Data Service (whose limited number of depositors overlap with standard depositors) has a higher information security classification for all materials and adds a requirement for SIP checksums which has yet to be mandated for all depositors

> *"SDS data depositors must be instructed to send the dataset as encrypted files, together with checksums, (with the password/pass-phrase sent separately by email), either via the University of Essex's ZendTo system, on SDS's secure FTP, or on a CD (or other removable storage device) posted to a named recipient."*

3.4.4 Authenticity evidence documenting the submission process

As noted above the majority of pre-deposit authenticity evidence is related to or collected during the acquisition negotiation and submission process. Some (but not all) of the provenance information related to the Acquisition process is formally preserved whereas other information is retained in acquisition administrative databases. The database of individuals and institutions is checked for accuracy with each submission but not formally preserved, however all relevant details of those involved are recorded in deposit forms and licences which are preserved.

The Acquisition negotiation is assigned a unique identifier in internally developed software and all materials are maintained. All new academic data offers are reviewed by the Acquisitions Review Committee (ARC). The ARC terms of reference notes that "All decisions taken at ARC meetings are minuted, noting details of why a decision was made and any issues which need to be flagged up with either the depositor or the processing team. Details of any embargo or access restrictions agreed by ARC are also recorded, as are any comments made. However these records relating to the deposit are not formally preserved.

During Acquisition negotiation access permissions are also agreed with any 'Special Conditions' being agreed based on a general Special Licence (based on one used with the UK Office of National Statistics) and additional access conditions attached to End User Licences which must be agreed to by consumers prior to access.

Upon delivery of the SIP a formal acknowledgement email is sent to the depositor and passwords are then requested and transmitted in line with standard procedures.

All information required is confirmed as received (licence, deposit form, data and supporting documentation) and copyright and consent issues are formally recorded as having been resolved. In addition to sign off of these actions special access permissions and any embargo periods agreed are recorded in local applications.

All copyright holders named in licences are cross-checked against all (co-)applicants of an award; if any are missing they are contacted; if they do not want to be on the licence, written confirmation is obtained.

3.5 THE INGESTION PROCESS

The UKDA's Preservation Policy states:

- the datasets it accessions are accompanied by adequate documentation to enable their use for analytical and research purposes;
- the datasets are checked and validated according to strict data ingest procedures;
- the data collections are professionally catalogued according to appropriate metadata standards;
- the data collections are indexed with keyword terms using the Archive's thesaurus;
- the datasets, documentation, metadata and other representation information are kept in conditions suitable for long-term archival storage;
- the authenticity, integrity and reliability of datasets preserved for future use are retained;
- the basic preservation actions undertaken by the Archive are uniform regardless of the perceived value of any dataset. (Events within the preservation process may differ from data collection to data collection but the actions specific to data preservation may not.)

The handover of control for a SIP from the Acquisitions team to the Ingest team is the last point in the process that a unique data collection identifier (Study Number) may be allocated. There is not necessarily a one to one relationship between an Acquisitions ID and an Data Collection Identifier as decisions on how to present (and therefore structure and identify) the data collection may be made during the Acquisitions process.

With the exception of Secure Data Service materials which are delivered into an independent system the application of checksums to SIP is not mandated. As the take up of fixity check advice is limited among depositors this is unlikely to occur until full lifecycle metadata standards adoption is complete and tools can be provided to support depositors in this activity.

Though no action is taken on the original SIP materials during the ingest process no system-wide controls are in place to prevent malicious tampering. Our collections development policy states that we seek data "of a type with which the Archive has expertise or may easily obtain expertise or expert advice" which reduces the risk of accidental amendments by non-expert staff and all released data collections are checked directly and against the released 'processing notes' by the original depositor.

Detailed procedures for all types of data collection are provided in the documents UKDA081 DS Quantitative Data Processing Procedures, UKDA093 DS Qualitative Data Processing Procedures, UKDA078 DS Documentation Processing Procedures and UKDA079 DS Data Processing Standards. These are made available to depositors and consumers as part of the Archive's commitment to transparent processing standards.

All processing actions are recorded in the Read (external access with the data collection) or internal-only Notes files unless included as standard fields in the internally developed ingest software.

Prior to any actions being taken on a data collection a pre-processing review is completed and documented which includes (but is not limited to) the following checks:

- completeness of study/data collection (i.e. whether all materials have been received)
- data and documentation confidentiality
- whether documentation coverage is sufficient

All processing plans and software reports generated as part of the ingest process are retained as part of the AIP as are all Read and Notes files.

The close relationship between the Archive and depositors extends into the ingest process. The examples of possible transformations and associated quality control here help to indicate the scope of the ingest task and the number of points in the process that problems could be detected that may be referred back to depositors.

This relationship is an important part of the services offered by the Archive but further blurs the custodial handover of the SIP in ways which will be addressed in part two of this case study.

Similarly due to the sheer number of potential processing steps and the absence of clear best practice guidance in the preservation community, and more importantly the absence of reference technical implementations of such guidance in software the transformations are documented at a high level with relevant outcomes reported rather than benefitting from clear event-based machine-controlled workflows.

The UK Data Archive maintains a limited list of acceptable deposit formats and maps these to acceptable preservation formats at a procedural rather than a per-data collection level.

Actions on different data types are outlined below but it should be made clear that mixed methodological studies with all referenced data types are increasingly common. The appendix outlines a limited number of possible transformations and quality control actions undertaken to support evaluation against the model.

3.6 THE LONG TERM DIGITAL PRESERVATION PHASE

The final stage of ingest processing differs from the standard OAIS model as AIP and component DIP files are created by the ingest team and submitted to Archival Storage together.

The Ingest Team transfers ingested files into a designated workspace.

3.6.1 Adding data collections to the Archival Storage

The Ingest Team submits a request for addition of the data collection to Archival Storage via a helpdesk request. The request has a standard Archival Storage designation.

Upon receipt of a Archival Storage Update request the following actions are undertaken:

- **Locking changes.** Running the structural validation script changes ownership of the relevant ingest data collection directory from the data services agent to the preservation control.
- **Structural Validation.** A Unix shell script structural validation which reports against the submitted data collection:
  - non-standard directory names (compares against a template of standard directory names)
  - deviations from appropriate file extensions in directory
  - checks study contents against the label file to ensure there is a single file label for each group of files with the same name but different extensions (file.txt, file.pdf, file.doc etc)
  - Checks filenames for illegal characters in the windows or Unix filenames including case matching

  Upon successful validation deletion parameters are added to standard archival storage scripts as follows:
- **Delete.** Submitted to remove files before an Update request. The help desk item requests the deletion of specific files or folders within a study
- **Update.** Files are copied into the existing study directory, overwriting any existing files of the same name.
- **Add New.** A request for the creation of a completely new study
- **Decat.** The study is 'decatalogued' by moving all data collection files into a Decat directory (see and up one level in the hierarchy
- **Archival Storage Preview.**
  - On screen preview of selected changes/editions/deletions
  - Human validation process prior to commit
- **Archive the Study.** The *Archive Study* script is run which:

- copies the appropriate study from the the working server area to the preservation server
- validates the copied content against the log file
- adds or updates a zero byte file to the dp level study directory with a current date time. Unix command forces update.

  Note: the UNIX touch program is used to add a 'current date time' touch because the files copied to the preservation server retain the last modified date stamps from the Data Services processing.
- Originals are deleted from working server area

- **Transfer DIP to Access.** The DIP files are generated by the ingest team rather than converted on the fly in response to access requests but generation and verification of the zip bundles available to users takes place from within the Archival Storage system**.** A separate script:
  - checks the directories daily
  - checks the touch times of the zero byte files to see if they have changed.
  - If a change is detected the script rebundles a new DIP zip file after comparing against the current downloads
  - Generates a CRC against the zip file
  - Adds the CRC to the end of the filename
    `L:DownloadService\[SN]\[SN]spss_[32CRC]`
  - Moves the new DIP zip to the access directory
- **Study Release.** A separate script releases to the web server for inclusion in the catalogue (no data, no access restrictions)

- **'Red Folders' (Acquisitions Material).** A separate script runs end of day to validate and transfer files of scanned hard copies and incorporate into the AIP as these will not be finalised until after the data collection is released and checked by the depositor. Material added includes:
  - Correspondence
  - Deposit and data review forms
  - Licences
  - Miscellaneous items

3.6.2 Controls performed to guarantee the integrity and the authenticity of the digital resources

Five versions of the complete preservation system are held: main near-line copy (on the main preservation server) and a shadow copy (on main preservation server). Both are held on the main area on the Hierarchical Storage Management (HSM) system and are presently accessed only by the dedicated preservation user. The access online copy (on the mirror preservation server) is held in a RAID 5 disc system and copies are generated for user access and dissemination. There are also a near-site online copy kept on a RAID 5 disc system on a server located in another building within the University of Essex, and an off-site online copy. Finally a disc-based offline copy exists, which is held in either DVD-R or CD-R copy.

Per Data Collection MD5 Checksums are generated on each of the preservation servers and regenerated and compared as matching each time a data collection changes. Ongoing random comparison of checksums across the servers is made on files which have not changed. Checksums are maintained in a separate database to include Server, directory, path, size, created, modified, directory Y/N, checksum data

DIP generated for access have a checksum generated and inserted in the filename to maintain non-guessable file paths to resources and to enable recipient validation of downloaded resources.

To date no transformations of preservation versions of digital objects have been mandated and implemented at an individual or bulk level directly in the Archival Storage system. The restrictions on deposit formats and the nature of the collections mean that the majority of preservation files remain accessible and format migration to newer 'versions' of formats like SPSS and Stata are enacted by the ingest team under the same procedures as for a standard new or revised ingest.

It is expected that as deposit formats expand and greater volumes of audio-visual material are developed that planned migration to new file formats will be undertaken in bulk at the Archival Storage level.

3.7 ADEQUACY OF THE MODEL

Discussions during the development of the first part of this case study indicate that the selected events prioritised in the proposed model for managing authenticity and provenance through the digital lifecycle are appropriate however as some aspects of the proposed events are not undertaken at the UK Data Archive we can provided limited feedback on the event structures proposed. At the stage the Archive has only a limited contribution to make to evaluation of the pre-ingest phase as outlined below.

It has been noted during internal debate over the proposed model that it assumes and requires the existence of a discrete, event based model for all archival actions and the Archive fully supports the concept that this is the best approach. However, since the formal adoption of OAIS principles at the Archive the focus has been on building the foundations for a more granular system of event-based information capture.

One barrier to adoption has been the absence of clear best practice from the archival community and this proposed model is an important step towards that best practice. In the absence of such an agreed model the Archive has focussed on areas where best practice does exist such as formal information security (against ISO27001) and formal records management (against ISO15489).

At this stage of our strategic development the Archive has a strongly procedurally based process for handling the disparate systems developed for data management over 40 years. Implicit within this system is a trust-based approach to actions taken on DRs and a confidence-based assessment of our activities by the designated community.

As full digital resource lifecycle-based data standards are adopted locally this procedurally based system will influence the adoption and development of new data management and metadata capture systems throughout the ingest and archival storage processes but the Archive considers procedural integrity to be a necessary precursor to such development. An inevitable factor in the future implementation of a granular, event-based control mechanism will be the cost implications in financial and time-investment.

In the following subsections the adequacy of the model will be discussed by comparing the current practices held at the Archive with the proposals in the model on the controls to be performed and the authenticity evidence to be collected.

3.7.1 Pre-Ingest Phase

3.7.1.1 Activities in the pre-ingest phase

It is clear from the attached response that the UK Data Archive has little or no control over the activities in the keeping systems where DRs reside in the pre-ingest phase. The focus of the organisation has been on providing guidance to the designated community with regards to best practice on data management and sharing, but there are two clear issues which impact a comparison with the model:

1. The disparity of opaque keeping systems in which a DR resides in the pre-ingest phase and the vast array of methodologies involved with DR development imply that the best current approach is for the UK Data Archive to continue to provide high level guidance on procedures and the adoption of appropriate data and metadata standards.

2. The absence of agreed best practice methodologies in the digital preservation community on the implementation of authenticity and provenance management means that adequate supporting tools have yet to be developed to support consistent data management in this early part of the DR lifecycle

It is clear that many of the DRs submitted to the UK Data Archive have simply been 'retained' by the original Producer or their institution in preparation for submission with few formal data management provisions and where such provision exist they are not communicated as part of the SUBMIT process. As the UK Data Archive implements underlying data storage standards designed explicitly for full lifecycle management and as the breadth of the collection changes it will become more practical to consider pre-ingest keeping systems and even managed interfaces between keeping and archival systems.

