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

D24.1 Report on Authenticity and Plan for Interoperable Authenticity Evaluation System

Document identifier:	**APARSEN-REP-D24_1-01-2_5**
Due Date:	2012-02-29
Submission Date:	2012-04-30
Work package:	WP24
Partners:	CINI, FORTH, SBA
WP Lead Partner:	CINI
Document status	FINAL
URN	urn:nbn:de:101-20140516151

Abstract:

The first part of the report is devoted to the state of art. We analyse the main international projects in the field, as well as the standards, recommendations and guidelines for keeping and preserving digital objects, with a special attention on the management of provenance and authenticity. The state of the art is completed by an extensive reference list and by an appendix where all the major projects in the area, their goals and their results are individually presented.

On the whole, the state of the art testifies that significant scientific contributions have been given, and that a good level of theoretical formalization has been achieved in this area, even if a large gap still divides the mostly theoretical results of the scientific community from the actual practices carried on in most repositories. This gap needs to be filled with more concrete guidelines and proposals.

Acting in this direction we propose a model of the digital object lifecycle, in order to identify the main events that impact on authenticity and provenance and to investigate in detail, for each of them the evidence that has to be gathered in order to conveniently document the history of the digital object.

The crucial problem to be addressed is, of course, interoperability, since along its lifecycle the digital object may go through several changes of custody, and therefore the authenticity evidence needs to be managed and interpreted by systems, both keeping and preservation systems, which may be different from the ones that gathered it. Thus, the authenticity evidence needs to comply with a common standard. Achieving such a standard is a quite an ambitious goal, but some basic guidelines can be developed. The model and the guidelines proposed in this report may be considered as a preliminary step in this direction, and a basis to derive operational guidelines to improve the current (and often very limited) practices in managing authenticity and provenance in keeping and preservation systems.

The report also documents other important activities that have been carried as part of APARSEN WP24. Interesting original results are presented about provenance interoperability and reasoning (described in detail at ID2401), which include a discussion of the mapping between different provenance models, and the proposal of a set of relevant reasoning rules for reducing the amount of provenance information that has to be explicitly stored and for making corrections easier. A further discussion is devoted to secure logging mechanisms, a specific aspect of the problem, which has a significant impact on managing authenticity.

Delivery Type	REPORT
Author(s)	Silvio Salza, Mariella Guercio, Monica Grossi (CINI); Stefan Pröll (SBA); Christos Stroumboulis, Yannis Tzitzikas, Martin Doerr, Giorgos Flouris (FORTH)
Approval	David Giaretta/Simon Lambert
Summary	Report on Authenticity and how to manage it over the whole lifecycle
Keyword List	Authenticity, provenance, digital resource lifecycle, authenticity evidence, secure logging, secure storage, provenance models, interoperability, reasoning, inference rules
Availability	☒ Public

Document Status Sheet

Issue	Date	Comment	Author
0.1	04/11/2011	Preliminary draft	Silvio Salza (CINI)
1.0	07/01/2012	Complete draft in line with comments from David Giaretta and incorporating the contributions from SBA and FORTH	Silvio Salza (CINI)
1.1	08/01/2012	Consistency checks	David Giaretta (APA)
1.2	15/01/2012	Incorporates David Giaretta's and Andreas Rauber's updates, and introduces changes in terminology in sect. 4.	Silvio Salza (CINI)
2.0	10/04/2012	Revised version, improved according to the reviewers' comments. Sections 5, 8 and 9 have been added, and sections 1 and 2.2.1.4 have been improved. Section 7 has been revised according to the new version of ID2401 (provenance interoperability and reasoning)	Silvio Salza (CINI), Yannis Tzitzikas (FORTH)
2.1	23/04/2012	Final draft, in which a new section (now sect. 9) has been added to discuss integration with other projects and the transfer of the results into practical environments.	Silvio Salza (CINI)
2.2	2012-04-28	Format updates and minor tidying	David Giaretta (APA)
2.3	2012-04-29	Minor amendments	Silvio Salza (CINI)
2.4	2012-06-28	Incorporates changes introduced to Section 7, consequent to comments in the M16 review.	Silvio Salza (CINI) Yannis Tzitzikas (FORTH)
2.5	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

Consistent with the plan defined for WP 24 and the specific objectives identified for tasks 2410, 2420 and 2430, this report, in the form of deliverable D24.1, investigates and discusses how best to capture evidence about authenticity and provenance and to evaluate authenticity and provenance in a common way that allows the interoperability required to support changes in data holders and processing [10, 16, 17, 19, 24, 43]. For this purpose, we propose a model for managing authenticity and provenance through the digital resource lifecycle which may constitute a sound basis for deriving operational guidelines for the actual practice of digital preservation.

The first part of the report (section 2) is devoted to the state of art with specific reference to the main projects, standards and recommendations on digital preservation with the aim of capturing the outputs relevant for building a consistent framework and related guidelines to assess the authenticity of digital resources, on the basis of a common and well recognized terminology. We first analyze in section 2.1 the main international projects in the field, with special regard to their results and their proposals related to the authenticity and provenance. Here InterPARES and CASPAR have been considered crucial because of the attention they have paid to the authenticity concepts analyzed as part of the chain of custody and with reference to the need of a coherent methodological approach. CASPAR in particular should be considered as the basis for our methodological investigation with respect to the definition of authenticity management procedures compliant with the OAIS reference model. Later, in section 2.2, we analyze the standards, recommendations and guidelines for keeping (i.e. the time before going into a preservation system) and preserving digital resources, with specific reference to the design, the management and the certification of electronic records management systems (crucial for the analysis of the legal authenticity) and for the long-term digital preservation systems. The ERMS recommendations (like ISO 15489 and MOREQ) provide requirements and principles based on the scheme that a digital resource should be efficiently and effectively handled in the course of its lifecycle according to the digital continuity principle. Similarly, the recommendations and the standards for the creation and the assessment of the digital repositories (see ISO 16363) consider the tracking of events at the submission phase and in the repository itself as central for the sustainability and the quality of the preservation.

On the whole, the state of the art testifies that significant scientific contributions have been made, and that a good level of theoretical formalization has been achieved in this area, even if a large gap still divides the mostly theoretical results of the scientific community from the actual practices carried on in most repositories. This is a gap that needs to be filled with more concrete guidelines and proposals specifically dedicated to identify, normalize and interconnect the various phases of the lifecycle of the digital resources and related information and responsibilities, by focusing on the transformations that may affect their authenticity and provenance.

The conceptual and methodological background is summarized in section 3, and this *basic framework* for authenticity and provenance management represents a solid basis and a good starting point to shift the investigation towards a more practical ground and at the same time to a more systematic and complete approach, an effort that is attempted in the following section.

The crucial point is concentrating on the digital resource lifecycle, since, in order to properly assess its authenticity and provenance, one must be able to trace back all the transformations the digital resource has undergone and the relevant events that have affected its management and keeping since its creation, and that may have affected its authenticity and provenance. And in connection with these transformations and events one needs to collect and preserve the appropriate evidence that would allow, at a later time, to make the assessment.

As generally acknowledged by all the relevant international projects, the authenticity assessment cannot be limited to a final verification of the bit-stream integrity, but requires a series of interrelated controls to be carefully performed and documented on a systematic basis along the whole digital resources lifecycle.

To this purpose in section 4 we develop a model of the digital resource lifecycle, in order to identify the main events that impact on authenticity and provenance and to investigate in detail, for each of

them the evidence that has to be gathered in order to conveniently document the history of the digital resource.

A main concern is, of course, interoperability, since along the lifecycle there may be several changes of custody, and therefore the evidence about authenticity and provenance needs to be managed and interpreted by systems, both keeping and preservation systems, which may be different from the ones that gathered it. Thus, to ensure interoperability, the authenticity evidence needs to comply with a common standard, based on shared terminology and on a consistent cross-domain framework of actions and procedures able to describe and document for assessment all the relevant aspects of the preservation function.

Achieving such a standard is a quite an ambitious goal, and a very complex process, and requires, of course, time, consensus and a thorough discussion. In section 5 we propose a preliminary step in this direction by translating the model presented in section 4 into practical guidelines which could be actually adopted in specific real-life environments, in order to improve the current (and often very limited) practices in managing authenticity and provenance in keeping and preservation systems. In our opinion these guidelines can be also considered as a first substantial step in the direction of interoperability, since they provide a uniform and systematic framework for the management of the authenticity evidence, therefore facilitating the exchange of information among heterogeneous systems.

Section 6 deals with a specific aspect of the problem which has a significant impact on managing authenticity: secure logging mechanisms. More specifically it investigates the problems related to performing semi-automated audits of log files in archives and to assessing the capabilities of an archive for providing evidence of interactions - be it regular or even malicious - with the system in a secure way.

Section 7 deals with provenance interoperability and reasoning. It discusses the mapping between different provenance models, as OPM (Open Provenance model) promoted by W3C and CRMdig an extension of the CIDOC CRM ontology for capturing digital resources, and proposes a set of relevant reasoning rules. A wider account of the activity summarized here has been presented in the internal deliverable ID2401.

Finally, in section 8 we describe how this work is related with the other work packages and tasks of APARSEN, in section 9 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 10 we give our concluding remarks.

Further material is presented in the Appendix that provides individual descriptions for all the research projects analyzed in the state of the art.

2 STATE OF THE ART

2.1 LONG TERM DIGITAL PRESERVATION RESEARCH PROJECTS

In this section we will concentrate on the aspects related to authenticity and provenance, as they have been treated in recent research projects on digital preservation. Our analysis begins with InterPARES and CASPAR, the most significant projects for the attention they have devoted to authenticity issues.

2.1.1 InterPARES[1]

The InterPARES research projects have addressed the creation, maintenance and preservation of digital records, with specific reference to authenticity. A major finding is that, to preserve trustworthy digital records (i.e., records that can be demonstrated to be reliable, accurate and *authentic*), records creators must create them in such a way that it is possible to maintain and preserve them. This entails that a relationship between a records creator and its designated preserver must begin at the time the records are created.

The InterPARES 1 research (1999-2001) was undertaken from the preserver's viewpoint [5, 63]. Three central findings emerged from it: 1) there are several requirements that should be in place in any recordkeeping environment aiming to create reliable and accurate digital records and to maintain authentic records; 2) it is not possible to preserve digital records but only the ability to reproduce them; and 3) the preserver needs to be involved with the records since the beginning of their lifecycle, to be able to assert that the copies that will be selected for permanent preservation are indeed authentic copies of the creator's records.

The InterPARES 2 research (2002-2006) took instead the records creator's perspective [17-19, 30]. The researchers carried out case studies of records creation and maintenance in the artistic, scientific and governmental sectors; they modelled the many functions that make up records creation and maintenance and records preservation, according to both the lifecycle and the continuum models; they reviewed and compared legislation and government policies from a number of different countries and at different levels of government, from the national to the municipal; they analysed many metadata initiatives and developed a tool to identify the strengths and weaknesses of existing metadata schemas in relation to questions of reliability, accuracy and authenticity; and, once again, they studied the concept of trustworthiness and its components, reliability, accuracy.