Given the proliferation of data management models pre-ingest the only practical approach for the UK Data Archive is to provide clear guidance to keeping institutions only for the **SUBMIT** and, in the absence of any formal authority over Producers to continue to provide support on best practice including minimal standards for **CAPTURE, INTEGRATE, AGGREGATE, DELETE, MIGRATE** and **TRANSFER** events to be retained during the pre-ingest phase of the DR lifecycle.

3.7.1.2 SUBMIT

At the UK Data Archive the content of a SIP in terms of data scope and file formats are strictly managed via the Acquisitions Review Committee and the actions of the Acquisitions review team. The structure of the SIP in contrast is extremely varied and it is currently accepted that extensive ingest time is spent on normalising that structure to local Archive standards.

A clear submission agreement in terms of Licence Form and Deposit Form exists but at present no record at the Archive formally records keeping system DELETE events post-deposit.

Analysis of the current practices in the Archive with regard to the schema proposed in D24.1 leads to the following remarks.

Agents

- *Owner agent* as designated in the model is analogous to the Rights Holder details contained in the Archive Acquisitions negotiation
- *Keeping System Administrator* is analogous to Depositor in that they are responsible for performing the submission of the DR

Input

- Analogous to the SIP for the UK Data Archive though as outlined in the case study above the close relationship between Owner and Archive extends into the Ingest process and the custody boundary is not entirely fixed

Output

- *DR' (i.e. the version of the DR which is kept in the keeping system)*. The Archive does not currently involve itself with the new version of the DR retained in the keeping system
- *DR''(i.e. the version of the DR which is kept in the LTDP system)*. The DR as SIP is analogous to the UK Data Archive Study

Authenticity Evidence Record

Though the model notes that "two different and independent systems are involved in the submission, the keeping system and the LTDP system. The corresponding authenticity evidence record must therefore contain the evidence produced, and conveniently authenticated, by the administrators of both systems" in the case of the UK Data Archive only the artefacts of the Archive-controlled Acquisition Negotiation are known and managed

- *Keeping system*

As noted above the Archive cannot comment on the AER of the keeping system but can confirm that a formal notification that the "DR has been received and accepted by the LTDP system" is provided to the Owner and Administrator of the keeping system.

- *LTDP system*

– *Event type*: submit-in
– *Identification data of the keeping system*. Recorded as a text string or in the case of known academic grants is recorded from a trusted system.
– *Identification data of the LTDP system*. At the point of submit-in an Acquisitions Identifier exists to control the entire negotiation and on submit a data collection identifier is assigned to the SIP. There is usually but not always a one to one relationship between these identifiers
– *Date and time the DR has been received from the origin system*. This information is available from one of the component systems (e.g. University of Essex ZendTo Service) if the transfer is electronic rather than physical but this information is recorded manually in the local system as a custody handover date.
– *Identification and authentication data of the LTDP system administrator*. The current Acquisition software does not uniquely identify the logged on individual for negotiation actions or assign a named individual. Specified managers exist for Government and Academic data and there responsibility is assumed but not specifically recorded with the PDI.
– *Assessment by the LTDP system administrator on the delivery of the DR by the keeping system and on the subsequent controls*. ARC records record reviews of sample data and decisions related to offers of deposit but these are currently managed records rather than formal contents of the PDI. All relevant negotiations regarding assessment of the actual DR delivered as formally preserved as correspondence records. Controls undertaken are governed by procedures and are listed under section 3.4.
– *Identification and authentication of the keeping system*. A managed database of individuals (Producers, Rights Holders, Depositors, and Authors) of associated material) exists and is linked to a controlled vocabulary of institutions. All identification details are retained in the Acquisitions system and preserved with the licence agreement and DDI2 XML catalogue record for the DR.
– *Trustworthiness of the data channel used for the transfer*. Adoption of best practice for digital transfer is limited outside of Secure Data Service data. Best practice methods for transfer are defined and transfer methods are agreed via correspondence but no explicit record of the transfer method or evaluation of its' trustworthiness is retained in metadata.
– *Integrity check performed on the digest produced by the keeping system administrator*. Adoption of best practice for digital transfer is limited outside of Secure Data Service data where fixity checks are required and validation procedures in place.
– *Digest of the of the DR authenticated by the LTDP system administrator*. No digest on receipt is mandated in current procedures. The reasons for this relate to the porous status of the SIP on arrival. As outlined in the case study extensive communications between the Owner and the Ingest team may continue throughout the ingest process and this may (in the case of complex anonymisation for instance) even impact the structure of the final AIP and access restrictions to the DIP. As this is such as vital part of the service offered by the UKDA investigations are

underway consider how future 'micro-submissions' of SIPS can occur with minimal impact on administrative overhead but any implementation is dependent on the DDI3 lifecycle metadata standard.

3.7.2 LTDP Phase

3.7.2.1 INGEST

Referring to the model, **INGEST** is currently the only event that may be clearly identified. Comparing the current practices in the Archive with the schema proposed in D24.1 we may make the following remarks about the authenticity evidence gathered and retained during the ingestion, and more specifically the content of the Authenticity Evidence Record, as proposed for the INGEST.

Authenticity Evidence Record

- *Original identifier of the submitted DR.* As above, retained.
- *New identifier of the DR in the LTDP system (if given.* Identifiers associated with Acquisition, and the SIP. Separate DOI registered for DIP on release
- *Date and time the DR has been accepted by the LTDP system.* As above, retained
- *Date and time the ingestion has been completed.* Analogous to release date time. Retained in catalogue metadata and stored within DDI2 XML
- *Identification and authentication data of the LTDP system administrator*. Administrator is identified by name for each event. This is not currently integrated with (Active Directory) authentication.
- *Assessment by the LTDP system administrator on the ingestion of the DR and the subsequent controls*. Assessment on format migrations, if any, including a statement that the intellectual content of the DR has not changed, specifying the criteria that have been adopted to perform the assessment. As noted above there is an extensive processing plan and guidance on actions implied for specific file types and anomalies encountered but these are not recorded by the system in an 'event' based method in line with the proposed model. Neither is an explicit statement that the intellectual content has not changed provided. Processes undertaken including the documentation of any issues from a format conversion are retained and made available to consumers.
- *Assessment on the authenticity and provenance evidence contained in the submitted DR*. With the exception of Acquisitions process metadata collected in the earlier phase all supporting materials are reviewed during ingest and may be extensively revised with the cooperation and agreement of the DR Owner to improve the understandability of the DR to the consumer. These amendments are approved by the Owner on release but the degree of potential makes it potentially complex to implement a granular event-based model for each change undertaken.
- *Digest of the AIP produced by the ingestion process.* A digest of the AIP is generated at the completion of the ingest process and managed in line with procedures outlined under section 3.6 above.

3.7.2.2 Other events in the LTDP phase

LTDP-AGGREGATE (one or several DRs stored in different AIPs, are aggregated in a single AIC)
As discussed above aggregation of AIP is not undertaken at the UK Data Archive. An AIP is entirely fixed during the preservation phase and the only possible amendment to the package involves a new ingest process and a new AIP. Some AIP are presented together in the Resource Discovery system but this aggregation is purely at the metadata level with the AIP defined as related object.

LTDP-EXTRACT (one or several DRs which are extracted from an AIC to form an individual AIP)

As discussed above extraction of AIP is not undertaken at the UK Data Archive. An AIP is entirely fixed during the preservation phase and the only possible amendment to the package involves a new ingest process and a new AIP.

LTDP-MIGRATE (one or several components of a DR are converted to a new format)
As discussed above while the maintainers of the Archival Storage system undertake a technology watch to identify obsolete formats this has yet to be an issue at a bulk archival storage level. Format migration of AIP content has been restricted to updated to DIP content in response to demand for newer formats (primarily for qualitative data applications). The significant properties of most of the SIP formats and the complexities of ensuring validity post transformation means that such migrations are undertaken by the ingest team and follow all the standard ingest procedures outlined above.

As the Archive is required to ingest a wider variety of less stable formats (especially for audio/visual data) we would expect to implement bulk transformation at the Archival Storage level.

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)

Though deletion is a vital component of any model the UK Data Archive has yet to reach the point that a stated preservation time expires. . Like The UK National Archives, the Archive distinguishes between 'soft deletion' whereby certain references to the withdrawn content are deleted, but not the content itself, and 'hard deletion' whereby the content and all references to it are deleted. In the case of soft deletion the data collection is only accessible to Digital Preservation and Systems staff. The Archive chooses soft deletion as the default method of withdrawal since it is too expensive to remove data collections, and their physical removal would present unacceptable risks to other parts of the collection. Soft deletion is handled through a 'de-accession' process as described in section 3.6, this involves a managed request from the ingest team to the Archival Storage system but involves the movement of the AIP, not its deletion.

LTDP-TRANSFER (a DR stored in a LTDP system is transferred to another LTDP system)

As noted in the case study above the UK Data Archive has yet to commit to the transfer of AIP to another LTDP system and in the case of many legacy datasets is not permitted to do so as a condition of the deposit licence. As authenticity and provenance information is united into agreed data standards from the various aspects of the ingest and archival storage workflows it is expected that we will develop a standard for Archival interchange, even without a clear remit to do so preparations will form part of our Business Continuity and Disaster Recovery plans in line with our risk register against ISO27001 for information security.

3.8 PLANNING FOR FUTURE PRACTICE

3.8.1 Implications of the Analysis

The UK Data Archive is fully supportive of the model proposed by APARSEN for the prioritisation of authenticity and provenance evidence capture during the digital object lifecycle. The items provided are the key high level events and elements of metadata required for collation into effective PDI. Our only caveat is that any best practice must explicitly include a mechanism for the capture and transfer of appropriate associated rights through the lifecycle. (See section 3.9.2.1 below in particular)

The future plans for the archive take account of this model and indeed propose to go further in the granular capture of events in the medium to long term but the UK Data Archive Case Study also raises a number of issues related to the deployment of these best practices into a production environment. As with a number of other Work Packages it is clear that the high level academic models of best practice are useful for general guidance but these remain somewhat opaque on the technical side (in much the same way that the OAIS aims to be a reference model rather than a technical architecture for a technical implementation). The APARSEN projects differs from other EC funded projects in the area

of digital preservation/continued access in that its goal is to create a network of excellence and, as with all best practice advice, there is a need to provide a path for those implementing digital repositories from the current state of the art in terms of production tools to the current state of the art in academic terms.

It has been acknowledged in [1] that current practice is some way from academic theory in this area and reaching the final goals will need to take into account the significant technical and financial challenges faced by the deployment 'in the field' of any standards. It has further been acknowledged that the areas of authenticity and provenance lack clear consistent best practice guidance and while this work package delivers some core priorities these logical requirements will take some time to reach maturity in technical implementation and tools available to the LTDP community. Further, these core priorities depend on some key foundations that fall outside the direct remit of this work package. These foundations are inherent in the current drive towards 'Trusted Repositories Standards (including the Data Seal of Approval, ISO16363 and DIN31644), in the inclusion of information security and governance requirements in those standards, and in the noticeable on-going alignment between the professional spheres of long term digital preservation and records management (in line with ISO15489).

The UK Data Archive is a founding board member of the DSA and has achieved the Seal of Approval. This process aligned with the on-going actions within the UK Data Archive to further formalise governance and records management procedures. Concurrently the Archive achieved ISO27001 (Information Security) status during the development of its Secure Data Service and under the APARSEN project has undertaken a self-assessment and test audit against the (at the time draft) ISO16363. On-going action at the Archive in all these areas is considered a necessary prelude to the kind of granular metadata capture proposed by this work package.

The UKDA works within a varied funding model with multiple sources and multiple individual projects which present challenges to maintaining a consistent technical solution across its many services. This is further complicated by the need to maintain a hybrid development environment due to its position with the University of Essex (a Microsoft Dot Net environment) and the predilection of the wider archival and academic community for open source solutions. The UKDA's specialist areas of Social Science and Humanities benefits from a committed community who have developed the Data Documentation Initiative (DDI) data standard and all future plans outlined must be understood in terms of our migration to the latest version of this standard DDI-L (Lifecycle) which permits the managed re-use of metadata and resources in a distributed manner as well as offering strong support for the full lifecycle of the digital object. These funding and technical complexities are by no means unique to the UK Data Archive and we hope this makes us a useful case study for evaluation. The long-tail of legacy collections managed by the UKDA, the local technical environment and the need to implement a nascent standard means their immediate development priority is to develop a bespoke ingest and rights-managed access system, it is expected that other data management tools will be 'plugged in' to this core system over time.

It is the policy of the UKDA to continue its on-going process of formalising governance and records management as a critical pre-requisite to delivering, granular, reliable and secure data and metadata capture and management throughout the digital object lifecycle, including those related to authenticity and provenance.