The case studies showed that record creation in the digital environment is almost never guided by considerations of preservation over the long term. As a result, the reliability, accuracy and authenticity of digital records either cannot be established in the first place or cannot be demonstrated over periods of time relevant to the "business" requirements for the records. These records cannot therefore support the creator's accountability requirements, nor can they be effectively relied upon either by the creator for reference or later action or by external users as sources. Furthermore, they cannot be understood within an historical context, thereby undermining the traditional role of preserving organizations such as public archival institutions.

The research undertaken in records and information-related legislation showed that no level of government in any country to date has taken a comprehensive view of the records lifecycle, and that, in some cases, legislation has established significant barriers to the effective preservation of digital records over the long term, most notably that regarding copyright.

It was the responsibility of the InterPARES 2 Policy Cross-domain research team (hereinafter "the Policy team") to determine whether it was possible to establish a framework of principles that could guide the creation of policies, strategies and standards, and that would be flexible enough to be useful

[1] The synthesis we present here is based on the appendix 1 published as part of the project "Digital Records Pathways: From Creation and Maintenance to Preservation" developed by InterPARES and ICA Section on Archival Education (2011), to be printed.

in differing national environments, and consistent enough to be adopted in its entirety as a solid basis for any such document. In particular, such a framework had to balance different cultural, social and juridical perspectives on the issues of access to information, data privacy and intellectual property.

The findings of the InterPARES 1 research were confirmed by the research conducted by the InterPARES 2 Policy team, which further concluded that it is possible to develop such a framework of principles to support record creation, maintenance and preservation, regardless of jurisdiction. This document, in combination with other products of the Project, especially the Chain of Preservation (COP) model, reflects this conclusion, while emphasizing the need to make explicit the nature of the relationship between records creators and preservers.

The Policy team developed two complementary sets of principles, one for records creators and one for records preservers, which are intended to support the establishment of the relationship between creators and preservers by demonstrating the nature of that relationship. The principles for records creators are directed to the persons responsible for developing policies and strategies for the creation, maintenance and use of digital records within any kind of organization, and to national and international standards bodies. The principles for records preservers are directed to the persons responsible for developing policies and strategies for the long-term preservation of digital records within administrative units or institutions that have as their core mandate the preservation of the bodies of records created by persons, administrative units or organizations external to them, selected for permanent preservation under their jurisdiction for reasons of legal, administrative or historical accountability. They are therefore intended for administrative units (e.g., a bank, a city or a university archive) or institutions (e.g., community archives or state archives) with effective knowledge of records and records preservation.

2.1.2 CASPAR

CASPAR was a 42 months project (2006-2009) with the goal of implementing, extending, and validating the OAIS reference model (ISO:14721:2003) in preservation, access and retrieval for different environments (cultural, artistic, and scientific), including specific attention to the authenticity issues [21].

Being aware of the dynamic profile of authenticity and the need for specific tools and methodologies to deal with it, using the OAIS reference model as a base (together with the outputs of the InterPARES Project, particularly those concerning the concepts of identity and integrity), the CASPAR Authenticity Team identified a set of attributes that allow the capture of information relevant to authenticity that can be collected along the lifecycle of the digital objects; the Team has also developed the required tools and procedures to manage this information.

CASPAR conceptual model of authenticity is founded on the concept of Authenticity Protocol, a process which is designed to assess the authenticity of a resource, and can be applied to specific domains and specific components, as confirmed by the other partners' testbeds.

According to the *CASPAR Conceptual model* (CASPAR-D1201-TN-0101-1_0)[2], 2007, authenticity is a key concept in digital preservation (unless one can prove that the data object is what was originally deposited then one cannot prove that digital preservation has been successful). Moreover authenticity is never limited to the resource itself, but it is extended to the whole information / document / record system, thus to the concept of reliability, that is to the control over the information / document / record creation process and custody.

CASPAR focused on what InterPARES Project called the *maintenance* of authenticity, that is related to records which "have been presumed or verified authentic in the appraisal process, and have been transferred from the creator to the preserver".

[2] Available from http://www.alliancepermanentaccess.org/index.php/practices/member-resources/documents-and-downloads/?did=18

The project has identified tools (*Authenticity management tools)* by detailing the steps of the conceptual model [20-23] able to plan/design, describe and evaluate the chain of evidence about the custodianship and treatment of the information (relating to the OAIS Preservation Description Information - PDI).

2.1.3 Other related projects on digital preservation relevant for authenticity and provenance

In this section we briefly discuss other research projects, not specifically dedicated to the authenticity issues, which nevertheless include some results useful to delineate a more complete scenario. The information is taken from the projects' websites and from the final reports and mainly concern the outputs relevant to support authenticity evidence in the form here investigated. The projects are listed according to a chronological order.

2.1.3.1 PLANETS (2006-2010)

The primary and general goal for Planets is to build practical services and tools to ensure long-term access to digital cultural and scientific assets. The specific objectives relevant for supporting authenticity evidence concern the development of preservation planning services that empower organisations to define, evaluate, and execute preservation and the definition of methodologies, tools and services for the characterisation of digital objects. Specifically a Planets Core Registry (PCR) has been implemented as a technical registry that stores core records for file formats, software, hardware, compression techniques, character encodings and storage media along with associated subsidiary records and reference information. The PCR also stores information about characterisation tools, which identify and measure the properties of digital objects. In this way it supports the automatic deployment of appropriate characterisation tools, and thus the validation of preservation actions, by enabling the significant properties of source and target objects to be measured and compared, to ensure authenticity. In the case of migration preservation actions, the resultant information about the instance properties of the source and target components, and any variance in the properties after migration, can then be associated with a particular migration pathway and stored within the PCR. Such a verification forms the core component of the Planets planning tool Plato [6, 7], where it is used to ensure the authenticity of any object with regard to changes stemming from the application of a preservation action. This, in turn, forms the basis of a preservation plan, documenting the effect of any action on a digital object's significant properties. Similarly, evaluation experiments on specific properties can be conducted and documented in the Planets Testbed. A tool and the related methodology developed for extracting significant properties can be fruitfully implemented as part of the process for documenting the nature of the digital resources. A combination of these tools may help in documenting an object's trail through specific events such as migrations.

2.1.3.2 InSPECT - Investigating the Significant Properties of Electronic Content Over Time (2007-2009)

The project investigates the concept of significant properties, determines which properties are relevant for specific types of object, assesses their importance for future representation, and finally proposes a general methodology that enables digital curators to determine the significant properties of classes of digital objects that must be preserved over time.

The project has analysed as distinct classes audio, email, raster image and structured text objects. It has concentrated its attention to the formats migration and to the metadata extraction with a high level of granularity in identifying the characteristics to be controlled and evaluated in case of migration. The research outputs concern the accessibility and the capacity of interpreting the digital resources behaviour when used by a different community of users.

2.1.3.3 PROTAGE - Preservation Organizations Using Tools in AGent Environments (2007-2010)

PROTAGE has addressed the challenges related to the preservation of digital resources of increasing volume and heterogeneity by developing tools allowing for more efficiency and self-reliance of preservation processes. For this purpose, PROTAGE researchers are exploring the value of a promising technology - software agents - for the automation of digital preservation processes. Based on the latest research on digital preservation strategies and on autonomous systems, the project intends to build and validate flexible and extensible software agents for long-term digital preservation and access that can cooperate with and be integrated in existing and new preservation systems to support various aspects of the digital preservation workflow such as the submission / ingestion of digital material, monitoring of preservation systems and transfer between repositories.

2.1.3.4 SHAMAN-Sustaining Heritage Access through Multivalent ArchiviNg (2007-2011)

The aim of SHAMAN was to develop a long-term next generation digital preservation framework and develop new solutions for analysing, ingesting, managing, accessing and reusing information objects and data in the librarian and archival sectors.

The essential goal of the project is to establish an open distributed resource management infrastructure framework enabling GRID-based resource integration, reflecting, refining and extending the OAIS model and taking advantage of the latest state of art in virtualisation and distribution technologies from the fields of GRID computing, federated digital libraries and persistent archives.

Three prototype application solutions have been implemented as a multilayer model in the domains of scientific publishing, parliamentary archives, industrial design and engineering applications, with an abstract representation that is independent of the implementation. In particular software has been created for capturing digital objects representations and their related workflows and maintain them in an abstract form which is implementation independent with the aim of preserving them for reuse in future unknown infrastructures. This specific goal can provide useful tools for sustaining authenticity evidence as collected at the ingestion phase. A second output related to the validation process based on the verification of the original bitstream in case of replication has specific value with reference to the dissemination action and when the preservation does not imply format migration.

2.1.3.5 PARSE.Insight - Permanent Access to the Records of Science in Europe (2008-2010)

The aim of the PARSE.Insight project has been to define a roadmap and recommendations for developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information.

The *Roadmap* (par. 8.3 *Authenticity of digital objects*, p. 26) recognizes the crucial role of standardized evidence for provenance and its availability for users when the digital objects trustworthiness has to be proved.

The relevant steps and tools identified by the project concern: the development of an authenticity formalism, the identification of international standards and common policies on authenticity and provenance; the creation of tools able to capture authenticity evidence; the maintenance of the chain of evidence through (automated) digital audit (provenance) trails by embedding support for capturing knowledge about the actual operations performed.

2.1.3.6 LiWA - Living Web Archive (2008-2011)

Aim of LiWA project (funded by EU) was to develop web archiving tools able to capture content from a wide variety of sources. With reference to the issues related to the authenticity evidence, specific attention is dedicated: to improve archive fidelity by capturing complete and authenticated version of web content thanks to the implementation of current tools; detecting and filtering out web spam and

traps that generate automatically fake content in archives and ensuring long term interpretability of web content by keeping track of the evolving terminology.

2.1.3.7 KEEP - Keeping Emulation Environments Portable (2009 - 2011)

The project aimed to develop an emulation access platform to enable accurate rendering of both static and dynamic digital objects: text, sound, and image files; multimedia documents, websites, databases, videogames etc. The overall aim of the project was to facilitate universal access to cultural heritage by developing flexible tools for accessing and storing a wide range of digital objects.

KEEP had a strong relation to other European projects relevant for planning and documenting the authenticity evidence, specifically Planets (with reference to its effort for developing a permanent access framework able to perform preservation planning, characterization and direct preservation actions like migration and emulation) and SHAMAN (which had "a similar approach of working with more natural systems utilizing abstract representation mechanisms").

A sub-contracted legal study was carried out to clarify the legal framework within which media transfer operations and emulation software currently operate. The study considered European Community legislation as well as the national laws of France, Germany and the Netherlands.

2.1.3.8 PersID - Building a persistent identifier infrastructure (2009-2011)

Aim of the project was to provide unambiguous identification of digital objects through a persistent identifier as a part of a long-term responsibility for preserving digital materials. According to the project the persistent identification of digital objects plays a crucial role in the life cycle approach in the cultural and scientific digital library and archival applications and for the scholarly community. It provides an important contribution to increase the trustworthiness and the reliability of the whole chain of preservation.

2.1.3.9 PrestoPRIME - Keeping audiovisual contents alive (2009-2012)

PrestoPRIME project aims to research and develop practical solutions for the long-term preservation of digital media objects, programmes and collections, and to find ways to increase access by integrating the media archives with European on-line digital libraries in a digital preservation framework.

The project adopts the OAIS model and provides an Audio Visual Data Model as a specialisation of the PREMIS-based digital preservation approach able to trace all the relevant preservation events (i.e. actions that have affected the objects structure or data/information about the objects' authenticity) and collect the related metadata. Special attention is dedicated also to the so-called "pre-deposit provenance events", created prior the ingestion by the repository and identified by the Audio Visual Data Model as elements to be captured and store as part of the AIP as provenance and authenticity information.

2.1.3.10 Wf4Ever (2010-2013)

Wf4Ever aims at providing the methods and tools required to ensure the long-term preservation of scientific workflows in order to support the scientific discovery process and the development of new scientific assets. The research project is intended to develop new models, techniques and tools for the preservation of scientific workflows. It includes the definition of the "Research Object" and the description of packages workflow. In its first phase it presents initial requirements for workflow integrity and authenticity maintenance and evaluation identified through a systematic analysis of the users' needs. To support the authenticity and integrity in the digital environment, the project focuses its attention to the provenance information (intended as a crucial part of authenticity evidence) and to the technical features required for its management and use over time. The definition of provenance accepted by the project mainly concerns "the origin information about a resource and the process that led to the specific state of that resource".

A special effort is dedicated to analyse the existing models to represent the provenance information as provenance ontologies and to extract a set of core terms to describe the related elements in a neutral standardized form (i.e. resource, process execution, agent, location, etc.). The research takes into consideration also the concept of *meta-provenance* (as expressed by the scientific community), that is "another group of provenance-related information that has been drawing growing attention in the provenance community, which provide annotation-like meta-statement about a provenance statement or a set of provenance statements". This type of *meta-provenance* provides an extra level of contextual information about provenance statements, e.g. describing who provided some provenance statements, when, under what circumstances. This extra contextual information is particularly useful when we want to verify the integrity and authenticity of research results based on provenance information.
The user-led methodology developed by the project is relevant for the present report: the identification of the requirements for provenance and meta-provenance information is based on the analysis of the concrete application domains like astronomy and bio-informatics.

2.1.3.11 SCAPE - SCAlable Preservation Environments (2011-2014)

This project intends to develop scalable services for planning and execution of institutional preservation strategies on an open source platform that orchestrates semi-automated workflows for large-scale, heterogeneous collections of complex digital objects. These services will be able to: identify the need for preservation within a repository and the role of the characterisation processes and the trend analysis; define responses to those needs using formal descriptions of preservation policies and preservation plans; allow a high degree of automation, virtualisation of tools, and scalable processing; monitor the quality of preservation processes.

2.1.3.12 TIMBUS (2011-2014)

Aims of the project is to analyse the preservation of digital objects as well as the related business processes supported by Software as a Service (SaaS) and Internet of Services (IoS) within which data is processed, analysed, transformed and rendered. Even if specifically concerned with the problems of continued accessibility, the project focus is dedicated to contextual information and workflows as elements required to validate digital information when used by consumers when the original environment will not be in place. The dependencies on third-party services and on the richness of available information are going to play a crucial role and have to be carefully controlled in the life cycle. From this perspective, the project confirms the basic assumptions of this deliverable that the business process in place when the resources are created is as important as the whole preservation function from the submission to the ingestion and dissemination activities.

2.1.3.13 ENSURE (2011-2014)

ENSURE is researching how current lifecycle management tools can be used to control the preservation lifecycle with specific attention to the use of emerging ICT technologies for solutions which are not only economical, but also capable of scaling over time to meet ever expanding amounts of data. Cloud storage is seen as a primary candidate for the underlying storage services. The project intends to analyse the related additional challenges connected, e.g., the migration of data from cloud to cloud, security issues, and the ability to perform preservation-related computing near the storage. Three use cases will be considered: healthcare; clinical trials and financial services.

2.1.3.14 SCIDIP-ES – SCIence Data Infrastructure for Preservation – with a focus on Earth Science (EU funding 2011-2014)

SCIDIP-ES is an EU e-Infrastructures initiative which aims to put in place long lasting services which will enhance the ability of archives to ensure the long term usability of their digital holdings. The services are based on those developed in CASPAR. A number of toolkits help in the creation of the metadata which the services use. Of particular importance for this work package is the Authenticity Toolkit which will carry forward a number of the ideas developed here. The initial "critical mass" user

community for SCIDIP-ES will be those connected with Earth Science, but other communities and disciplines will also be worked with in the course of the funding period and beyond through the APA.

The APA will act as what is termed the exchange/guarantor node of the e-infrastructure components, which will fit well with the Virtual Centre of Excellence which we will be developing in APARSEN.

2.2 STANDARDS, REQUIREMENTS AND RECOMMENDATIONS

2.2.1 The role of recommendations, standards and guidelines for the creation and keeping of digital objects

2.2.1.1 Introductory remarks

The standards, recommendations or guidelines we have considered in this chapter (apart from ISO 14721:2003 or later – OAIS [31] - which is the basis of our common understanding in building an open framework for digital preservation and for this reason implicitly assumed in this report as its essential reference) are generally dedicated to the creation and keeping of accurate, complete and reliable records in the e-government environment. Even if intended for a specific domain, these rules are relevant for the preservation of any type of objects, if their evidential value has to be considered for building a detailed framework (and this is the basis for the identification of the workflow model as presented in section 4).

The attention to the assessment of the evidential value of the digital object and its persistency is crucial in the e-government environment, due to the legal and administrative function of the electronic records [58, 59], and the need for the continuing capacity to prove their legal authenticity. For this reason the standards in the recordkeeping sector pay attention to these aspects and therefore play a relevant role in our analysis. We also consider the recommendations for digital transfer developed by the International Council on Archives and implemented by the United Nations, which are specifically intended for documents/records, but can be easily extended to any type of digital objects. A final sector is dedicated to the metadata for preservation [25, 49], mainly to PREMIS standard [51, 52], used in preservation processes and with some basis on OAIS.

The principle of the records lifecycle or of the records continuum (as named in the Australian recent tradition) is based on the capacity to define – at the right degree of detail – the responsibilities, the actions, the events and the representation information to be handled in each phase of the digital records management, and to maintain in each transfer (to another recordkeeping system or to the preserver) appropriate documentation of all the main processes involved and of all required metadata.

The level of granularity required is still an open question, in particular with reference to what has to be documented in term of events and responsibilities, and with reference to the descriptive information that is required. To answer these issues, the concept of significant properties [28, 42], introduced by scholars and developed in international projects, can play an important role if extensively interpreted. In particular reference [22] details the limitations of previous work and applicability of the ill-defined nature of the term “significant properties” and the revised version of OAIS introduced the more precisely defined and more widely applicable “Transformational Information Property” of which some definitions of significant properties can be regarded as a sub-set which may be adequate for rendered digital objects such as simple documents and images. Given this proviso, the concept has been detailed by the recommendations and standards on e-government, as a well-defined set of properties, attributes and representation information to be taken into account to provide authenticity evidence for electronic records. Nevertheless, as already pointed out, the specifications that we are considering can be applied to any kind of other digital objects, and for these reasons they are proposed here as a crucial part of a methodological approach (see PARSE.Insight project in section 2.1.3.4).

It is of course impossible to consider in detail all these regulations, so we will concentrate on the most important ones, those issued by ISO, specifically by TC 46 SC 11 Archives/Records Management. Specifically the standard ISO RM 15489-1: 2001 *Information and Documentation – Records Management. Part 1: General* [32] and the standards ISO 23081-1:2006 *Information and*

Documentation – Records Management Processes – Metadata for Records, ISO 23081-1:2009 *Information and Documentation –Managing Metadata for Records – Part 2: Conceptual and Implementation Issues* and , ISO 23081-1:2011 *Information and Documentation –Managing Metadata for Records – Part 3: Self-Assessment Method* [33-35], and those issued by the international and multinational institutions, like the International Council on Archives and the DLM Forum (MoReq2 and MoReq 2010), with the aim of providing specifications for building electronic records management systems [46, 47]. The recommendation for digital transfer [64] has been also considered for its relevance in carefully documenting crucial aspects of the digital object lifecycle (UN/CEFACT: *Business Requirements Specificatio:. Transfer of Digital Records*, Version 1.0 (2008).

2.2.1.2 ISO 15489-1: 2001 Information and documentation – Records management

The ISO 15489 [32] is significant for our purpose because it provides the list of actions and responsibilities relevant for the accurate and reliable records creation and keeping. With reference to the authenticity and its evidence, accuracy and reliability are based on the complete information and documentation of all the activities involving changes on digital contents and contextual information. The capacity to create and maintain records, which can be presumed authentic, is considered one of the main goals of a records management system.

The list of the actions to be qualified and tracked includes:

- to *capture* and *create*: the creation of the digital object should be able to define persistent and verifiable relationships between a record, its author and all the relevant contexts (legal, administrative, functional)[3];
- to *register*: the digital object has to be persistently and uniquely identified at the moment of its creation or at least at the time of its inclusion in the documentary/archival system with a double aim of proving for evidential reasons its existence at a specific and documented time and for its retrieval; the authenticity evidence is based on the neutrality of this action, that is on the implementation of automated procedures;
- to *classify*: the digital object is not relevant as a single entity, but as part of a collection/set of entities; the connections have to be determined and maintained with the aim of being retrieved for access and for evaluation, but also for building control on the security levels, for establishing management and keeping responsibilities,
- to *store*: since their formal creation the digital objects have to be safely kept and protected against non-authorized access, loss, destruction; this action implies the capacity to copy and/or migrate without losing the level of reliability, usability, accessibility and *authenticity*; each change should be tracked;
- to *make accessible* and *secure*: a specific policy should ensure the definition of access rights and the controls against any abuse with reference to system security, freedom of information, privacy and reference rules; the policy should be periodically evaluated;
- to *track* the use and the transfer: these controls concern both the internal and the external use and transfer, to identify the actions on the digital objects, to allow retrieval, to prevent loss and to ensure proper documentation of relevant actions, but also to identify the functional provenance for each object in case of system migration or incorporation;
- to *implement disposition*: a systematic and periodical approach should be taken to define a retention plan, with the aim of correctly identifying the terms of selection, preservation and destruction.

We will consider most of these actions when analysing the digital object lifecycle in section 4.

[3] In fact, the term ‘to capture’ as explained in the standard includes actions like ‘to register’ and ‘to classify’, which are separately indicated by the ISO standard.