The UKDA will pursue a policy of maintaining the high levels of confidence from across its designated community including depositors (the contact for negotiating a deposit agreement is not usually the original Producer) and users. Until best practice guidance in the areas of authenticity and provenance have moved from the academic to the practical such that they are effectively deployed within archival tools that do not vastly increase the time/cost of processing we will pursue a policy of maintaining a trust-based system with close relationships with depositors and highly trained ingest staff with a direct understanding of the needs of the designated community. Each time a decision is taken to ask ingest team members to 'manually' record an archival process event (in the absence of

open source or commercially available systems to automatically capture the event) this has a direct impact on the time to ingest a data collection and a related financial impact.

The series of interviews and on-going feedback from our designated community outlined in section 3.8.3.8.2 suggests a different prioritisation of supporting evidence than that of the LTDP community. Until the understanding of the concepts of authenticity and provenances (in LTDP rather than general terms) have driven demand for increased presentation of such evidence alongside every DIP it is difficult for us to justify the extra cost in transforming the evidence within our systems into evidence suitable for their consumption. In the medium term the Archive will maintain a procedurally-based evidence system which focusses on the transparent sharing of our standard procedures for data management rather than on providing end user access to detailed actions taken on a per-data collection or per-file basis.

As clarified by the future plans outlined in this case study none of the above precludes the Archive's support of the lifecycle event information capture proposed by this work package, it simply attempts to clarify the position of that information capture within the practicalities of general Archival and UKDA-specific deployment of technical solutions in a production environment.

3.8.2 Needs of the Designated Community

The UKDA is in continuous close contact with its designated community (both producers and users) but involvement in this work package raised a number of specific questions. With this in mind we arranged a series of short interviews with some of the expert users on our advisory committee to:

1. evaluate their interpretation of the concepts of authenticity and provenance;
2. consider what evidence of authenticity and provenance needs to be presented to end users (in addition to its management within the repository).

UKDA interviewed both users and producers from within the social science domain. It was agreed that interviews would not be named or directly quoted in detail. Interviews were an open expression of opinion based on a brief outline of the work package goals provided. They do not claim that the interviews represent controlled evidence appropriate for statistical analysis however they provided further support for the UKDA's staged approach. The broad outcomes of the interviews are provided here as they provide some important context for the Archive's response to this deliverable.

There were a variety of interpretations of authenticity and provenance; as with other discipline specific repositories we do not expect our user community to be either familiar or even comfortable with OAIS terminology, but there were a number of consistent themes in response.

Both producers and users have a strong confidence (to avoid the formal term 'trust') relationship with the UKDA. There is general understanding that there is less 'control' over data pre-ingest but that the UKDA is seen as the managing curator who will standardise what is available. Data Producers are aware of and unhappy with the current common scenario where data sets are circulated (and incrementally aggregated, re-weighted, labelled, revised etc) among researchers with no formal control before eventually being used as the basis for secondary research and publication. The Producers consider increased direct use of the 'authentic' LTDP instance for research as the immediate priority. Increased mandated use of data citations is expected to improve this situation while also providing increased credit to the data producer and ease of tracing the DR through the lifecycle.

Authenticity is 'assumed' from the point in the lifecycle where the UKDA has custody. There is an understanding that exerting control earlier in the lifecycle would be desirable but this level of consistent administration is difficult to standardise and enforce and would be costly and would limit the time available for preparing more research-relevant contextual information. Provenance in terms of custody history and transformation history was considered a lower initial priority by both creators and users. The LTDP descriptive/resource discovery metadata always includes the funding body and researcher details and this was considered sufficient.

The variety of data creation environments, funding environments and keeping systems were all raised as issues with the complex pre-ingest phase, especially given the absence of standard tools. From the user perspective the priority when accessing data is the availability of sufficient Context information. Given the large quantity of information already supplied in a DIP anything which does not have immediate relevance to the use, interpretation and analysis of the data is given a low priority. Information which would allow the user to make some ad hoc personal confirmation of accuracy is relevant but not a high priority. There was little appetite from the user end to receive additional Authenticity and Provenance data captured by the UKDA as this would further complicate the interpretation of the download. Simple clarification as to whether an earlier analysis is replicable on the latest version of the data is available is considered important but this should be short and clear i.e. manually edited into useful prose from more granular technical change history information.

A number of interviewees mentioned an increasingly prominent issue for the UK Data Archives: the increased availability of and demand for administrative data from government departments and other official sources. It is expected that there will necessarily be a greater focus on Authenticity and Provenance when increased quantities of administrative data (not developed with a research focus) are collected and archived. This will have significant cost implications but still the priority among respondents was for sufficient context about the business processes used in generating the administrative data, not on formal authenticity and provenance evidence in OAIS terms.

We feel that while these responses might be mistakenly understood to reduce the importance of the OAIS concepts or the validity of the model proposed by WP24 but in a limited funding environment these reactions support our current approach of a firm governance foundation to support the management of future change. Granular authenticity and provenance collection throughout the process is not only a function of academic agreement on the underlying concepts; it is also a function of providing low cost, low effort tools to support the seamless collection and transmission/submission of the data. In many ways this must be driven by proofs of concept and production line application within LTDP systems first and migrated into other areas of the lifecycle. Even once this information is formally collected by an Archive there is an additional cost to making is truly ‘useable’ by the designated community so (again in a funding-restricted environment) we must be demand-led in considering what information to add to an already complex DIP.

3.9 APPLYING THE MODEL TO FUTURE PLANS

3.9.1 Pre-Ingest Phase

If we define ‘keeping system 1’ as the last keeping system prior to the LTDP phase (the Depositor/Rights Holder the Archive interacts with directly) it may be useful to consider the different characteristics of the prior keeping systems (keeping systems 2-N) in terms of the level of influence and control the destination LTDP system has in the relationship.

Keeping System 2 to Keeping System N

Without direct contact the UKDA has no *control* on the practices of these systems in procedural terms. They have some *influence* on authors and keeping system administrators in terms of the guidance we provide on “Research Data Management and Sharing”[1] and in cases where the original author may have retrieved data from the Archive they are implicitly influenced by the structure and content of the data they receive.

Keeping system 1 has a closer relationship with keeping systems 2-N. It is expected that the keeping systems 2-N will only advance significantly once we have good practice in Keeping system(s) 1.

[1] The term Research Data Management is increasingly used to cover activities carried out by the Producer before data reaches the repository. It should not be confused with Data Management in OAIS terminology.

Keeping System 1

The people we deal with directly at the Producer Archive Interface in negotiating a submission agreement.

The number of types of system here are clearly not as wide-ranging as exist among keeping systems 2-N but they remain extremely varied. Due to long-standing relationships with many of the funding bodies and government departments involved the Archive has some influence over interactions with these keeping systems but certainly no power to refuse deposits unless there are critical issues with data such disclosure risks, lack of consent or rights issues.

3.9.1.1 Events prior to SUBMIT

As noted under the analysis of current practice the UKDA has no current requirements for the detailed capture of events prior to **SUBMIT** and no power to enforce their capture or transmission. To our knowledge for most of the data collections we accession these processes are not specifically controlled during this part of the DR Lifecycle.

Increased adoption of critical authenticity and provenance metadata collection by the keeping systems is desirable but not practical unless it is aligned with agreed best practice advice and the adoption of appropriate standards, without alignment to support standard automated ingest the additional metadata is potentially an extensive additional burden on the Archives ingest processes. The Archive would expect that such adoption of best practice and standards will only arise when compliant tools to support these keeping system processes reach critical mass.

The DDI-L standard does provide for the full lifecycle model and the UKDA would expect to ingest metadata including authenticity and provenance related material once the model is integrated into software tools that support general data management throughout the lifecycle. The demand for such tools is expected to increase as international funding bodies require data management plans alongside initial applications for funding.

The events in the pre-ingest phase which are likely to receive priority attention align closely with those from the DDI-L controlled vocabulary (See http://www.ddialliance.org/Specification/DDI-CV/LifecycleEventType_1.0.0.html). With a DDI-L system the UKDA will be able to capture and ingest pre-ingest events including:

- Study Proposal
- Funding
- Study Design
- Instrument Design
- Questionnaire Translation
- Questionnaire Adaptation
- Interviewer Training
- Ethics Review
- Legal Review
- Sampling
- Instrument Pre-testing
- Pilot Study
- Data Collection
- Data Collection Reports
- Data Processing
- Data Processing Coding
- Data Processing Classification

- Data Processing Interview Transcriptions
- Data Processing Weighting
- Data Processing Aggregation
- Data Processing Composite Measures
- Data Processing Derivation
- Data Processing Data Quality Checks
- Data Processing Data Integration
- Data Processing Disclosure Limitation
- Data Processing Imputation
- Metadata Production
- Metadata Editing
- Metadata Translation
- Final Report
- Evaluation

Note that the UKDA works closely with depositors to enrich and improve submitted data collections so many of these events overlap with events which occur during the ingest process. We are far more likely to persuade depositors to provide these context information details than, for instance a full custody history or a detailed history of transformations with inputs/outputs/associated checksums etc.

3.9.1.2 SUBMIT

As noted elsewhere all medium and long term amendments to the UKDA's systems are dependent on a deployment of the DDI-L (Lifecycle) standard.

Given the response from our designated community (section 3.8.3.8.2) and the absence of tools to streamline their capture it seems unlikely that detailed Authenticity and Provenance information in the pure LTDP sense will be available from most keeping systems for some time.

As noted elsewhere the UKDA is not in a position to comment on any Submit record maintained by Keeping systems but formal, standard acknowledgments of receipt and ingest approval are transmitted to the keeping system as part of the standard submit process.

D24-1 Notes the following aspects of the Submit process:

- "Event type: submit-in
- Identification data of the keeping system
- Identification data of the LTDP system
- Date and time the DR has been received from the origin system
- Identification and authentication data of the LTDP system administrator (the person who was in charge of the destination keeping system when the transfer took place)
- Assessment by the LTDP system administrator on the delivery of the DR by the keeping system and on the subsequent controls:
  - Identification and authentication of the keeping system
  - Trustworthiness of the data channel used for the transfer
  - Integrity check performed on the digest produced by the keeping system administrator
- Digest of the of the DR authenticated by the LTDP system administrator"

All of these form part of existing data acquisition processes or future DDI-L requirements. The UKDA's submit-management is already more granular than the proposed model (as it includes sample data evaluation pre-submit and a number of formal communication acknowledgements) but on-going

developments will formalise the current process into a standard event model in line with PREMIS with metadata which is capable of structured data exchange (initially in XML).

Current agent identification for external agents aligns with the models proposals of 'Owner' and 'Keeping Systems Administrator' but extends into more granular identification of funding bodies, primary investigators, contributors and rights holders. These will be extended into a formal, extensible controlled vocabulary. Current internal agent identification relies on self-reporting, i.e. any authorised LTDP system administrator may log in to the software and they self-identify actions they are associated with. Future software systems will be designed from the ground up to record which staff (local agents) have amended a Submit or Ingest record using information available via the Microsoft Active Directory authentication system and internally developed authorisation systems.

Plans are underway for a revision of our unified management of 'people and places' which will encompass external agent management of individuals and organisations. The system will include provision for persistent author identifiers and organisation identifiers. These identifiers will support capture of consistent information from Keeping System 1 but are unlikely to extend further back in the DR Lifecycle in the medium term.

In the short term 'deposit forms' which accompany all submission events will be updated to comply with DDI-L and with our current strategy of standardising the transfer of consistent rights information relating to copyright, consent and the rights to preserve and provide access. A *Rights and Access Management Strategy* is currently being implemented and will guide the development of integrated systems to receive permissions from depositors and grant them to end users with maximum central control.

In the medium term deposit forms will be extended to automate fixity checks and the development of detailed (but optional) forms for recording free text metadata on a per file or a per directory basis; this will be driven by a METS (Metadata Encoding and Transmission Standard) structural map derived from the directory tree of the submitted files (see also 3.9.2.1). This level of structural metadata extraction accompanied by file format recognition (initially on Ingest, not Submit) will support more granular AIP definitions and machine-validation of the submission agreement.

Fully functional deposit forms will in the longer term be offered as downloadable applications. In this way it is hoped they will also be adopted by the 2-N keepings systems. We would expect that developing such a product and including seamless application of checksums is the only way the UKDA can obtain consistent fixity checks on SIPs.

One challenge is the transfer of emails exchanged in the course of negotiation into a preservation environment such that they can be searched during future interactions rather than simply maintained. Where possible standard information exchanges (e.g. confirmation of receipt letters/emails for SIP) will be automatically generated by the system to avoid the manual creation of event records.