2.2.1.3 ISO 23081-1:2006 Information and documentation – Records management processes – Metadata for records

The family of ISO standards 23081 [33-35] is dedicated to the control of the metadata relevant for the records management processes, but it could be applied to any type of digital objects. Specifically, the metadata identified by the standard concern the records identity, their validation (with reference to the records themselves and the processes, actors and systems involved) and their contextualization. The standard declares the relevance of an accurate management of metadata in the digital objects lifecycle, especially in the creation and keeping phases, for protecting the evidential value of the records/digital objects, for evaluating their authenticity[4].

The specific areas implemented by the standard (not in the form of schemas whose nature depends upon the environment and the designated community, but as a guide for implementation and use) include:

- the digital objects (i.e. the persistent identifier, the reference code, the date),
- their relationships (i.e. the aggregations),
- their connections with processes and events (i.e. workflows),
- the access limitations,
- the digital objects description for further handling, keeping and preserving.

The framework is compliant with the standard ISO 15489 and the connected actions (to capture and create, to keep, to store, to make accessible) and identify five categories which could be considered before or after record capture. These categories include:

" a) metadata about the record itself;
b) metadata about the business rules or policies and mandates;
c) metadata about agents;
d) metadata about business activities or processes;
e) metadata about records management processes"

The requirements related to the authenticity and fixity of metadata themselves (8.3.9.2) are crucial with reference to the specific purpose of this deliverable, as clearly stated by the standard:

"Records management metadata are as much subject to authenticity rules or criteria as the records to which they are linked in order to make them trustworthy. Agents should therefore document all policies and rules relating to metadata and developments therein. Changes in structures for metadata, either conceptual or physical, should also be documented. An important element for ensuring authenticity of metadata and proper metadata management over time is the requirement that captured

[4] "Metadata support business and records management processes by:
a) protecting records as *evidence* and ensuring their accessibility, and usability through time,
b) facilitating the ability to understand records,
c) *supporting and ensuring the evidential value of records*,
d) *helping to ensure the authenticity*, reliability and integrity of records,
e) supporting and managing access, privacy and rights,
f) supporting efficient retrieval,
g) supporting interoperability strategies by enabling *authoritative capture of records* created in diverse technical and business environments and their sustainability for as long as required,
h) providing logical links between records and the context of their creation, and maintaining them in a structured, reliable and meaningful way,
i) supporting the identification of the technological environment in which digital records were created or captured, and the management of the technological environment in which they are maintained in order that *authentic records can be reproduced as long as they are needed*, and
j) supporting efficient and successful migration of records from one environment or computer platform to another or any other preservation strategy" [*the italic has been added by the authors of this report*].

metadata are fixed. Records management metadata need to be maintained as they are and, in case change is needed, rules should be in place to govern the process. These should include rules to document the reasons for the changes, the changes themselves, and the authorized agents involved. These requirements apply over time and to any organization responsible for the records involved. Metadata providing details about the creation of, or change to, the metadata record itself should be maintained. This should include information about any agents associated with the creation or change and the type of activity that was undertaken, for example: created, modified, checked, deleted. In addition, the version of the metadata schema used to define and populate the metadata elements should be identified".

With reference to the present deliverable, all the requirements have to be considered to develop a robust methodology, even if this standard has not the aim of defining detailed schemas. A granular list can (has to) be deployed only within each identified environment and specific domain. The model we describe in section 4 is compliant with the main categories of the standard ISO 23081 which have been identified for their relevance in assessing the authenticity or providing content to its evidence.

2.2.1.4 Recommendations and guidelines for the functional requirements for ERMS

The recommendations and guidelines for the definition of functional requirements for planning and implementing electronic records management systems have been developed with specific attention to the application analysis since 1995 (as testified by many national rules and standards approved in Australia and North America). It is impossible to analyse all of them. We will concentrate on the European guidelines Model Requirements for Electronic Records Management Systems, whose three versions have been approved in 2001 (*MoReq1*), 2008 (*MoReq2*) and 2010 (*Moreq2010*) [46, 47]. The last version is more an integration of MoReq2 for private domains or for small administrations than a new stand-alone version, even if special and new attention has been dedicated to the definition of data modelling for sustaining exporting functions. This family of recommendations can play a crucial role in implementing methods and translating them into detailed lists of measures for assessment and for this reason is here examined in some details.

The MoReq specification has the ambition to make a complete model for an electronic records management system available planned as a consistent set of principles, methods and processes. The guidelines are relevant also for the richness of their details and for the goal of implementing an operational certification environment.

Specific attention is dedicated to all the actions and actors relevant for the records creation and capture, access, use and selection. The requirements for controls and security of the systems are very detailed, specifically in the last version of MoReq2010 whose aim of building a data model is explicitly consistent with the need – relevant for digital preservation – of ensuring interoperability among systems and export functionality.

The MoReq2010 specification dedicates attention to the integrity and to the authenticity as a general goal of an ERMS, but it does not provide a specific model related workflows nor does it define responsibility for explicitly supporting authenticity evidence. The reason to this can be found in the main focus of the specification: to build a data model for interoperable record systems. The authenticity implication of this effort is implicit and it is related to the detailed control list of relevant events which track the record system functions and are able to impact on the integrity of the record system itself and on the provenance of each record or its related component.

To provide elements and metadata compliant with ISO standards for record management (ISO 15489 and ISO 23081) the specification has developed a very specific information model (chapter 14) organized in five categories: entity types (aggregation, class, component, contextual metadata element definition, disposal hold, disposal schedule, entity type, event, function definition, group, metadata element definition, record, role, service, template, user), data structures (access control entry, access control list, metadata change entry), system metadata element definition (107 elements are specified), function definitions (196 functions are identified and described with reference to the aggregation, classification, identification of components and of context, disposal, inspection, roles management, records handling). The information model is limited to define the records system functionality.

The specification pays particular attention to the need for event history, made up of a series of events that have occurred to that entity and require to be carefully selected on the basis of their relevance. (2.2.8 and 6.2.8): to be self-contained or atomic (for supporting interoperability and allowing the transfer of entities between different records systems) it is stressed that a record's event history has to be established to confirm/document which key events have occurred since the record was created.

For exporting records the specification points out the crucial role of :

- system metadata,
- contextual metadata,
- entities referred to by system identifiers,
- entities significant for system management (like records disposal schedule),
- the access control list and the users roles,
- the events in the entity's event history and their related metadata.

A chain of timestamps and persistent system identifier in the form of UUID (universal unique identifier) is also considered crucial to provide integrity and interoperability.

In conclusion the specification provides a detailed list of elements to be considered for an integrated approach to the preservation even if the model is only focused on the record creation and no mention is made to the OAIS architecture.

2.2.1.5 The recommendations for the transfer of digital objects: UN/CEFACT, Business Requirements Specification. Transfer of Digital Records, Version 1.0 (2008)

This recommendation approved by the International Council on Archives and implemented in 2008 by the United Nations as UN/CEFACT *Business Requirements Specification – BRS. Transfer of Digital Records* [64] describes the transfer of custody of digital objects (specifically, records) from one keeping or preserving system to another. It concerns the formal transfer of responsibilities which can happen within the same organization or between different entities.

The rules have the aim of reducing the risks of loss involved in the transfer and of limiting the overall costs, but also of allowing the re-use of and the correct access to the digital objects transferred to other organizational and technical contexts. They are based on the approval of a transfer agreement which includes (according to an OAIS model) the specifications related to the digital objects to be transferred, the terms for the transfer, the access conditions and the standards to be applied for identifying the digital objects, their structure and their contexts.

The main focus is on the localization of the transferred resources and the most relevant part concerns the definition of the information related to the transfer operation itself as managed through *transfer sessions.* A special attention is dedicated to the accuracy and the completeness of the transfer documentation, to the analysis of the mechanisms for securing the transfer and verifying its quality and correctness mainly in term of responsibilities and controls.

It includes the description of the following scenarios: Proposal/Manifest agreement, Reject Transfer Session, Transfer, Signal Transfer Status, Finalise Transfer Session.

The recommendation has been approved by Archives New Zealand, Bibliotheque Nationale de France, Bundesarchiv, Direction des Archives de France, National Archives of Australia, National Archives of United Kingdom, Public Record Office of Victoria, Rahvurarhiiv Estonia, Riksarkivet in Stockholm, University of Michigan Bentley Historical Library, US National Archives and Records Administration.

The deliverable has considered this recommendation in building the model framework in section 4.

2.2.1.6 Metadata for digital preservation: the role of PREMIS

PREMIS. Preservation Metadata Implementation Strategies[5] [25, 51, 52] is a US Library of Congress standard approved in 2008 with the aim of providing a set of core elements, easily extensible, for the preservation of digital objects. It is based on the OAIS standard and has been developed as an XML schema and a data dictionary. The preservation metadata identified by the standard include the technical and the administrative information that a digital repository manages for ensuring the digital preservation process with specific attention to those characteristics which support – among others – the authenticity and the identity of digital objects in a preservation context. A special attention is dedicated to standardize the documentation related to provenance, fixity and digital object structure. The PREMIS data model is intended for clearly defining the meaning of each element or semantic unit, but not for providing a reference architectural model. The units are organized in four entity types: object, actor, event and right. Tracing actors and events in the preservation environment is crucial for authenticity, as clearly expressed in the introduction to the standard:

"The Event entity aggregates metadata about actions. A preservation repository will record events for many reasons. Documentation of actions that modify (that is, create a new version of) a digital object is critical to maintaining digital provenance, a key element of authenticity. Actions that create new relationships or alter existing relationships are important in explaining those relationships. Even actions that alter nothing, such as validity and integrity checks on objects, can be important to record for management purposes."

The PREMIS working group emphasizes the relevance of fixity, integrity and authenticity by stressing that "objects that lack these features are of little value to repositories that have the mission to protect evidentiary value or indeed to preserve the cultural memory".

2.2.2 The role of recommendations and standards for the certification of digital repositories

The standards aimed at defining rules for evaluating and measuring the quality of the preservation are crucial for the success of the authenticity assessment. The effort for developing a normalized process for certifying the quality of the repositories has taken more than a decade [4, 29, 56]. The final outputs, based on previous guidelines and checklist [14, 54, 55], are the standards ISO/DIS 16363 *Space Data and Information Transfer Systems – Requirements for Audit and Certification of Trustworthy Digital Repositories*) and 16919 (ISO/DIS 16919 *Space Data and Information Transfer Systems - Requirements for Bodies Providing Audit And Certification Of Candidate Trustworthy Digital Repositories*) [36, 37] produced by the Mission Operations and Information Management Services Area (MOIMS) of the Consultative Committee for Space Data Systems (CCSDS), the same body responsible for the creation of the OAIS model. The common bases for these two standards and for the other rules in the certification area under approval by other standard organizations are: ISO/IEC 17021: 2006, *Conformity assessment – Requirements for bodies providing audit and certification of management systems* [38], RLG/OCLC Working Group on Digital Archival Attributes, *Trusted Digital Repositories. Attributes and Responsibilities*, 2002 [55] and RLG-NARA Task force on digital repository certification: *Audit Checklist for Certifying Digital Repositories*, 2004 [54].