The Acquisitions team which handles deposit negotiations will have access to a subset of the Checklist application described under section 3.9.2.1, standard events will include all of the items listed above and will be extended to confirm receipt of all required supporting metadata and the formal handover of required rights.

Other events in the acquisition process including application for deposit, evaluation by the Acquisitions Review Committee, reviews of sample data received etc. will also be recorded as pre-ingest/pre-submit events.

3.9.2 LTDP Phase

3.9.2.1 LTDP-INGEST

LTDP-INGEST: a DR delivered from a producer is ingested by the LTDP system and stored as an AIP.

D24-1 Notes the following aspects of the Ingest process:

- Event type: ingest
- Original identifier of the submitted DR
- New identifier of the DR in the LTDP system, if given
- Date and time the DR has been accepted by the LTDP system
- Date and time the ingestion has been completed
- Identification data of the LTDP system
- Identification and authentication data of the LTDP system administrator
- Assessment by the LTDP system administrator on the ingestion of the DR and the subsequent controls:
  - Assessment on format migrations, if any, including a statement that the intellectual content of the DR has not changed, specifying the criteria that have been adopted to perform the assessment
  - Assessment on the authenticity and provenance evidence contained in the submitted DR;
- Digest of the of the AIP produced by the ingestion process

All of these form part of existing data acquisition processes or future DDI-L requirements. In line with the 'main actions' outlined under 24-1 ingest involves:

- the application of a locally unique identifier (a DOI is also created but not until release of the DIP)
- A full assessment of the SIP in the form of a 'processing plan'
- Format conversions to approved LTDP formats in line with our published standards and procedures

Format conversions and outcomes are currently recorded in prose (see Read and Notes files details in 'Current Practice') rather than as structured data associated with a particular file in the AIP.

Agents represented include the head of the ingest team and the various members as well as the resource discovery team responsible for updating 'catalogue' information in Data Management.

For actions within the ingest system logs of events (commits and amendments to the metadata surrounding an object) will be recorded wherever possible using agent information available via the Microsoft Active Directory authentication system and internally develop authorisation systems.

One of the key issues for the UK Data Archive is a result of the close relationship with producers/depositors and the long standing practice of working with them during acquisition negotiation and ingest to correct and improve submitted data as well as to enrich deposited supporting metadata and documentation and to create the standard Archive metadata common to all studies. The result of this situation is a Producer/Archive interface which is somewhat porous and some or all of a SIP may be amended and resubmitted by the depositor or may be amended by the Archive with the approval of the depositor.

The challenge is to improve granular control over submit events without risking the level of service to depositors due to unacceptable administrative burdens.

Though the implementation of a DDI-L ingest system will improve support of wider lifecycle events we will not, at least in the medium term, have an ingest applications which 'calls' (and can therefore log) every external validation and transformation script run on standard file types. In addition there will always conversions and validations which may only be run through commercial applications (such as SPSS) which cannot be integrated for control and automated logging. Therefore there will be an additional administrative burden in manually recording each step taken on a data collection or one of its component files.

Given the impracticality of fully machine-mediated event deployment and management during the ingest process some sort of compromise is required. The Archive is currently developing plans for

snapshot comparisons (SIP vs AIP and AIP vs DIP) with associated file linking and metadata collection. As functionality is implemented we will consider deploying similar functionality to depositors to support improved data management earlier in the DR Lifecycle. The process will be applied alongside more integrated use of the DROID/PRONOM file format identification system.

An initial snapshot of an SIP—on the assumption that no similar deposit tool exists to provide this pre-ingest—would involve:

- Locking the SIP
- Recursive fixity checks
- Full directory tree/directory name/ file name extraction
- Generation of a full manifest
- Generation of a METS-compliant structural map of the full SIP

On completion the ingest system provides appropriate user entry forms for metadata including standard required metadata and the option to annotate the structural map.

Note that this functionality, if available to Keeping systems would exert more fine-grained controls and greatly simplify the deposit and ingest processes.

On completion of the AIP for Archival storage a second snapshot would be taken. The system makes a 'best guess' at relationships between files in the SIP and those in the AIP. The Ingest team confirm or correct the relationships between files to the point that we can identify:

- All new files
- All unchanged files (fixity check comparison)
- All changed files

This information is sufficient to generate a bespoke 'checklist' to record actions taken on the SIP to generate the AIP with associated checkboxes to confirm that appropriate procedures (including quality assurance) have been followed.

- All unchanged files require no action
- New files should have been generated in line with standard procedures
- All changed files must be identified in terms of:
  - Merge or split from originals in the SIP
  - File name changes
  - File content changes
  - File format Changes

Amendment timestamps on AIP files permit some limited automated time logging of event sequence.
Accompanying this dynamically generated checklist a standard 'worklog' event checklist would allow formal recording of standard ingest processes.
The increased structure and control outlined above is a carefully considered medium term goal to permit the output of standard compliant (DDI-L, PREMIS) event records without overburdening the ingest team with administrative processes.
A number of the contextual artefacts produced by the Archive to support data collections (see details on read and notes files in the Analysis of Current Practice) can be machine-generated as a result of the process above.
The analysis of current practice details the numerous steps taken during current ingest procedures, the following items are standard DDI-L lifecycle processes that could occur during ingest as well as during the pre-ingest phase:

- Data Processing
- Data Processing Coding

- Data Processing Classification
- Data Processing Interview Transcriptions
- Data Processing Weighting
- Data Processing Aggregation
- Data Processing Composite Measures
- Data Processing Derivation
- Data Processing Data Quality Checks
- Data Processing Data Integration
- Data Processing Disclosure Limitation
- Data Processing Imputation
- Metadata Production
- Metadata Editing
- Metadata Translation

In addition to the above all of the following items will be output to a DDI-L record during ingest

- Original Release
- Deposit
- Preservation Package Production
- Dissemination Package Production
- Data Analysis Reports
- New Version Release

3.9.2.2 LTDP-MIGRATE

As noted in the Analysis of Current practice the Archive's current migration process occurs through a re-iteration of the ingest process rather than through bulk migration of files within the Archival Storage system.

As the variety, size and number of files and formats increases over time the Archive expects to integrate automated file format migrations into its workflows over time but this will careful consideration. At present the Archival Storage is kept isolated from any change process other than that mandated under standard requests from Ingest to Receive Data. It is expected that bulk transformations would require a separately managed workflow to retain this integrity.

D24-1 Notes the following aspects of the Submit process:

- List of all the components of the DR affected by the migration, specifying for each of these:
  - the reason why the migration has been performed;
  - the input format;
  - the output format;
  - the procedure and the application used to perform the conversion;
  - criteria that have been used to verify the result of the conversion, e.g. the Transformational Information Properties which in the judgment of the administrator are adequately carried over into the new format.
- Statement, for each DR affected by the migration, that the intellectual content of the DR has not changed, specifying also the criteria that have been adopted to perform the assessment

The reason for each migration is recorded unless it forms part of standard procedures. Input and output formats follow standard procedural workflows and similarly applications used are recorded within

procedures. All procedures are formally managed meaning that the Archive can always identify which standard procedures were in place during a given ingest (in this case re-ingest to migrate) event. Formal per-format migration records will be investigated as part of the workflow analysis for the new worklog system.

D24-1 mentions Transformation Information Properties which are particular to the latest version of the OAIS standard and are conceptually related to Significant Properties. There is a general understanding within the community that the definitions of "Significant Properties" vary widely and the Transformation Information Property has yet to gain common currency in production environments. All evaluations of transformations at the Archive involve detailed procedures for human-mediated and machine-checked (where possible) validation but these are not benchmarked against a set list of transformational properties. It is expected that, like file format identification, the resource intensive nature of managing such properties across the vast range of file formats means it will not be practical in most production environments until some form of web services-enabled registry is available from a trusted provider.

3.9.2.3 LTDP-DELETE

Delete requests from Ingest to Archival Storage on Receive Data events (as described under the Current Practice part of the case study) does not actually involve a formal deletion from archival storage, only a stage in updating the 'current' AIP used to generate the DIP. As previously outlined the Archive favours soft-deletion over hard deletion (removal the artefact from access, not from archival storage). In the extremely unlikely event that a critical delete were required (e.g. were we to accidentally ingest data with an information classification level of 'Classified') this would be undertaken through the standard change management procedures for the archival storage system which apply to other areas such as media refreshment. These procedures are ISO27001 Information Security compliant.

3.9.2.4 LTDP-TRANSFER

As noted under Current Practice there is no current process for transferring data to other LTDP systems. However the implementation of DDI-L and the use of METS for structural metadata and PREMIS for event management will mean that we can generate self-describing Repository Exchange Packages (http://wiki.fcla.edu/TIPR/21). It is likely that RXP will be generated to further support PDI in the first instance but these could support transfer if necessary.

4 SCIENTIFIC EXPERIMENTAL DATA REPOSITORY IN HEP

4.1 THE DESIGNATED COMMUNITY

4.1.1 The data producing community

High-Energy Physics (HEP) is a global research community, known to the wider public through big research labs and experiments. Especially the hunting for the Higgs boson fascinates the public and media regularly report about successes and future steps in hunting for the Higgs.

It is a rather small research community of about 30,000 to 50,000 researchers, one fraction being "experimental physicists" working in big collaborations to experiment, the second fraction studying physics in theory. This division is important to highlight in the eyes of research workflows, research data production, preservation, distribution and reuse [2, 5].

Within experimental physics, research is done in big research facilities allowing to study in/on big experiments such as ATLAS[2] or CMS[3] on the LHC[4] at CERN[5]. Similar to the outstanding size and complexity of these experiments, the research data output is also extraordinary in its size and complexity. The data output from the LHC experiments is organized and handled via the Worldwide LHC Computing Grid (WLCG) and thus distributed worldwide in so called "Tier" centres[6]. Many parts of the workflows are automated, but monitored by researchers as well. Due to automation, the size and complexity of the experiments and collaborations, data authenticity and integrity are of highest importance to the research community when using and reusing the data. This becomes even more relevant when considering the multiple processing steps in the data lifecycle in HEP. In the past, it was apparent that the community does care a lot about their research data, and its current availability worldwide, so that solutions like the grid have been developed. But research data preservation as an independent activity had not yet taken off in the community. One reason to this is the ever growing data flow in which newer data superseded the old constantly. Another one is certainly due to a marginal tradition of reusing research data in HEP and thus only little awareness in that regard. Thus, any kind of preservation action focused on solely technical aspects.

4.1.2 Data preservation in HEP

The immense and complex data flow in HEP should not mislead over the important fact that many of the studied phenomena are unique as well as are the instruments they are measure with/at. That means that many of the experiments in HEP cannot be repeated, may it be due to the complexity/size or funding of the big experimental facilities.

This is one of the reasons why the community initiative "Data Preservation in High-Energy Physics"[7] (DPHEP) has been started. They studied the complexity and diversity of the research data output in HEP. This is in particular important as there are many community standards and individual solutions within experiments. They distinguished four different levels of research data in HEP (Table 1 after South, 2011 [6]). The most complex layer includes "basic level data" including simulation/analysis software in order to maintain full potential of the experimental data. The "simplest" layer of research data in this model comprises research data associated to a publication.

Preservation Model	Use case

[2] http://atlas.ch/
[3] http://cms.web.cern.ch/
[4] http://lhc.web.cern.ch/lhc/
[5] http://www.cern.ch
[6] http://lcg.web.cern.ch/LCG/public/
[7] https://www.dphep.org/

1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

Table 1 - The four different levels of research data in HEP listed in order of increasing complexity after (South, 2011)

It is important to remark that research data in HEP is not only produced in big collaborations, but also in smaller collaborations and experiments, as well as in theoretical physics. Some of these data are actually *in silico* simulation of physical processes which have to be preserved or exchanged to be compared to future or present experimental results. Depending on the setup of the experiment and/or the research question complexity and size of the datasets can vary tremendously.

4.1.3 The data (re)users

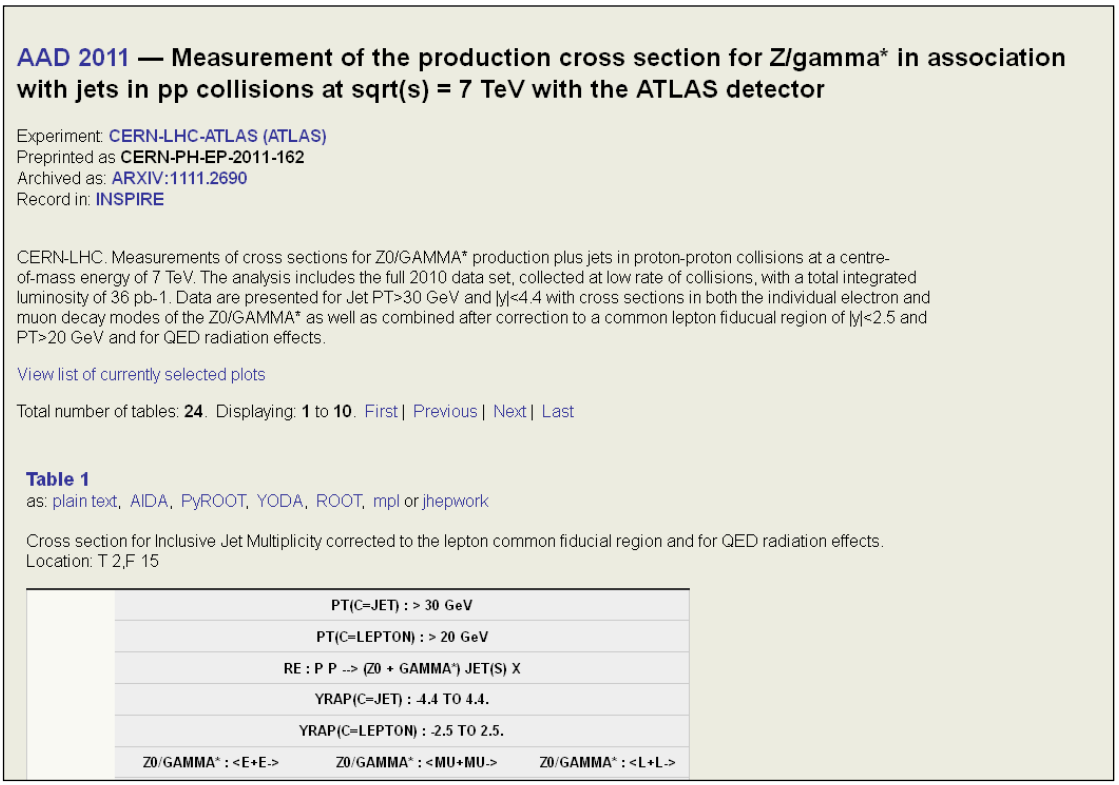
Consumers of shared data and preserved data, to date, vastly overlap with those who produced it. However, implementation of a solid strategy for data preservation, appropriate tools and a culture of sharing, would allow a growth in the community, as those who are not part of the original research that has been conducted could use the data, too. As an example: they could be theoretical physicists looking for experimental data for data-simulation comparisons. But they are mainly part of the HEP community and familiar with the community standards. Data produced within the big collaborations could be (re)used by different groups within one collaboration, sometimes at a different time. It is of course also possible that data would be "consumed" on a cross-experimental scale. It has to be noted though, that data reuse is an emerging activity in HEP and this case study will deal with the first example of making HEP data available for others.

The High-Energy Physics community is a highly dynamic and international community. This means that researchers move regularly and change positions frequently, e.g. in order to do experimental work at one of the main laboratories. With such a fluctuation in personnel it is even more important that data production is well documented and preserved. In that regard, authenticity and provenance has a crucial role in order to facilitate reliable future reuse of materials.

4.2 TYPOLOGY OF THE PRESERVED DATA

This case study will focus on a data repository making available the simplest layer of research data. It is the only well established source for open data across experiments and laboratories in HEP. Since the 1970s, it is hosted at a University, not at a particular laboratory. There, researchers started to collect data as presented in publication but encoded in an electronic format so that they could easily re-use it, at later points in time, to compare through numerical algorithms to their theoretical studies. Following this, it slowly emerged as a central facility for handling of high-level data in the discipline. By the end of 2011 it comprises 7210 records (only considering data associated to publications, see Fig. 4.1). In regard to the DPHEP classification scheme, only the uppermost layers are covered in this data

repository, 1 but increasingly some bordering between 1 and 2. The more complex data levels are covered by experiment specific solutions and databases.



AAD 2011 — Measurement of the production cross section for Z/gamma* in association with jets in pp collisions at sqrt(s) = 7 TeV with the ATLAS detector

Experiment: CERN-LHC-ATLAS (ATLAS)
Preprinted as CERN-PH-EP-2011-162
Archived as: ARXIV:1111.2690
Record in: INSPIRE

CERN-LHC. Measurements of cross sections for Z0/GAMMA* production plus jets in proton-proton collisions at a centre-of-mass energy of 7 TeV. The analysis includes the full 2010 data set, collected at low rate of collisions, with a total integrated luminosity of 36 pb-1. Data are presented for Jet PT>30 GeV and |y|<4.4 with cross sections in both the individual electron and muon decay modes of the Z0/GAMMA* as well as combined after correction to a common lepton fiducial region of |y|<2.5 and PT>20 GeV and for QED radiation effects.

View list of currently selected plots

Total number of tables: **24**. Displaying: **1** to **10**. First | Previous | Next | Last

Table 1
as: plain text, AIDA, PyROOT, YODA, ROOT, mpl or jhepwork

Cross section for Inclusive Jet Multiplicity corrected to the lepton common fiducial region and for QED radiation effects.
Location: T 2,F 15

	PT(C=JET) : > 30 GeV		
	PT(C=LEPTON) : > 20 GeV		
	RE : P P --> (Z0 + GAMMA*) JET(S) X		
	YRAP(C=JET) : -4.4 TO 4.4.		
	YRAP(C=LEPTON) : -2.5 TO 2.5.		
Z0/GAMMA* : <E+E->	Z0/GAMMA* : <MU+MU->	Z0/GAMMA* : <L+L->	

Figure 4.1 - Example of a record in the data repository, showing the connection with publications on INSPIRE[8] and arXiv[9], and also presenting the different formats available.

The records represented in the data repository of this case study are usually tables and text files containing data shown in a plot or table in a publication (see also Figure 4.2). There is also additional data that is not directly discussed in a publication. In this case, the scientific results of a sub-experiment run by a large collaboration depend on theoretical parameters and assumptions, which could assume several different values. The experimental collaborations will in this case publish an article which contains the methodology and some example of results, and then make available in numerical format a vast amount of different options to allow other experimental physicists or theoretical physicists to re-use this information, possibly in the future. These complex matrixes (see Fig. 4.3) appear in a non-standard format on the data repository- the format has been decided on together with the community so that reuse of the material is facilitated.

Regarding the size of data files, the data in the repository are rather small and can be handled with little disk space. These data are authentic - but there are some rare cases in which data might have to be corrected: a rare example being if the transmission was not accurate, or incomplete. A more likely

[8] INSPIRE is the new information platform for HEP, realized by CERN, DESY, FERMILAB and SLAC. http://inspirehep.net

[9] The arXiv is an online accessible archive for electronic preprints of scientific papers in the fields of mathematics, physics, computer science, quantitative biology, quantitative finance and statistics. It started in 1991 and expanded then to other fields. http://arxiv.org/

scenario is that data have been processed further, e.g. a better statistical analysis has been conducted and thus these data need to replace the old data. This results in the fact that sometimes different versions of one record have to be stored.

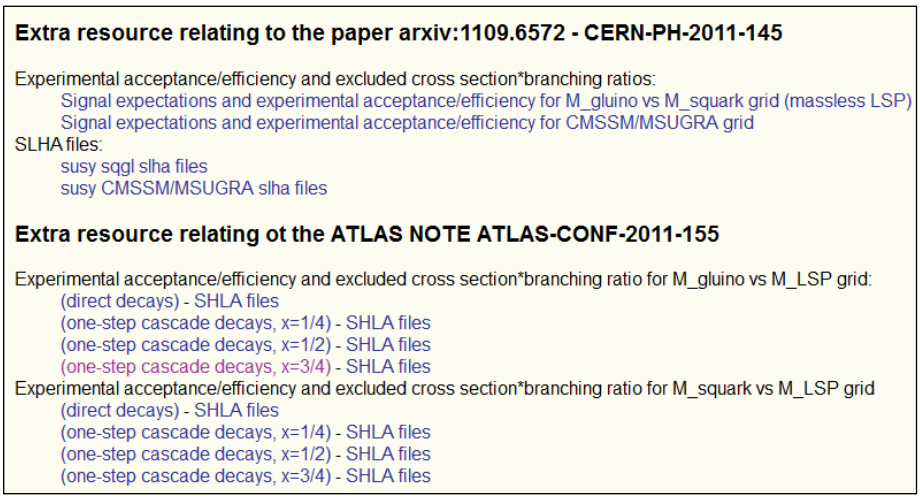


Figure 4.2 - Complex materials associated to a publication

The data submitted to the data repository is already highly processed. It reflects a certain stage in the data's lifecycle that is captured in a publication and the additional data, tables or figures are preserved in the data repository. The underlying raw data is kept available for usage by the community in their own facilities, and is in principle not useful for wider analysis and is preserved at the bit-to-bit level as part of the mission of the large experimental facilities where experiments are conducted. This is not the scope of this case study. One example, however worth mentioning, is the accessibility of data at individual centers through WLCG infrastructure, where trust is guaranteed through Grid-certificates[10] issued by the experiments.

4.3 THE SUBMISSION TO THE REPOSITORY

There are two ways and also two types of data that are generally available on this data repository. Firstly, researchers from an experiment approach repository staff in order to submit a Digital Resource (DR) (research data) associated to a publication to the repository. Secondly, repository staff extracts data that is shown in tables or plots from preprint papers and made available for reuse on the repository and creates Submission Information Packages (SIP) themselves. These two types of the DR are distinguished in the Archival Information Package (AIP) and Dissemination Information Package (DIP) and the respective provenance information is stored in the Preservation Description Information (PDI).

Any submission procedure is done by email which turned out to be a secure submission procedure during the years. Tailored solutions (such as FTP to trusted user or web transmission) are provided in case of bigger file sizes (e.g. the ATLAS SUSY scans mentioned above). Evidence on the authenticity and provenance of the DR is twofold. On the one side, for data retrieved by the repository from a publication, is automatic, as the ingestion happens on the repository side. On the other side, which is the case for newer, more complex, records, as well as numerical information supplementing the one of figures, the data are supplied by individually known individuals on behalf of their collaboration, re-using existing trust relations in a relatively small and tightly-knit community.

[10] http://lcg.web.cern.ch/lcg/digital.htm

Usually the data repository links to external digital resources such as papers, proceedings, presentations (e.g. on the digital library INSPIRE) in order to provide a comprehensive picture on authenticity and provenance. In some cases further explanatory text for documentation is provided by the data repository staff alongside the data. As mentioned above, the data repository staff is connected in the community that they know the submitters in person which eases the identification process of submitters. Alternatively, large experimental group appoint dedicated staff to liaise with the repository, and those relations of trust are inherited when a contact changes.

4.4 INGESTION AND LONG TERM DIGITAL PRESERVATION (LTDP)

After data have been submitted, the repository staff is in charge of the following steps of the preservation process. Many of the preservation workflows are triggered or done manually.

Upon ingestion to the repository, the datasets receive a unique repository data identifier. Submitted or extracted materials are checked for their technical coherence and completeness. (**LTDP-INGEST** in D24.1)

Most of the data are rather complex as can been from Figure 4.3 and 4.4; there are several small data sets or tables related to each other. The majority belongs to a single publication. Data subsets are stored in AIPs and all the data sets belonging together are then aggregated in Archival Information Collections (AIC). [**LTDP-AGGREGATE** in D24.1]

The PDIs provide comprehensive information about the data stored in the AIPs and the AICs, such as reference information like an assigned DOI (the preparation for DOI assignment to data from the HEP repository is work in progress) or a tailored abstract for the data entry provided by the repository staff. Furthermore, context information such as links to the associated publication on INSPIRE is stored in the PDI**.**

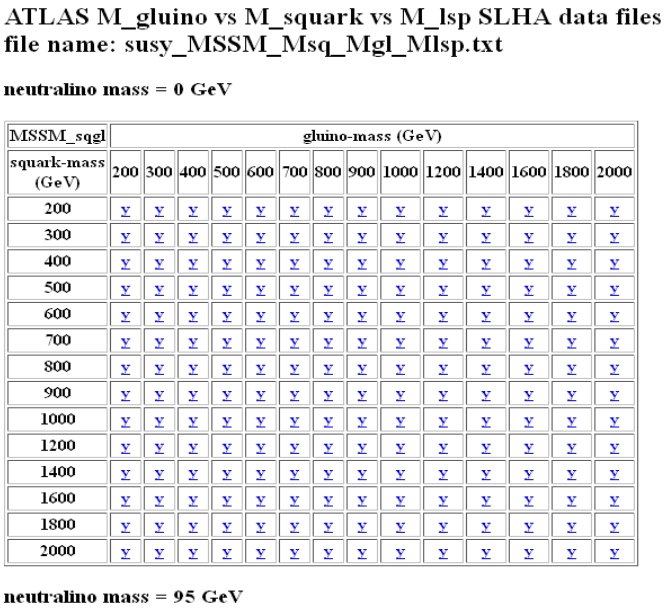
In terms of dissemination, the content of the data repository has always been linked to the content of SPIRES[11]. The latter is now replaced by its successor INSPIRE. This new digital library will allow an integration of the datasets into this community platform so that the connection between publication and data becomes even more visible (this is work in progress).