These three standards (ISO/IEC 16363, 16919 and 17021) provide a detailed and well-articulated framework for assessing how a repository performs its preservation function. In particular it specifies:

- rules and guidelines for certifying digital repositories also in the form of self-evaluation,
- a technical basis for developing operational tools to measure quality and capacity in the field,
- guidelines to auditors.

ISO 16363 [36] does not specify explicitly how the evaluation of the authenticity is done or how the developments of steps and workflows for ensuring different degrees of authenticity evidence are carried out. As has clearly emerged from the test audits (see APARSEN deliverable D33.1B) – the

[5] The first version has been released in 2005.

definition of metrics and the list of well documented tasks and requirements, as expressed by the standard, can provide a relevant support both to the repositories and to the auditors because it can identify the areas and the phases to be strictly documented for providing elements to evaluate the authenticity and the integrity (and supporting the quality) for the audit function.

A good example is the definition of *clear responsibilities* as expressed by the requirements expressed under 3.1. of the standard ISO 16363: 3.1.1. Mission statement, 3.1.2. Strategic plan or 3.3.3. Documented history of changes to its operations, procedures, technologies.

A similar role is played by the requirements related to

- the *integrity control and accountability* (like 3.3.5. Documentation of the repository integrity measures, 3.3.6. Regular schedule for self-assessment and external auditing, 3.4.1. Transparency of the financial practices, 3.5.1. appropriate contracts or deposit agreements for preserved digital materials, 3.5.2. track of restrictions on use of digital contents);
- the *control on the content acquisition* (4.1.1. identification of the content information and its property, 4..1.2. information associated to the SIP at the time of ingestion, 4.1.4. mechanisms to appropriately verify the depositor of all materials, 4.1.5. ingest process which verifies each sip for completeness and correctness, 4.1.6. sufficient control over the digital objects to preserve them at bit level at the moment of transfer, 4.1.8. contemporaneous records of actions and administration processes that are relevant to content acquisition);
- the *control on the AIP preserved* (4.2.1. an associated definition for each AIP or class of AIPs preserved by the repository, 4.2.2. a description of how AIPs are constructed from SIPs, 4.2.3. documentation of the final disposition of all SIPs, 4.2.4. a convention for the generation of persistent, unique identifiers for all AIPs, 4.2.6. documented processes for acquiring preservation description information (PDI) for its associated content information, 4.2.8. verification of completeness and correctness of each AIP at the point it is created, 4.2.9. independent mechanism for verifying the integrity of the repository collections/content, 4.2.10. contemporaneous records of actions and administration processes that are relevant to AIP creation);
- the *control on the preservation* **(**4.3.1. documentation of the preservation planning, 4.3.2. documentation of the monitoring mechanisms on the preservation environment, 4.4.1. specifications for how the AIPs are stored down to the bit level, 4.4.2. contemporaneous records of actions and administration processes that are relevant to storage and preservation of the AIPs, 4.5.2. minimum descriptive information and ensure that it is associated with the AIP, 4.6.2. policies and procedures that enable the dissemination of digital objects that are traceable to the originals, with evidence supporting their authenticity);
- the *control on the risk management* (5.1.1. identification and management of the risks to the preservation operations and goals associated with system infrastructure, 5.1.2. management of the number and location of copies of all digital objects, 5.1.3. Delineated Roles, Responsibilities, And Authorizations Related To Implementing Changes Within The System, 5.1.4. suitable written disaster preparedness and recovery plans, including at least one off-site backup of all preserved information together with an offsite copy of the recovery plans).

We plan indeed to consider these elements as part of a specific metric, in a further work when analysing case studies, later in the project. The results will be reported in Deliverable D24.2.

3 A BASIC FRAMEWORK FOR AUTHENTICITY AND PROVENANCE

The term "authenticity" is adopted in different disciplines. Even if the general meaning is common, the tools and the methods employed are specific to the disciplinary domains. In this report the term will be explored with reference to the juridical sector, to the documentary/archival environment (*archival diplomatics*, according to Luciana Duranti) and to the historical point of view. In all these areas, and specifically from the point of view of the objects preservation, the term is closely related to the presence of crucial requirements and controlled processes able to ensure their integrity, their trustworthiness and their reliability during the creation phase, their migration and the production of authoritative copies.

From the legal point of view, authenticity consists of the capacity of proving the imputability of a digital object (generally a record) to the specific person responsible for its creation. In a context of civil law, an object/record is considered *authentic* when its author (that is its provenance or *origin*) is undoubtedly recognized. In synthesis according to the legal semantics the concept of authenticity is related to the certainty of the record's provenance (G. Belli, *Autenticazione*, Novissimo digesto italiano, 1957).

With reference to the archival diplomatics three different concepts have been developed:

- *genuineness*: the authenticity from the diplomatics point of view; a record is genuine if it is created as part of regular procedures and its logical articulation and its configuration are compliant with the prescribed (and verifiable) procedures;
- *authenticity strictu sensu* or legal authenticity: a record is authentic if it has the required forms (conceptual organization and configuration) and validation elements necessary to provide full faith and evidence to the content;
- *veracity* or the historic authenticity: a record is veracious if the represented facts are consistent with the reality.

Specifically, according to the archival diplomatics, an authentic record is a record that is what it purports to be. According to Black's Law Dictionary (1968) "this term (i.e. *genuine*) means that they (i.e. the written instruments) are truly what they purport to be, and that they are not false, forged, fictious, simulated, spurious or counterfeit". In 1931 Harry Bresslau [9] specified that diplomatics must strictly control the processes of transmission of a document, analysing the intermediate events through which authentic copies of documents/records are generated, with the aim to verify that those procedures have not affected their genuineness.

This goal is still valid in a digital environment and is not limited to the records but can (has to) be applied to the whole complex of digital objects. Moreover, the activity of examining and tracking events that could attest the genuineness (i.e. the authenticity in the diplomatics sense of the term) is an essential component in the process of evaluating the legal evidence of a digital document/record: in the *common law* juridical system the integrity of a document is strictly related to its *unbroken custody* in a trusted repository, as necessary condition for the admissibility of the document as evidence in legal proceedings.

The basic framework for authenticity and provenance presented here has been developed by considering the outcomes of the research projects previously analysed (2.1) and by developing the concepts of the archival diplomatics. This framework will be used as a starting point for modelling standardized workflows for preservation (3.2) and for providing principles, concepts and a methodological approach to the assessment of the authenticity evidence at a more systematic and operational level.

The main assumptions are summarized according to the reference project here considered:

- **OAIS** [31] is the reference model to be implemented with specific reference to the definition of the common and well recognized architecture required to manage workflows of information concerning the responsibilities for digital preservation, the representation information and the PDI;

- **InterPARES** [17-19, 30] constitutes the main conceptual framework related to the authenticity (according to the distinction described in the introduction): it includes the interrelating principles and methods to compare and assess quality and consistency of the digital practices for authenticity. InterPARES makes a clear proposal about the assessment of authenticity, when the records are transferred to archival custody, and about maintenance of authenticity, when it is necessary to produce authentic copies as part of the custodial requirements (i.e. in case of migration).
- **CASPAR** [20-24] provides the methodological approach for the implementation of a standardized set of tools able to integrate and document the main events and information related to the preservation function, specifically with the aim of functionally supporting authenticity evidence. As clearly stated by the CASPAR position paper, this analysis is built on the conclusion that authenticity cannot be handled as a static quality of the object, but as a complex process that requires the creation and the preservation of well-structured documentation, a conceptual and systematic model for handling the events and their flows and the notion of authenticity protocol aimed at representing the procedures to be followed to assess the authenticity of objects.
- **ISO 16363** and **ISO 16919** [36, 37] are included with specific reference to their identification of measures relevant for presuming authenticity and creating its evidence.

The (updated) OAIS revision defines “authenticity” as: "the degree to which a person (or system) may regard an object as what it is purported to be. The degree of authenticity is judged on the basis of evidence". Note that, as stated in the introduction, in principle authenticity does not have degrees: it is a binary characteristic, an attribute of a digital object that could just be authentic or not according to the manifestation of its author (provenance) and to the preservation across the time of some elements and attributes that contribute to prove its provenance (context of creation and management over time). What OAIS refers to is the degree of certainty a person has in his/her judgment of that binary evaluation. This may be termed the *presumption of the authenticity*, the evidence for which is based upon a set of information, and of a consistent and reliable documentation for each event that has occurred to the digital object in the course of its life. Those information support the presumption of authenticity, can be assessed and the assessment can be more or less supported by the preservation system.

According to InterPARES, the concept of *authenticity* is based on *identity* and *integrity* and both concepts are defined on the basis of complex and integrated requirements identified at any transfer of the digital objects from one responsibility to another (*benchmark requirements*) and in the archival repository (*baseline requirements*). OAIS refers the preservation (or the evidence) of the *authenticity* to the *whole process of preservation* and specifically identifies it as part of the Preservation Description Information (PDI). As a consequence, also the evaluation of the authenticity refers to the whole process of preservation.

The concepts of integrity and identity have to be further investigated, with the aim of identifying the actions types, the specific events and the representation information to be maintained over time for authenticity evidence.

According to InterPARES, the *integrity* of a resource refers to its wholeness. A resource has integrity when it is complete and uncorrupted in all its essential respects. The verification process should analyse and ascertain that they are consistent with the inevitable changes brought about by technological obsolescence. The original bit stream can be compromised, but the content structure and the essential components must remain the same. InterPARES refers the *identity* of a resource not only to its unique designation and/or identification. It refers to *the whole* of the characteristics of a resource that uniquely identify it and distinguish it from any other resource, i.e. it refers not only to its internal conceptual structure but also to its general context (administrative, legal, documentary, technological, some could even add social).

By translating these concepts into the OAIS framework, the crucial role of PDI – Preservation Description Information appears clearly for all the types of information involved (*Reference Information*: mechanisms used to provide assigned – internal and/or external – identifiers for the

Content Information; *Context Information*: the relationships of the Content Information to its environment i.e. why it was created, how it relates to other Content Information objects, etc.; *Provenance Information*: the history of the Content Information as the origin or source, any changes since it was originated, who has had custody of it, etc.; *Fixity Information*: data Integrity checks or validation/verification keys used to ensure that the particular Content Information object has not been altered in an undocumented manner. *Access Rights Information* has been added as part in PDI in the revision of OAIS.