Based on the documentation on the data repository and the cited publications thereon the data consumer is provided with a comprehensive Dissemination Information Package (DIP) on the dataset (taken into account that (s)he is familiar with community standards etc.). This context information is important for the community to be able to contribute to authentication of the data and quality control. The DR provided in the SIP is usually ASCII tables and files and they will be transformed into several community standard formats such as PyRoot[12] or YODA[13] (see also Fig. 4.1). However, all these formats are not stored in the AIP but generated on the fly when the user demands them and therefore part of the DIP. This service produces ongoing costs for the repository but it has got the advantage of less effort and risk when migrating to other formats in the future. In regard to this outstanding role of the community in terms of authenticity and provenance control, we therefore propose to extend the current model by a DIP (see also next chapter - Adequacy of the model).

[11] The SPIRES-HEP database has been run by the Stanford Linear Accelerator Center(SLAC) since the late 1960's as a database of particle physics literature. SLAC also collaborates with other physics institutions around the world on SPIRES. In 1991 it became the first web-site in North America.

[12] PyROOT is an extension module to the scripting language Python that allows bindings for the ROOT class library. ROOT is an object-oriented software developed at CERN and therefore especially designed for particle physics data analysis.

[13] YODA is an abbreviation for **Y**et another **O**nline **D**ata **Analys**er. It enables the user to access experiment data, perform almost arbitrary processing on the data, and then create statistically accumulating objects like 1- or 2-dimensional histograms and define 1- or 2-dimensional cuts graphically during analysis.



ATLAS M_gluino vs M_squark vs M_lsp SLHA data files
file name: susy_MSSM_Msq_Mgl_Mlsp.txt

neutralino mass = 0 GeV

MSSM_sqgl	gluino-mass (GeV)													
squark-mass (GeV)	200	300	400	500	600	700	800	900	1000	1200	1400	1600	1800	2000
200	y	y	y	y	y	y	y	y	y	y	y	y	y	y
300	y	y	y	y	y	y	y	y	y	y	y	y	y	y
400	y	y	y	y	y	y	y	y	y	y	y	y	y	y
500	y	y	y	y	y	y	y	y	y	y	y	y	y	y
600	y	y	y	y	y	y	y	y	y	y	y	y	y	y
700	y	y	y	y	y	y	y	y	y	y	y	y	y	y
800	y	y	y	y	y	y	y	y	y	y	y	y	y	y
900	y	y	y	y	y	y	y	y	y	y	y	y	y	y
1000	y	y	y	y	y	y	y	y	y	y	y	y	y	y
1200	y	y	y	y	y	y	y	y	y	y	y	y	y	y
1400	y	y	y	y	y	y	y	y	y	y	y	y	y	y
1600	y	y	y	y	y	y	y	y	y	y	y	y	y	y
1800	y	y	y	y	y	y	y	y	y	y	y	y	y	y
2000	y	y	y	y	y	y	y	y	y	y	y	y	y	y

neutralino mass = 95 GeV

Figure 4.3 - Screenshot showing the complexity of new ATLAS SUSY data submitted to the data repository

In terms of dissemination, it should be mentioned that the information from the data repository will be integrated in the digital library INSPIRE (work in progress) so it will be made available and preserved in two different community systems. [LTDP-Transfer in D24.1]

4.5 ADEQUACY OF THE MODEL

Scientific information structures like in HEP are traditionally community-driven. Such tools and services do serve a particular need for a project or experiment. With initiatives such as DPHEP at hand, the community's awareness of preservation issues as a whole and in practices is emerging.

So far preservation standards have often been very community specific, maybe even project specific. But with the emerging awareness for interoperability and preservation across the discipline changes occurred. The data repository of this case study represents the data preservation enabled platform of the enhanced publication infrastructure in HEP. This allows interoperability scenarios with other platforms and a long term view.

Any improvement in regard to this data repository that is being considered needs to be studied crucially in regard to the specific community needs and habits. In the eyes of the strong connection between the data repository and the community in HEP, it might be worthwhile to extend the model presented in D24.1 by a dissemination phase.

- LTDP-Dissemination[this is not covered in D24.1]: Preparing the DR for (public) dissemination. This might be very relevant also in regard to authenticity and provenance of the DR as the designated user community might be able to report on possible improvements etc. and thus contribute to the quality assurance in authenticity and provenance (see also the extension of the model by a DIP above for more information).

As a general remark, it is to say that the workflow in the HEP data repository is driven by human beings more than it is focused on technical driven identification and authentication processes. It is an ingrown and community based approach and worked fine during the last decennia. With the emerging awareness of the community in regard to data sharing, it is expected that the data repository shall host and preserve more data with increasing complexity in the future.

mgl	msq	Expected # signal events (A,B,C,D,E)	Acceptance * efficiency (A,B,C,D,E)	Excluded (A,B,C,D,E)
200	200	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	n,n,n,n,n
200	300	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	n,n,n,n,n
200	400	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	n,n,n,n,n
200	500	1.13e+02,2.24e+02,1.79e+04,4.99e+02,1.74e+02	1.54e-04,3.05e-04,2.44e-02,6.80e-04,2.37e-04	n,n,n,y,n
200	600	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	n,n,n,n,n
200	700	1.17e+02,1.17e+02,1.24e+04,0.00e+00,1.17e+02	1.75e-04,1.75e-04,1.86e-02,0.00e+00,1.75e-04	n,n,n,n,n
200	800	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	n,n,n,n,n
200	900	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	n,n,n,n,n
200	1000	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	n,n,n,n,n
200	1200	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	n,n,n,n,n
200	1400	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	n,n,n,n,n
200	1600	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	n,n,n,n,n
200	1800	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	n,n,n,n,n
200	2000	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	0.00e+00,0.00e+00,0.00e+00,0.00e+00,0.00e+00	n,n,n,n,n
300	200	1.86e+02,5.52e+02,2.61e+04,4.86e+02,2.55e+02	2.91e-04,8.62e-04,4.08e-02,7.59e-04,3.98e-04	y,y,y,y,y
300	300	3.33e+02,8.17e+02,1.90e+04,8.42e+02,2.79e+02	1.39e-03,3.40e-03,7.91e-02,3.50e-03,1.16e-03	y,y,y,y,y
300	400	1.94e+02,4.61e+02,1.71e+04,5.32e+02,7.11e+02	1.40e-03,3.32e-03,1.23e-01,3.83e-03,5.12e-03	y,y,y,y,y
300	500	2.17e+02,7.46e+02,1.29e+04,7.24e+02,8.64e+02	2.19e-03,7.52e-03,1.30e-01,7.30e-03,8.71e-03	y,y,y,y,y
300	600	1.51e+02,4.79e+02,9.28e+03,4.95e+02,5.47e+02	1.85e-03,5.87e-03,1.14e-01,6.07e-03,6.71e-03	y,y,y,y,y

Figure 4.4 - Details of an exemplary data repository record

In addition, the data from the repository will be integrated in INSPIRE. It is work in progress to assign DOIs to the datasets, so that they will be made citable. With the increased visibility and added value it is expected that data submission might increase even more over time. Therefore, the workflow has to be adjusted to the new scale of data and more processes have to be automated. This is one of the reasons why improvements on the current workflow will be introduced in the next part.

4.6 SUGGESTED IMPROVEMENTS

In summary, regarding the focus on authenticity and provenance management in a data repository storing complex data associated to publications, the events **CAPTURE**, **SUBMIT** and **INGEST** are crucial. Therefore they will be specified in this part of the case study. In addition, improvements on these events of the workflow will be proposed meeting the needs of the community.

4.6.1 CAPTURE

Description

Experimental researchers in HEP mainly work in collaborations. In smaller study groups they work on a particular measurement or search for phenomenon, using a common data sample but performing some different methodological approaches to identify different characteristics, or select a particular sub-sample of data, data is processed accordingly and usually a paper is being written. The publication is submitted to arXiv (and a journal for peer review). This is often internally peer-reviewed by the entire larger group and signed by all. At the same time the researcher or a group of researchers produces the final version of the datasets, strongly related to the tables and figures appearing in the article, that are intended to be submitted to the data repository. As part of the big collaborations this workflow becomes much more complex as this published materials needs to be reviewed by corresponding boards in the experiment's hierarchy. This means that the materials are sent for a detailed internal review in order to get the approval for publication. Again, it needs to be highlighted that all materials are validated in regard to authenticity and provenance within the experiment. During these steps internal platforms for data keeping are used, as they are for the ordinary analyses workflow, which are outside the scope of this paper.

The data submitted to the data repository is a highly processed version that will not be changed – in very rare cases, it will be updated (e.g. due to the external peer review in a journal). Then, there will be an updated version added to the system (which will then be indicated as a link to the new version in the metadata of this DR). The researcher provides the context(s) in which the data were produced e.g. through a publication or information about the experiment when the data have not been published yet.

Agents

- Researcher as author of the record. In most of the cases this is a researcher on behalf of a collaboration working on an experiment.
- Researcher as administrator of his/her own data storage that serves as keeping system until the data will be submitted to the data repository which is the LTDP system

Controls

- Assessment on the identification and authentication of the author.
- Assessment on the integrity check: before submission a thorough internal peer review and control check takes place. In regard to data taking this is mainly automatically, but in regard to publication related data this is done via publication boards and a thorough internal peer review process etc.

Authenticity evidence record

- Identity of the DR
- Date and time the DR record has been created
- Context information associated to the DR

Suggested improvements

The data creators could be provided with a workflow or guidelines that help them to capture information about provenance within the processes happening internally in the experiment or collaborations until the point the data are sent to the repository.

The cooperation and collaboration between the experiment platforms and the data repository could be stronger. By collaborating even earlier in the data lifecycle potential any kind of misunderstandings or so that would result in extra work upon data submission to the data repository could be avoided.

In addition, that would improve interoperability especially concerning format standards. The researchers could help designing workflows and give feedback on the developments made by the data repository (e.g. the submission interface which is described in detail in the next section).

4.6.2 SUBMIT

Description

The DR is transferred from the keeping system – the researcher's or experiment's own data storage – to the LTDP system, the data repository. This process is authorized by the submitter as he has to actively submit the data to the repository.

Agents

- Researcher as data creator and contact person for more information on provenance
- Researcher as administrator of his/her own data storage that serves as keeping system
- Data repository staff as administrators of the LTDP system

Controls

- Identification and authentication of the submitter of the DR
- LTDP administrator checks the submitted content with the papers it belongs to. Only additional data will be preserved in the repository as the data published in the paper is preserved with the paper.

Authenticity evidence record

- Identification and authentication data of the LTDP system administrator

Suggested improvements

It is being considered to provide a submission interface, which would come with various advantages. First of all, the researchers should be asked to submit metadata about their data which could be one

way to get more provenance information directly from the submitters. In addition, there could be a quicker manual quality control by the data repository staff to validate the input. As usual they will contact the submitter for more information needed to ensure the authenticity of the data and to provide comprehensive PDIs for the AIPs later on. For a submission interface, one needs to consider some kind of identification process to prevent too many uploads of non-quality data that then results in more work for the repository staff instead of making the workflow easier. One possible identification method is the arXiv single sign-on (SSO) process.

arXiv is the platform where most of the publications in HEP are published first and in order to submit papers to arXiv, an author has to be verified. So if an author uploaded the paper to arXiv, we will trust him/her as submitter of additional material to the paper. That way authentication of the data and identification of the submitting researcher could be ensured.

Another SSO could be provided for accounts related to the experiments submitting data to the repository. An association to a HEP experiment would then be considered to be trustworthy to submit authentic data from this experiment.

Furthermore, the submission interface would solve the problem of tailored solutions for bigger files. In cooperation with the researchers, it could be structured in a way that even complex data sets could be uploaded without needing further engagement of the data repository staff.

4.6.3 INGEST

Description

The DR submitted in the SIP receives an identifier in the LTDP system. It is checked for technical coherence and completeness as well as for authenticity and provenance by the data repository staff. If the latter are not sufficient, the data submitters will be contacted for additional proof. If necessary, the data will be migrated into the LTDP system standard format. At the end of the ingestion process, AIPs with as comprehensive PDIs as possible are created.