On the basis of these considerations and principles, many other assumptions – relevant for the definition of the authenticity evidence and its assessment – can be made:

- It is not possible (feasible) to preserve electronic resources as original unchanged resources: we have only the ability to reproduce them in the form of *authentic copies* thanks to the preservation of authentic copies of digital components.
- Authenticity *cannot be recognized as given once and for ever* within a digital environment: a clear distinction should be made between the authenticity of the preserved record/resource (not necessarily the same objects as those originally deposited) and the procedure of *evaluating* and *validating* the same object.
- Not only is the digital preservation a dynamic process but also the profile of the authenticity has to be considered as a *process* aimed at gathering, protecting and/or evaluating information/set of attributes mainly about identity and integrity of the digital object, of its components and of the related data relevant for handling the content and packaging it.

The consequence of this reasoning – based on the fact that the digital objects curation is increasingly based on the concept of *trust* – brings to the centre of the future implementation the principle of trustworthiness. In the dictionary (Merriam-Webster, s.v.) *trust* is identified as "a charge or duty imposed in faith or confidence or as a condition of some relationship", a sort of "glue which binds that relationship together", whose ingredients have to be identified and described for effectiveness of the custody. The core concepts concern the creation of a multilayer approach able to verify the integrity and authenticity of the resources at various levels of analysis. Authenticity and integrity could be evaluated *as inference on the basis of the trustworthiness of the document/information system* in which the documents/information exist.

These concepts are supported by the analysis of the weaknesses of the mechanisms put in place in the market to develop digital validation of bit-stream finalized to ensure them evidential value overtime. A good example is providing by ADOBE analysis of the existing mechanism in this field and their efficacy. Adobe proposes many strategies for ensuring the persistency of the evidential value of the digital records:

- the incorporation of the validation controls within the object: this is considered a light solution;
- the incorporation of the date and time as part of the hash in the digital object creation: this is considered a temporary solution not practical for the long-term validation issues;
- the *archival validation***:** based on the definition of the authentication process and its controls for identity and integrity as part of the creation and keeping of the digital objects; the metadata (representation information and all the relevant data) are preserved as part of the management and preservation system able to be recorded and verified over time as far as the archival controls and systems last.

As a first conclusion – accepted by all the projects considered here and not yet contested – we can state that evidence for the assessment of authenticity has to consist both of *technical* and *non-technical* elements. The technical elements include controls on the integrity and can be defined as tools for validation like digital digests in the case that the bit sequences are expected to have been unchanged and/or the unchanged content can be attributed to an author: according to OAIS model the significant information can be provided by fixity but also by provenance information. If the bit sequences have been changed then the Transformational Information Properties and the recorded judgments of the custodians are the key pieces of information. The non-technical elements vary from the identity of the

author and set of custodians to the elements able to provide evidence of the reliability of the creation system and of the trustworthiness of the custodian.

Similarly, from the point of view of the representation information relevant for supporting the authenticity verification of digital objects, authenticity and integrity could be evaluated:

- on the basis of the information elements present on the *face/form* of the resource and its attributes /metadata as part of the creation process;
- from the circumstances carefully documented and tested through metadata related to its maintenance and preservation: "an unbroken chain of responsible and legitimate custody is considered an assurance of integrity until proof to the contrary";
- from the integrity of essential information related to the resources handling and preservation as a further requirement for attestation of integrity and authenticity:
  - individuals/offices involved in the relevant processes as producer/preserver/consumer,
  - indication of annotations, of technical changes, of presence or removal and their time of digital signature and other digital seals, the time of transfer to a trusted custodian, the time of planned deletion, the existence and location of duplicates outside the system, etc.

4 A MODEL FOR MANAGING AUTHENTICITY AND PROVENANCE THROUGH THE DIGITAL RESOURCE LIFECYCLE

4.1 THE DIGITAL RESOURCE LIFECYCLE

As we have already anticipated in the previous sections, in order to properly assess the authenticity and the provenance of a Digital Resource (DR)[6] we must be able to trace back, along the whole extent of its lifecycle since its creation, all the transformations the DR has undergone and that may have affected its authenticity and provenance. For each of these transformations one needs then to collect and preserve the appropriate evidence that would allow someone or something, at a later time, to make the assessment, and that we shall call therefore *authenticity evidence*.

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

- *Phase 1: pre-ingestion phase*

This phase begins when the DR is delivered for the first time to an intermediate system (referred to here as a *keeping system*) and goes on until the DR is submitted to a Long Time Digital Preservation (LTDP) system. During the pre-ingestion phase, the DR may possibly be transferred between several keeping systems and may undergo several transformations. For instance it may be aggregated with other digital resources, its content and metadata may be integrated with additional information or some of its components may be migrated to a different format etc. It is therefore of the utmost importance to understand how these transformations may affect the authenticity and the integrity of the DR, and to collect the proper evidence. This phase may in some cases be indistinguishable from Phase 2.

- *Phase 2: LTDP phase*

This phase begins when the DR is submitted to a LTDP system and goes on as long as the DR is preserved. As for the pre-ingestion phase, also during the LTDP phase the DR may undergo several transformations, as for instance format migrations, aggregations and disaggregations. Moreover it may be transferred from one LTDP system to another one. The main difference with respect to the keeping phase, when collecting authenticity evidence, is that we can make quite more precise assumptions about the internal organization of the LTDP systems, and therefore the evolution of the DR during this phase is therefore is more predictable and controllable.

Each *transformation* a DR undergoes during its lifecycle is connected to an *event*, which occurs at a precise time and 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.

A very ambitious goal would be to try to determine 'all' possible events that are relevant with regard to the authenticity of a DR, and to draw precise guidelines to specify which authenticity evidence should be collected for each of these events, and how to organize it. This would be indeed a very interesting result since, as we have seen, a DR moves along its lifecycle from system to system, and therefore these systems, when they exchange a DR, need to interoperate in order to exchange also the related authenticity evidence. Interoperability means agreeing on a common ground, and therefore common guidelines would form the basis that would allow such systems to interoperate.

Unfortunately, the variety of events that may occur during the pre-ingestion and the LTDP phases is very large and depends, at least in part, from the specific environment in which a DR is produced and managed. It may therefore not be reasonable to try to draw a comprehensive list of all possible events. Nevertheless, it is possible to consider at least a *core set* of events that includes the most important

[6] To avoid confusion, we decided to use, as a generic term, Digital Resource (DR) instead of Digital Object, a term which is given in the OAIS Reference Model a specific (and different) meaning. According to the OAIS terminology the appropriate term should be Content Information, which is used to designate "The set of information that is the original target of preservation". However, this term is specifically related to the structure of the Information Package (SIP and AIP), and hence using is, in our opinion, inappropriate when discussing the 'pre-ingestion' phase of the digital resource lifecycle.

ones, as well as the ones which are likely to occur in most of the environments in which DRs are produced and managed.

In the rest of this section we will try to sketch out such a core set of events for the pre-ingestion and the LTDP phases, to formalize the transformations a DR undergoes in connection with these events and to specify the evidence that must be collected.

More precisely, we will represent each event according to a uniform schema, where we shall specify:

- the *description*, i.e. the circumstances and the actions connected to the event and the transformations that are induced on the DRs that are involved;
- the *agent*, i.e. the person(s) under whose responsibility the transformation occurs;
- the *input*, i.e. the pre-existing DR(s) that are the object of the transformation;
- the *output*, i.e. the new DR(s) that are the result of the transformation (possibly new versions of input DR(s));
- the *Authenticity Evidence Record (AER)*, i.e. the set of information that must be gathered in connection with the event to support the tracking of its authenticity and provenance.

As a DR progresses along its lifecycle through a sequence of events, an incremental sequence of authenticity evidence records, 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 keeping and/or LTDP systems, and will accompany it along all its lifecycle.

When the DR enters the LTDP phase, the Authenticity Evidence History collected during the pre-ingestion phase provides crucial information to generate the Preservation Description Information (PDI), i.e. the component of the Information Package (SIP and later AIP) where is recorded the information which is necessary for adequate preservation of the DR, and which can be categorized as Provenance, Reference, Fixity, and Context information.

Later, during the LTDP phase, the authenticity evidence collected in conjunction with further transformations of the DR may lead to the generation of new AIPs in order to include the related updates in the PDI. Though not strictly necessary, it could be very useful to maintain also in the PDI the incremental structure that we have suggested for the AEH in the pre-ingestion phase, in order to record orderly the whole sequence of transformations that have affected the DR during its lifecycle.

The authenticity evidence records need therefore to be managed and interpreted by systems (both keeping and LTDP systems) which may be different from the ones that gathered them. Thus, to ensure interoperability, their content and their structure should comply with a common standard.

Achieving such a standard is a quite an ambitious goal, and requires, of course, time, consensus and a thorough discussion. The model we propose here should therefore be considered only as first step in this direction, and a starting point. From a practical point of view, we believe that it may be a sound basis to derive operational guidelines to improve in a significant way the current (and often very limited) practices in managing authenticity and provenance in keeping and preservation systems.

4.2 PHASE 1: PRE-INGESTION

The pre-ingestion phase begins when the DR is *created* by its *author* and terminates when the object is transferred to a LTDP system.

The *author* of a DR is the person who, individually or as the representative of an institution, takes the responsibility of the content of the DR and of the descriptive information associated to it when the DR is *created*, i.e. delivered for the first time to a *keeping system*, a term by which we mean any kind of system where the DR is kept, once it has been created, until it enters the LTDP phase.

This definition encompasses a large variety of situations. For instance in a scientific experimental environment, where a DR is a collection of experimental data, the author is the scientist in charge of the experimental measures, who certifies the authenticity and the integrity of the data and of the associated descriptive information, and the keeping system is the computer system used to store and managed the experimental data, for instance a database centred system. Similarly, in a document management environment, where the DR is an electronic document, the author is the person who

prepares the final version of the document, and the keeping system is the Electronic Record Management System (ERMS) where the document is kept.

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

In the next subsections we will analyse in more detail the core set events of the pre-ingestion phase, i.e. the most important ones and those more likely to occur, and discuss which actions and which provisions should be taken for each of these events in order to properly collect the related authenticity evidence. In the section 4.3 the same analysis will be carried-out for the core set of events of the *LTDP phase*.

In our proposal, the core set for the keeping phase comprises the following events:

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

4.2.1 CAPTURE

Description The DR author produces the 'final' or 'stable' version of the DR, i.e. a version that (s)he considers as complete and no longer subject to changes, and delivers it to a keeping system. After the capture, the original content of the DR is kept unchanged through its whole lifecycle, but additional information (content, metadata or descriptive information) may be associated to it at later times. Moreover, while entering the keeping system the DR may be associated to one or several contexts that depend from the author and/or from the circumstances in which the DR has been delivered, and/or have been explicitly indicated by the author.

- **Agents**:
  - author: the physical or juridical person who delivers the DR to the keeping system
  - keeping system administrator: the person who has accepted the DR in the keeping system
- **Input:** none
- **Output:** the captured DR
- **Authenticity evidence record:**
  - Identity of the DR
  - Date and time the DR has been created by the author (may be different from the delivery date and time)
  - Date and time the DR has been delivered to the keeping system
  - Identification and authentication data of author(s)
  - Identification data of the keeping system
  - Identification data of the keeping system administrator
  - Digest of the of the DR produced by the author
  - Assessment by the keeping system administrator on the delivery of the DR and the subsequent controls:
    - Assessment of the identification and authentication of the author

  - Assessment of the integrity check based on the digest produced by the author
- Context information associated with the DR that may depend on the author and/or on the circumstances in which the DR has been delivered, and/or has been explicitly indicated by the author.
- Digest of the of the DR produced by the keeping system administrator

4.2.2 INTEGRATE

Description To integrate a DR means to add content or metadata information to a DR already stored in the keeping system. The integrated DR is a new version of the original DR that includes the original version of the DR and the integration. When accessing the DR it is therefore possible to access both the original and the new version. Multiple integrations lead to a DR with a layered structure, in which the original DR and all the subsequent versions can be individually accessed.

- **Agents**:
  - author: the physical or juridical person who has added the new content to the existing DR (i.e. has created the new version of the digital resource.)
  - keeping system administrator: the person who has accepted the new information into the keeping system
- **Input:** any DR in the keeping system
- **Output:** the new version of the integrated DR
- **Authenticity evidence record:**
  - Event type: integration
  - Date and time the integration has taken place
  - Identification and authentication data of the author
  - Keeping system administrator identification data
  - Digest of the of the new version of the DR after the integration (generated by the keeping system, or manually by the keeping system administrator)
  - Assessment by the keeping system administrator on the delivery of the integration and the subsequent controls, including the identification and authentication of the author

Example 1. A record in a medical archive

The original DR is a record in a medical archive. An incorrect piece of information in the record has been detected, and it is therefore important to correct it. The author of the integration is the person who generates the new version of the record. Users subsequently accessing the record see the new version, but have full visibility of the original version as well and of all the details of the integration (author, date, etc.)

Example 2. A record concerning the result of an exam in a university database

In many universities the results of the exams are no longer recorded in paper registries. If incorrect information has been recorded, it is possible to correct the record by integrating it with additional information. The whole process, and the two versions of the record, must be fully documented and visible. Therefore it is necessary to carefully collect the authenticity evidence connected with this event.

4.2.3 AGGREGATE

Description An aggregated DR is a collection of DRs, each of them individually and independently preserved in the keeping system. The aggregated DR has its own identity and its own author, and is originally created as an empty container by its author. The event AGGREGATE corresponds to changing the composition of an aggregated DR, by adding or removing a single or several component DRs. This could be regarded simply as the creation of a new DR but the provenance of the new DR must (logically) include the provenance of its components.

- **Agents**:

  - owner: the physical or juridical person who has created the aggregated DR and has the authority to change its composition
  - actor: the person, entrusted by the owner, who actually changes the composition of the aggregated DR (possibly the owner himself)
  - keeping system administrator: the person who has accepted the change in the aggregated DR into the keeping system
- **Input:** an aggregated DR, and a set of DRs to be added or removed from the aggregated DR
- **Output:** the same aggregated DR with an updated composition
- **Authenticity evidence record:**
  - Event type: aggregate
  - Date and time the aggregation has taken place
  - Identification and authentication data of the actor
  - Identification data of the keeping system administrator
  - Identity of the component DR(s) that are added or removed
  - Digest of the of the aggregated DR after the aggregation
  - Assessment by the keeping system administrator on the transformation and the subsequent controls, including the identification and authentication of the actor

Example 1. An archival file in an archive or in a document management system

The aggregated DR is the archival file. The owner is the person having the authority of managing the file. Component DRs are the individual documents, already preserved in the archive, that are filed in the file, possibly at different times. The keeping system administrator is the curator of the archive or the administrator of the document management system, who has the responsibility of enforcing the rights management and access control policy. The authenticity evidence is typically automatically recorded by the system, but in some cases may be manually recorded by the keeping system administrator. Any change in the composition of the file corresponds to a distinct event and therefore produces a separate authenticity evidence record. The sequence of the authenticity evidence records allows one to trace back the evolution of the file.

Example 2. A collection of experimental data sets in a repository of scientific data

The collection is composed of several experimental data sets, related to a given experiment or to a given series of experiments. Individual data set are delivered at different times, and possibly by different people, and are kept as independent DRs in the repository. The collection is managed by a scientist in charge of the experiment (the owner), who may change its composition by adding or removing the individual data sets. The owner may perform the action directly or by means of authorised collaborators. All changes in the collection must be adequately documented.

4.2.4 DELETE

Description To delete a DR from a keeping system means to remove its content from the keeping system and to destroy it. This is typically done according to precisely stated policies that require the system to preserve some given classes of DRs for only a limited amount of time, and/or to destroy it when the time has expired. While the content of the DR is destroyed, the keeping system generally needs to preserve the evidence that the DR has existed and possibly part of its metadata, and to document the circumstances in which it has been destroyed. This information becomes the content of the new version of the DR preserved by the keeping system.

- **Agents**:
  - owner: the physical or juridical person who has the authority of order the DR deletion
  - keeping system administrator: the person who performs the action of removing the DR from the keeping system
- **Input:** any DR in the keeping system

- **Output:** the same DR with a different content:
  - the original content of the DR is removed;
  - the identifier and eventually part of the metadata are preserved.
- **Authenticity evidence record:**
  - Event type: delete
  - Date and time the deletion has taken place
  - Identification and authentication data of the owner
  - Identification and authentication data of the administrator
  - Digests of the of the DR before and after the deletion
  - Assessment by the keeping system administrator on the transformation and the subsequent controls, including the identification and authentication of the actor

Example 1. A digital document in an administrative archive

For each class of administrative document, the law states a minimum time the record should be preserved. When the time expires the document may be destroyed. This may happen either because the owner of the document (the person responsible of the administrative procedure to which the document relates) issues an explicit order, or because a specific policy has been given to the keeping system administration to delete all the documents of a given class when their time expires.

When a document is deleted a minimal set of metadata (author, creation time, etc.), including a digest of the last version of the document, is preserved as a new digital resource. Preserving the digest may be interesting, since, for instance, it may be used to authenticate a copy of the document preserved in an another archive, even after the original has been destroyed.

4.2.5 MIGRATE

Description To migrate (more accurately - transform) a DR means to change the data format of one or several of its components. This is generally triggered by technical obsolescence, but may be as well the result of adopting more restrictive policies on the formats accepted in the keeping system. Depending on the circumstances, the migration may or may not require the consent of the owner of the DR. As a result of the migration a new version of the DR is created, which is supposed to preserve the intellectual content, despite the format migration. The most delicate part of this transformation, is to verify, and to assess, that the integrity of the DR has been maintained, i.e. that its intellectual content has not changed. Depending on the circumstances, producing an assessment may involve the owner of the DR, the keeping system administrator, or both.

- **Agents:**
  - owner: the physical or juridical person who originally created the DR, or, in general, anyone who has acquired the right to manage the DR.
  - keeping system administrator: the person who is responsible of performing the migration.
- **Input:** any DR in the keeping system
- **Output:** a new version of the DR
- **Authenticity evidence record:**
  - Event type: migration
  - Date and time the migration has taken place
  - Identification data of the keeping system
  - Identification and authentication data of the owner
  - Identification and authentication data of the keeping system administrator
  - 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.
- Digest of the of the new version of the DR after the migration (generated by the keeping system, or manually by the keeping system administrator)
- Statement that the information content of the DR has not changed, by the owner and/or the system administrator, specifying also the criteria that have been adopted to perform the assessment

Example 1. Accepted and converted formats

Even if the administration of an archive decides that only a limited set of data formats are suitable for the DR that have to be kept in the archive, it may still be very difficult to be certain that users deliver to the archive only documents in these formats. Therefore some archives accept DR in a broader variety of formats, *accepted formats*, and eventually perform a format migration to an *internal format* immediately after a document has been delivered to the archive. For instance the archive may accept documents in MS Word 98 format, an accepted format, but they are later converted to PDF/A which is the corresponding internal format.

Reasonable criteria to perform the assessment about the integrity of the DR after the format migration may be: checking that the number of pages has not changed, comparing the word content of the same page in the two versions (for every page or by sampling), etc. Since the migration is performed immediately afterwards the DR is delivered to the archive, the statement that the intellectual content of the DR has not changed may be issued by the author of the DR, who is required to compare the two versions, according to clearly specified criteria. All this information needs to be clearly specified in the authenticity evidence record.

For data files the checks would include the comparison of data values and things like coordinate systems and units.

As for the content of the DR, it may be wise to maintain both versions: before and after the format migration. At a later time, even if the accepted format is no longer supported by that specific archive, it may still be supported by some other system, and therefore the original version may be a crucial element in assessing the authenticity of the DR.

Example 2. Complying with a new policy about formats

A new policy is defined in an archive banning some data formats that were previously used as internal formats. The archive administrator then performs the conversion of all DRs which have some component in the banned format. A long time may have passed since the creation of the migrated DRs, and their authors are therefore presumably not at hand: the assessment has then to be performed by the system administrator.

4.2.6 TRANSFER

Description A transfer occurs when a DR is moved from a keeping system (origin system) to another keeping system (destination system). The transfer needs to be authorized by the owner of the DR, and involves also the responsibility of the administrators of both keeping systems. After the transfer, the DR may eventually be deleted in the origin system, but this action should be considered a separate event. According to the circumstances, and the different policies adopted by the two repositories, the DR identity may be maintained or may change in the transfer.

A transfer may indeed be considered as the sequence of two separate steps: i) preparing the DR for shipping in the origin system and delivering it to the destination system; ii) accepting the DR in the destination system. As a consequence, two distinct new versions of the DR are produced: DR' which is kept in the origin system, to preserve memory that the DR has been transferred, and DR", the new version that will be kept in the destination system.

- **Agents:**

- owner: the physical or juridical person who originally created the DR, or, in general, anyone who has acquired the right to manage the DR and is acknowledged by the keeping system administration as the person responsible for it.
- origin system administrator: the administrator, or the person who is responsible in the origin system, on behalf of the administrator, of performing the transfer.
- destination system administrator: the administrator, or the person who is responsible in the destination system, on behalf of the administrator, of performing the transfer.

- **Input:** any DR in the origin keeping system
- **Output:**
  - DR': the new version of the DR which is kept in the origin system
  - DR": the new version of the DR which is kept in the destination keeping system
- **Authenticity evidence record:**

Two different and independent keeping systems are involved in a transfer, the corresponding authenticity evidence record must therefore contain evidence produced, and adequately authenticated, by the administrators of both systems. As a consequence, the authenticity evidence record is divided in two separate parts, which are generated and preserved in the two systems.