Agents:

- Data repository staff as administrators of the LTDP system

Controls:

- Assessment on format migrations if the data was not permitted in a format that can be stored in the LDTP; statement that the content of the DR has not changed
- Assessment on the authenticity and provenance evidence contained in the submitted DR

Authenticity evidence record:

- New identifier of the DR in the LDTP system
- Original identifier of the submitted DR if existent
- Date and time the ingestion has been completed
- Digest of the AIP produced by the ingestion process
- Identification data of the LTDP system
- Identification and authentication data of the LTDP system administrator

Suggested improvements

As a consequence of the improved submission process, the check for technical coherence and completeness could be done automatically. This might be useful as with the emerging awareness within the HEP community there are more and more complex data submitted and it will be impossible for the data repository staff to check them by hand in the future.

In addition, there could be a standardized format for the DR in the keeping system that can be stored in the LTDP so that there is less risk that information might be lost concerning the content as well as the authenticity and provenance information stored in the PDI.

This becomes even more relevant when considering complex dataset that cannot be fitted into the "classical" data repository record and thus need extra treatment (as can been by the ATLAS SUSY data). The growing complexity and size of datasets being submitted should not be underestimated in terms of long term digital preservation. Another important aspect to consider in terms of ingestion will be authenticity and provenance in versioning of materials which might become more important due to increased reuse.

4.6.4 Summary of all the improvements proposed

To sum it up, improvements are above all to be done in two ways – one in regard to an even stronger collaboration of data repository and community, another one in in regard to automation of the workflow as the current workflow is based on human beings.

As the community's awareness for data preservation is increasing, there will be more data to store and it is not feasible that all of this is done by the data repository staff. As the community is well-connected, it might not be a big problem to develop solutions that are satisfying and practicable for both the researchers and the staff taking care of the LTDP though.

Overview over important HEP community platforms

SPIRES: The SPIRES-HEP database has been run by the Stanford Linear Accelerator Center(SLAC) since the late 1960's as a database of particle physics literature. SLAC also collaborates with other physics institutions around the world on SPIRES. In 1991 it became the first web-site in North America.

INSPIRE: INSPIRE is the new information platform for HEP, realized by CERN, DESY, FERMILAB and SLAC. It will replace SPIRES as main database for HEP literature http://inspirehep.net

arXiv: The arXiv is an online accessible archive for electronic preprints of scientific papers in the fields of mathematics, physics, computer science, quantitative biology, quantitative finance and statistics. It started in 1991 and expanded then to other fields. http://arxiv.org/

5 ARTICULATION WITH THE REST APARSEN WPS AND TASKS

Here we describe how this work is related with the other work packages and tasks of APARSEN.

WP	**Notes**
WP11 Common Vision (M1-M18)	The results of the current deliverable can be related with the following candidate objectives for the common vision: a/ Identification of a common terminology for handling the preservation of digital resources with the aim of tracking information related to the events and the actors in the DR lifecycle b/definition of a conceptual framework related to the ingestion and to the preservation phases able to provide a comparable set of elements for assessment integrity and authenticity
WP13 Coordination of common standards (M4-M48)	The results of the current Deliverable are related to T1310 (Analysis of current standards) since we refer to: - ERMS standards as developed by ISO (15489, 23081) and by DLM Forum (MOREQ) - Standards for trusted digital repositories: ISO 16363 / DIN 31644, Data Seal of Approval - PREMIS as common dictionary
WP22 Identifiers and Citability	An authoritative link is a crucial part of the authenticity assessment and is handled as part of the authenticity evidence record for any component of the digital resource and in any phase of its lifecycle. The quality and persistency of the identifier will contribute to reinforce the authenticity assessment itself.
WP25 Interoperability and Intelligibility (start: M20-M33)	The results of the current Deliverable, specifically the mappings, as well as the integration approaches and systems/tools that are discussed, are important for achieving provenance interoperability. Furthermore, provenance can be used to interpret data, an element which is essential in the preservation of knowledge, therefore the results of this deliverable also relate to Intelligibility. The guidelines for authenticity assessment are specifically dedicated to develop a model based on a common terminology able to make interoperable and comparable the information provided in the whole digital resource lifecycle. The schema here developed and the detailed explanation for each activity, event and actor could also play a role for providing a contextualized knowledge and intelligibility for the preservation function.
WP26 Annotation, reputation and data quality	In many cases annotations are used for documenting the authenticity and the provenance of the various artefacts. The models here discussed can be adopted for that need.
WP31 Digital Rights and Access Management	Provenance and authenticity are a crucial aspect of digital rights, therefore the discussed models and their mappings are strongly related to this.

(M27-M38)	
WP35 Data policies and governance (M27-M38)	The issues discussed in the deliverable are strictly related to the policies applied for handling the preservation function and can strongly contribute to guarantee the sustainability of the repositories.

6 INTEGRATION AND OUTREACH

In planning and carrying out the activities of WP 24, that are documented in this deliverable and in the companion deliverable D24.1 *Report on authenticity and plan for interoperable authenticity evaluation system* [1], two major concerns have been, on one hand to provide adequate integration with other research projects and standardization initiatives in the area, and on the other hand to make sure that the results of the RTD activity could be actually translated into practice. In this section we shall briefly address these issues. This same section is to be found in the companion deliverable as well, since it actually refers to both deliverables which are strictly interconnected, being the first one the formulation of the methodology and the second one the discussion of case studies that we have carried out to test on the field its effectiveness.

A consistent effort has been devoted to investigate the literature and to develop a comprehensive state of the art, in order to properly defragment the several different proposals that have been made in the literature and to get to the definition of a simple model of the relevant events in the digital resource lifecycle and to the specification of the authenticity evidence that should be gathered in connection with each of them. To do that, we have reviewed about twenty major research projects and the most relevant standards, recommendations and guidelines for keeping and preserving digital resources (see D24.1 sect. 2).

Our main connection is certainly with CASPAR and InterPARES, without any doubt the two projects that have devoted the most attention to the problem and produced the most significant results. We have taken from InterPARES the central role of the lifecycle in the management of the authenticity of digital resources, and from CASPAR the crucial concept of authenticity protocol, i.e. the need to introduce formal procedures for the gathering of the related evidence.

We have based our proposal on the standards as well, on OAIS of course, which has been the main reference for the preservation part of the lifecycle and for the transformations that the digital resource undergoes during that phase, but also on standards and recommendations for recordkeeping systems, as for instance ISO 15489 for the need of documenting record transactions and action and location tracking. Similarly we have tried also to harmonize with the MoReq2 and MoReq2010 recommendations, since we are convinced that, for a proper management of the authenticity, one needs also to carefully tackle all the transformations that a digital resource undergoes during the recordkeeping phase, that is before it enters long-term preservation.

With specific reference to the MoReq specifications, and to MoReq2010 in particular, our proposal can contribute to provide normalized workflows for supporting the interoperability, not only among different ERMS but also with future long-term preservation repositories. Moreover, the functional framework we refer to for assessing authenticity and for producing authenticity evidence records compliant with OAIS, is based on a categorization of events and actors which is meant to be compatible with recordkeeping system based on MoReq specifications.

As for the ability of successfully transferring the results of the RTD activity to real life environments, a problem not often enough addressed by the academic community, our main principles have been usability and flexibility. Usability means that the model and methodology one intends to propose should not indulge in theoretical narcissism and self-praise, but should be instead limited to a minimum core of information, controls and actions. That would make it acceptable to people who operate in real life environments and are willing to accept only what they can actually understand and rate important enough to be worth the price of changing their current practices in order to accommodate the innovation.

In our case usability arises from the simplicity of model of the digital resource lifecycle (see D24.1 sect. 4), which is based on a limited *core set of events* that correspond to the relevant transformations affecting the authenticity and the integrity of a digital resource. For each event we define an *Authenticity Evidence Record (AER)*, that is the set of evidence item that should be collected and preserved to allow assessing the authenticity of the digital resource at a later time. The AER should be

not intended as a mandatory list, but rather as a *template*, that is a general reference to be adapted to each specific case.

By flexibility we mean the ability to formulate a proposal that could be tailored to meet the requirements of a specific environment. This has indeed turned out to be a central issue in the case of authenticity, since different communities may have different needs and attach to this concept different meanings. The balance between cost and effectiveness may therefore have quite different points of equilibrium. To allow flexibility, we have devised a set of guidelines (see D24.1 sect. 5) whose purpose is to guide the process of adapting the model and the AER templates to the specificity of the individual environment, and to define the *Authenticity Protocols*, that is the procedures that should be followed to perform the controls and to collect the proper evidence.

So far the strategy, but, thanks to the results of case study analysis presented in this deliverable, we may actually claim that the outcome of the field test of our approach has been encouraging. The guidelines have proved to be helpful and effective in two ways. On the one hand, the reference model and the templates for the AER have been an effective tool in analysing the current practices in the repositories that we have studied, by providing a guide to model the workflow and a sort of checklist to understand which authenticity evidence was/should have been collected. On the other hand, the guidelines have helped in adapting the general templates to the specificity of the context and have provided an operational guide to the definition of the authenticity protocols.

We may therefore say that the results of the RTD activity in WP 24 are well suited for dissemination and to be translated into practice to improve the current (and often very limited) practices in managing authenticity and provenance presently held in keeping and preservation systems. In the future we plan to further disseminate of our approach both within APARSEN and in the larger user community outside the project, by replicating the process we have already successfully tested in the case studies to improve the practices currently used in their repositories. For instance, SFTC, an APARSEN partner that manages a number of large repositories, is willing to cooperate and to involve the repository managers in discussing how the results of WP 24 can be used in their repositories. Other smaller organizations may just take these ideas on board in their plans for system upgrades.

We also plan as a further development, in cooperation with SCIDIP-ES project which is part of the worldwide Earth Science Long Term Data Preservation program, to incorporate our methodology into the SCIDIP-ES Authenticity Toolkit, which is part of the services and tools that the project proposes to implement for Earth Science data preservation.

7 CONCLUSIONS

The main purpose of the activities documented in this report has been to test the model and the guidelines for the management of authenticity evidence developed within APARSEN WP 24 *Authenticity and Provenance* against real life environments. This has proved fruitful and interesting, and anyway a necessary complement to the other activities carried out in the WP and to the more formal and theoretical results presented in the companion deliverable D24.1 *Report on authenticity and plan for interoperable authenticity evaluation system* [1].

Shifting to a practical ground and facing the actual problems that arise in the management of a repository has indeed been an important move, since, as we have already remarked, the heart of the problem in this area is to try to fill the gap that still divides the mostly theoretical results of the scientific community from the actual practices carried on in most repositories, and to reduce the fragmentation among the different approaches that prevents interoperability.

The model and the guidelines that we have proposed, and then checked in the case studies, have proved to be helpful in both regards. From one hand they have shown to be easily applied and well understood in all the test cases, and from the other hand the simple and yet rigorous concepts of *Authenticity Evidence Record (AER)* and *Authenticity Protocol (AP)* may provide a common ground for the management of authenticity evidence and for exchanging it among different systems.

In at least one of the case studies, the Vicenza public health care system, the guidelines have been applied to their full extent, i.e. from the preliminary analysis, to the identification of the relevant lifecycle events, to the detailed specification of the AERs. Moreover for a specific event the process has been carried out to the formal definition of the authenticity management policy, that is to the specification of the AP.

In the case of the UK Data Archive a long history of ingesting heterogeneous data collections has led to procedure-based controls over processes and they remain some distance from implementing the granular information model implied by [1]. Their evaluation of the model has provided a good analytical tool to help direct their current approach of integrating procedural standards into machine-actionable rule sets as they transfer their holdings to the lifecycle-focussed DDI-L standard. The Archive has a clear strategy of aligning their organisation processes with appropriate high-level standards including the Data Seal of Approval and ISO27001 as a precursor to integrating the model into their repository systems to ensure a sustainable implementation which can respond to changes in the evolving best practice in this area.

In the field of HEP, as one of the case studies showed, the model has been implemented to a rather small extend yet as research data management is community driven and the community was not aware of guidelines in the past. However this changes with scale and complexity of the output produced and improvements and adjustments will have to be done in the workflow.

In each case study, the specificity of the legal requirements and the different awareness and risk perception of stakeholders and users have had an impact in evaluating the role and the consequences produced by the events analysed and, even more, in deciding the specific content of AERs. In general, it could be said that the case studies have been able to provide a consistent picture of actions to be taken for assessing authenticity and provenance and a general framework for comparing different preservation models. In all cases, referring to the guidelines has provided valuable help, both in pointing out some weakness in the current practices and in providing a reasonable way to fix the problem.

Finally, it is worth to point out that, on the whole, the contribution of WP 24 has concentrated on one aspect of the preservation process that has not always been given sufficient attention: the fact that users should be able to evaluate the degree of trust provided when the DR are disseminated, not only on the basis of the trust they feel for the repository, but also (and mainly!) from the documentation accumulated in the course of the DR lifecycle, and made available to them as part of their Authenticity Evidence History.

REFERENCES

1. APARSEN Project: Deliverable 24.1. Report on authenticity and plan for interoperable authenticity evaluation system. (2012)
2. Gentil-Beccot, A., Mele, S., Holtkamp, A., O'Connell, H. B., & Brooks, T. C. (2009). Information resources in High-Energy Physics: Surveying the present landscape and charting the future course. Journal of the American Society for Information Science and Technology, 60(1), 150-160. doi:10.1002/asi.20944
3. Giaretta, D.: Advanced Digital Preservation (specifically chapt. 13 and sect. 17.11). Springer-Verlag, Berlin-Heidelberg (2011)
4. Linee guida per la dematerializzazione della documentazione clinica in laboratorio e in diagnostica per immagini normativa e prassi , 19, March 2007, http://bancadati.digita-lex.it/public/files/pdf/0143_5-Linee-Guida-Dematerializzazione.pdf
5. Parse.Insight Project Members. (2010). Insight into digital preservation of research output in Europe: Deliverable D3.3 - Case studies report. Distribution (p. 248). Retrieved from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-3_CaseStudiesReport.pdf
6. South, D. (2011). Data Preservation in High Energy Physics. Arxiv preprint arXiv:1101.3186, (May), 1-18. Retrieved from http://arxiv.org/abs/1101.3186
7. UNI 1386:2010 SInCRO - Supporto all'Interoperabilità nella Conservazione e nel Recupero degli Oggetti digitali, http://webstore.uni.com/unistore/public/searchproducts.

APPENDIX - INGEST AT UK DIGITAL ARCHIVE

This appendix describes some of the actions and transformations undertaken during ingest at UK Digital Archive to support the core description of a (currently) procedurally-driven approach.

Quantitative data

In practice, the vast majority of quantitative microdata files are deposited in SPSS format and it is also by far the most popular dissemination format. Processing a quantitative study therefore typically entails:

- converting the data into SPSS .sav format where appropriate, if that is not the deposit format;
- performing integrity and validation checks on the data according to its processing creating dissemination and preservation formats (usually SPSS, Stata and tab-delimited text).

Descriptive statistics and Data dictionary

In combination with a visual examination of the data, the descriptives and data dictionary output from SPSS conversion provides the basis for the following checks:

- unlikely or impossible values for interval variables;
- undefined or incorrect values for nominal (categorical) variables;
- completeness and interpretability of value labels for nominal (categorical) variables;
- missing values appear sensibly and consistently defined (for example, if 'refused' is defined as missing for one variable, is it defined as missing for other variables?).

Frequency distribution generation to supplement the above checks

The *Quantitative Data Processing Procedures* note that "The SPSS frequencies command provides useful information in addition to that provided by the descriptives command. For example, consider a nominal (categorical) variable that is supposed to range between 1 and 8, but the descriptives output shows a maximum value of 18. From the descriptives output it cannot be seen whether there is a single case with a value of 18 (in which case, it's most probably a data entry error), or whether there are many values between 8 and 18, in which case, a more substantial problem exists (either incorrect mapping of value labels or very 'dirty' data)."

Addition of display labels

Many studies contain variables that are used to perform weighted statistical analysis (a technique typically used to make a sample representative of some important criteria, such as population figures). However, it is desirable to take off the weighting for ingest processing and archiving, unless the depositor has specifically requested that the weight remain on (this is recorded in the Note file for information).

Documentation is also checked to ensure that comprehensive information on weighting variables and their construction and use are given. The depositor is contacted and extra information requested if that provided is not adequate.

Identification of cases with anomalous values

For example, the documentation and value labels in SPSS for the variable mstatus (marital status) may be described both in the documentation and the SPSS value labels as 1=married, 2 = divorced, 3 = single; yet the descriptives command reveals a maximum value of 6, and the frequencies command reveals a substantial number of cases of values 4 and 5 as well. As a result, the meaning of codes 4-6 is unknown. They may either be invalid (i.e. data errors) or they may be incorrectly labelled (i.e. the data are correct but the codes 4 to 6 have not been defined by value labels or in the documentation).

Tests for non-integer values so cases can be examined where this would not be appropriate

Examination of Print and Write formats, variable width and decimal places

Print and write formats may display variables with many decimal places as an integer (e.g. 1.348726 being displayed as 1), or with only one or two decimal places. This may confuse naïve end users of the data, though the data are correct

The processing script currently in use at the Archive creates STATA from .sav format to avoid issues with conversions which rely on .por format

Variable widths may need to be adjusted in source material to avoid incorrect conversion to STATA or Nesstar formats.

If such formats inevitably lead to rounding, truncation or loss of data upon conversion to other preservation or dissemination formats, they must be altered prior to conversion

Known bugs with commercial software applications

Such as display issues with SPSS software reporting the incorrect number of cases

Rounding errors in material with an Excel or Access

Addition and editing of variable and value labels. Labelling is usually carried out in SPSS prior to data format conversion. In order to enable 'tracking' of the edits undertaken, the labels are added using a syntax file rather than being added directly to the file via the SPSS graphical interface. The syntax file is then archived with the study.

When variable and value labels have been added/edited as necessary, the results are checked by running frequencies to check the amended variables. If the additions/edits have been successful and all other errors in the data fixed or noted, conversion from SPSS to dissemination and archival formats (usually STATA and tab-delimited text) may be undertaken.

SPSS to STATA transfers

Most problems are evident from very basic post-transfer checks, but not all; information on more complex problems that may result during transfer (and their solutions) are available internally for Archive staff, covering incorrect value transposition, incorrect missing value transfer and problems with variable name transfer.

Basic information on potential truncation during transfer to STATA is given in the RTF 'SPSS to STATA' RTF document generated as part of the processing script outputs, based on transfer to STATA version 8.0. The likely effects of transfer are as follows:

- String variables in the SPSS file with a defined width of >80 characters (the standard STATA limit) or >244 characters (the STATA Special Edition (SE) limit) will be truncated.
- Variable labels in the SPSS file of >80 characters (the STATA limit) will be truncated.
- Value labels in the SPSS file of >32 characters (the STATA limit) will be truncated.
- String variables that have value labels in the SPSS file will lose these in STATA.
- Non-integer values that have value labels in the SPSS file will lose these in STATA.

Tab delimited files

The Archive's current preservation format is tab-delimited text (either of the ASCII or UNICODE character set). The following checks are made, using the original SPSS .sav file and RTF data dictionary file (created by the (internally developed) script as a guide:

- all cases have transferred successfully
- the variable names are included in the top line of the tab-delimited file
- all variables have transferred successfully.

Qualitative Data

Details are provided in *Qualitative Data Processing Procedures*.

Qualitative data collections may contain a variety of materials, but currently the majority of deposits comprise a set of interview transcripts and accompanying documentation.

Quali Software

Qualitative software packages such as CAQDAS, NUD*IST, ATLAS-ti and WinMax have export facilities that enables one to save a whole 'project' consisting of the raw data, coding tree, coded data and associated memos and notes. For archival purposes the raw data, the final coding tree and any useful memos should be exported prior to deposit. Depositors are requested to do this at the acquisition stage and checks should be made prior to processing to ensure it has been done. Coded data are retained but not preserved, as they cannot as yet be exported in a common non-proprietary format.

Interview Templates

A large proportion of qualitative digital data collections are deposited in MS Word format. However Rich Text Format (RTF) is the standard UKDA preservation format for this kind of text, and qualitative data are also typically distributed in RTF, for better cross-platform usability. To produce the standard Qualidata interview transcript format, an RTF interview transcript template is constructed for the addition of:

- standard header information
- Interviewer/Respondent demarcation tags

Post-preparation the transcript is read thoroughly to check for:

- items that could be added to enhance the data list (e.g. details of employment, education (suitably anonymised));
- logical consistency (e.g. ‘find and replace’ errors);
- correctness of formatting;
- confidentiality (see below).
- During this thorough reading of each transcript, notes are made on potential index terms for the list of keywords (resource discovery)

Confidentiality and anonymisation of textual data

Although UKDA users sign a legally binding access agreement to re-use data, and in that undertake to respect guarantees of anonymity, consistent with the original investigator’s undertaking, no information that clearly breaches the confidentiality of the respondent or any other person or entity may be present in the dissemination version of the data collection.

Confidentiality is of paramount importance and depositors are thus requested to edit material prior to deposit. The task of the processing officer is to check this has been done well and consistently.

In some cases it can be very difficult to disguise the identity of participants without introducing an unacceptable distortion into the data, and so full anonymisation may be impossible. Alternative solutions may be available, such as the restriction of user access to certain interviews within the data collection.

Automated search and replace techniques may be used (such as MS Word’s ‘Search and Replace’ function), but additional proofreading is always be carried out as automated processes are not foolproof

Pseudonyms and pseudo place names must be the same as those used in any prior publication by the depositor/principal investigator;

A file for internal UKDA use only, containing details of cross-referencing for pseudonyms to the original names (i.e. a 'key') is compiled and preserved.

The Data List

The Data List accompanying a qualitative data collection lists the key demographic characteristics of interviewees that define the sampled population, such as year of birth/age, gender, and perhaps geographical region. Depositors are routinely asked to collate and supply this information themselves prior to deposit of the data collection, which they may already have done in the course of conducting fieldwork or analysing data. The elements to be listed depend upon what has been recorded by the original researcher and research by processing staff which may pick out trends, so most listings will vary. Completion and review of the Data List also supports quality assurance in term of missing unique identifiers, missing data or if there is some variation in materials; such as a mixture of interview transcripts and focus group transcripts, fully transcribed files or summary notes which must be recorded in descriptive metadata.

Documentation

The ingest of documentation provided with the data is covered in *Documentation Processing Procedures*

Most documentation is currently deposited in MS Office formats, i.e. Word, Rich Text Format (RTF) or Excel. The primary documentation dissemination format created at the Archive is Adobe Portable Document Format (PDF). Conversion to preservation file formats may include generation of bookmarks to aid navigation followed by manual quality assurance of the converted files and software validation of the file format as PDF/A

Read notes and files

Two metadata files, named 'Read' and 'Note' files, are compiled during study processing as noted in the section on ingest controls above. They are held in the Calm processing database. Both files contain information about processing history - checks carried out, problems discovered, etc., but are created for different purposes

- the Read file is for external display on the UKDA website, and is distributed to the user with the data collection download package;
- the Note file is for internal use only.

The details are entered under broad headings and are output as html files.

Standard metadata recorded with the files include:

- Acquisition ID
- Data Collection ID
- Acquired Date
- Released Date
- Service (Core, HDS, SDS etc)
- Depositor name
- Date released for Ingest
- Assigned to (ingest team member)

Free text data entry headings include:

- Details of data collection received
- Details of data files
- Details of documentation

- Details of hardcopies
- Filename changes

For qualitative data collections, information is added regarding work carried out on the data collection, e.g. clear notes on anonymisation techniques used and the replacement of original identifiers. As the Read file is visible to users, it must not contain any confidential information that may have been included in the Note file, such as keys to replaced names.

Mixed methods data collections must include information on whether and how files within the data collection inter-relate and may be linked between quantitative and qualitative files (e.g. how a case in an SPSS questionnaire file may link to the same respondent's RTF interview transcript)

Label files

Once all data and documentation files and formats are complete, a text file is automatically generated containing the name of each file for distribution to users. Processing staff add a short (<60 character) description of each file available to consumers. This file is used by the online documentation table in the Archive/ESDS catalogue and a formatted rtf version is included in the download package

With the exception of some descriptive catalogue information designed to support resource discovery all Preservation Description Information including Reference, Provenance, Context and Fixity information collected at all stages of the acquisition and ingest process is included in the AIP.

As the Archive moves towards DDI3 implementation the metadata collected will be refined and extended and the LifeCycle event model from the standard will be used to map information collected directly to the OAIS information model for preservation description information.

Persisten identifiers

Once Ingest staff have requested the addition of a data collection to the Archival Storage system the online catalogue record is updated. Under recent developments in partnership with DataCite and the British Library each new data collection is accompanied by the assignment of a digital object identifier.

The Archive distinguishes between two forms of alteration post ingest:

- new version (Definition: when there is a change to the preserved metadata);
- new edition (Definition: when there is change to data or documentation).

All new versions and editions are supported by a descriptive change log of the amendments made since the last release available on a change history page for the data collection, in addition a new DOI is generated for each new edition.

DIP generation

Once ingest processing is complete and a catalogue record is published the downloadable DIP is generated by running scripts against the completed AI