Origin system

- Event type: transfer-out
- Identification data of the origin keeping system
- Identification data of the destination keeping system
- Date and time the DR has been prepared for migration and shipped to the destination system
- Identification and authentication data of the owner of the DR who has given the authorization for the transfer
- Identification and authentication data of the origin keeping system administrator
- Evidence that the DR has been received and accepted by the destination system; this item is added at a later time, when the transfer process is completed
- Digest of the DR produced and authenticated (digitally signed) by the origin keeping system administrator

Destination system

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

4.2.7 SUBMIT

Description A submit occurs when a DR is transferred from a keeping system to a LTDP system. The submit needs to be authorized by the owner of the DR, and involves also the responsibility of the

administrators of both the keeping system and the LTDP system. As for the content and the structure of the (set of) Submission Information Package(s) (SIP) that contains the DR and is delivered to the LTDP system, it should comply with a submission agreement established between the keeping system (i.e. the Producer in the OAIS reference model) and the LTDP system (the OAIS). After the submission, the submitted DR may eventually be deleted in the origin system, but this action should be considered a separate event. The DR identity is maintained in the keeping system, but a new identity may be given to the DR in the LTDP system.

Similarly to a transfer, a submission may indeed be considered as the sequence of two steps: i) preparing in the keeping system the DR for shipping; ii) receiving and accepting the DR in the LDTP system. As a consequence, two distinct new versions of the DR are produced: DR' which is kept in the keeping system, to preserve memory that the DR has been submitted, and DR", the new version, conveniently restructured in a SIP, that is accepted into the LTDP system.

- **Agents:**
  - owner: the physical or juridical person who originally created the DR, or, in general, anyone who has acquired the right to manage the DR and is acknowledged by the keeping system administration as the person responsible for it.
  - keeping system administrator: the administrator of the keeping system, or the person who is responsible in that system, on behalf of the administrator, of performing the submission of the DR.
  - LTDP system administrator: the administrator of the LTDP system, or the person who is responsible in that system, on behalf of the administrator, of accepting the submitted DR.
- **Input:** any DR in the keeping system
- **Output:**
  - DR': the new version of the DR which is kept in the origin system
  - DR": the new version of the DR, restructured as a SIP, accepted into the LTDP system and ready for ingestion.
- **Authenticity evidence record**:

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. As a consequence, the authenticity evidence record is divided in two separate parts, which are generated and preserved in the two systems. As for the LTDP system the authenticity evidence record will become part of the PDI.

Keeping system

- Event type: submit-out
- Identification data of the keeping system
- Identification data of the LTDP system
- Date and time the DR has been prepared for submission
- Identification and authentication data of the owner of the DR who has given the authorization for the submission
- Identification and authentication data of the keeping system administrator (the person who was in charge of the origin keeping system when the transfer took place)
- The evidence that the DR has been received and accepted by the LTDP system; this item is added at a later time, when the submission process is completed
- Digest of the DR produced and authenticated (digitally signed) by the keeping system administrator

LTDP system

- 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

4.3 PHASE 2: LONG TERM PRESERVATION

The long term preservation phase begins when the DR is delivered to a LTDP (Long Term Digital Preservation) system and goes on as long as the DR is preserved. We assume that during this phase the DR may be transferred between different LDTP systems.

During the long term preservation phase, the DR may undergo several kinds of migrations. OAIS distinguishes refreshment, replacement, re-packaging and transformation as types of migration. Only transformation changes the bit-sequences of the Content Data Object i.e. change of format. In addition there may be transfers between different preservation systems. As far as the authenticity and provenance of the DR are concerned, the target that should be pursued during this phase is to collect any further evidence that, together with the evidence that has been originally delivered to the LDTP system when the DR has been submitted, would allow tracing the authenticity and provenance history of the DR to its creation. That means being able to verify and check that the completeness, the accuracy and the reliability of the DR have not been altered within the preservation repository.

As for the pre-ingestion phase, we do not pretend to draw a comprehensive list of all possible events that may happen during this phase, but we restrict to a *core set* of events, that represent the most important ones, as well as the ones which are likely to occur in most of the environments in which DRs are preserved.

More precisely, in our proposal, the core set for the long term preservation phase comprises the following events:

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

In the next subsections we will analyse in more detail the core set events of the LTDP phase and discuss which actions and which provisions should be taken for each of these events in order to properly collect the related authenticity evidence. In the discussion we will refer to the OAIS reference model whose main concepts and terminology we give for granted.

According to the OAIS reference model, many delicate and complex activities are carried out in connection with each of these events, but we will focus our discussion on the sole aspects related to authenticity and provenance of the DR. Therefore for each event we shall point out:

- which of the activities carried out in connection with the event may have some impact on the authenticity and the provenance of the digital DR;

- which information (authenticity evidence) has to be gathered and preserved in the PDI (Preservation Description Information), and more specifically in the Provenance, Context and Fixity components, in order to conveniently document the history of the DR.

A final remark about the terminology. Technically an OAIS manages Information Packages (SIP, AIP, DIP), and all transformations connected to the events we are discussing concern information packages. This does not mean that the DRs that have been submitted to the OAIS do not exist anymore: they are simply stored and preserved as part of information packages. Therefore, in the description of the events of the LTDP phase we may still refer to DRs, which is correct, since these, and not the information packages, are the objects that we are tracing through their lifecycle, and for which we are collecting the authenticity evidence.

4.3.1 LTDP-INGEST

Description According to the OAIS terminology, one or more SIPs contain the version of the DR delivered by the submitting keeping system and accepted by the LTDP system (DR" in sect. 4.2.7). The main actions performed during the ingestion that may affect authenticity and provenance are:

- *the DR may be given a new identifier, if the original one is not compliant with the LTDP system standards;*
- *a general assessment is performed on all authenticity and provenance evidence associated to the submitted DR;*
- *format conversions may be performed, if necessary, according to the policies of the LTDP system.*

- **Agents:**
  - LTDP system administrator (or the person who is responsible, on behalf him, of performing the ingestion)
- **Input:** a SIP, i.e. the DR submitted and accepted by the LTDP system (DR" in sect. 3.3.7)
- **Output:** an AIP and updates to the information in the Data Management
- **Authenticity evidence record:**
  - 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

4.3.2 LTDP-AGGREGATE

Description According to the OAIS reference model, an Archival Information Collection (AIC) is composed of several AIPs, which are aggregated according to some predefined criteria. The aggregated AIPs may be added to an existing AIC, or a new AIC may be generated as a result of the aggregation. In order to be able to trace the authenticity and the integrity of the DRs contained in the aggregated AIPs, it is necessary to gather the following information:

- *the identity of the AIPs involved and of the AIC, if already existing;*
- *the circumstances of the aggregation;*

- *the criteria that triggered the aggregation;*
- *the digest of the resulting AIC.*
  - **Agents:**
    - LTDP system administrator (or the person who was responsible, on behalf of him, of performing the aggregation)
  - **Input:** a set of AIPs, and possibly an existing AIC
  - **Output:** a newly created AIC, or an updated version of an AIC
  - **Authenticity evidence record:**
    - Event type: aggregate
    - Date and time the aggregation has taken place
    - Identification data of the LTDP system
    - Identification and authentication data of the LTDP system administrator
    - Description of the criteria according to which the aggregation was performed
    - Identity of the aggregated AIPs
    - Identity of the AIC
    - Digest of the AIC generated (or modified) by the aggregation, authenticated by the system administrator

4.3.3 LTDP-EXTRACT

Description As a result of the extraction, one or several AIPs that were part of an AIC are removed from the collection to be preserved as independent objects. As a consequence, a new version of the AIC is produced as well. In order to be able to trace the authenticity and the integrity of the all the DRs originally contained in the AIC, it is necessary to gather the following information:

- *the identity of the AIC and of the AIPs that are extracted;*
- *the circumstances of the extraction;*
- *the criteria that triggered the extraction;*
- *the digests of the resulting AIC and AIPs.*
  - **Agents:**
    - LTDP system administrator (or the person who was responsible, on behalf of him, of performing the extraction)
  - **Input:** an AIC
  - **Output:** one or several AIP and a new version of the AIC
  - **Authenticity evidence record:**
    - Event type: extract
    - Date and time the extraction has taken place
    - Identification data of the LTDP system
    - Identification and authentication data of the LTDP system administrator
    - Description of the criteria according to which the extraction was performed
    - Identity of the AIC
    - Identity of the extracted AIPs
    - Digests of all the extracted AIPs and of the new version of the AIC.

4.3.4 LTDP-MIGRATE

Description To migrate/transform an AIP means to change, during its preservation, the data format of one or several components of the DR(s) contained in the AIP. This is generally triggered by technical obsolescence, but may be as well the result of new policies adopted by the LTDP system about accepted formats. As a result, of the migration a new version of the DR(s) affected by the migration is generated, which is supposed to preserve the intellectual content, despite the format migration. The most delicate part of this transformation is to verify that the integrity of the individual DR has been maintained, i.e. that its intellectual content has not changed.

- **Agents:**

    - LTDP system administrator (or the person who is responsible, on behalf of him, of performing the migration)
- **Input:** one or several DR contained in an AIP or in an AIC
- **Output:** a new version of the AIP or AIC
- **Authenticity evidence record:**
    - Event type: migration
    - Date and time the migration has taken place
    - Identification data of the LTDP system
    - Identification and authentication data of the system administrator
    - 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.
    - Digest of the of the new version of each affected DR after the migration
    - 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
    - Digest of the of the new version of the AIP produced by the migration

4.3.5 LTDP-DELETE

Description To delete a single or several DRs from a LTDP system means to remove from the system the related content information and to destroy it. This is generally done according to precisely stated policies, that have been negotiated with the producer as part of the submission agreement, that may require the system to preserve a given classes of DRs, to which the deleted objects belong, only for limited amount of time, and to destroy them when the time has expired. While the content information of the DRs is destroyed, the LTDP system needs to preserve the evidence that the DRs have existed and have been preserved in the system for a given extent of time, possibly to preserve part of their metadata, and to document the circumstances in which they have been destroyed. Even if all the DRs contained in an AIP are deleted, the AIP is preserved, as the evidence of the deletion becomes part of its PDI.

- **Agents:**
    - LTDP system administrator (or the person who is responsible, on behalf of him, of performing the deletion)
- **Input:** one or several DRs contained in an AIP
- **Output:** evidence that the DRs have existed and have been preserved for a given extent of time
- **Authenticity evidence record:**
    - Event type: delete
    - Date and time the deletion has taken place
    - Identification data of the LTDP system
    - Identification and authentication data of the system administrator
    - Identification of all the deleted DRs
    - Statement by the system administrator specifying the motivations and the circumstances in which the DRs have been deleted
    - Digest of all the deleted DRs before the deletion

4.3.6 LTDP-TRANSFER

Description A transfer occurs when a DR or a set of DRs are moved from a LTDP system (origin system) to another LTDP system (destination system). The transfer involves the responsibility of the administrators of both LTDP systems, and should refer to a submission agreement negotiated between the two LTDP systems. After the transfer, the transferred DRs may eventually be deleted in the origin system, but this action should be considered a separate event. The DRs' identities are generally maintained in the transfer, since persistent identifiers should be used in LTDP systems. In principle the transfer consists in moving a single AIP, containing all the DRs, between the two LTDP systems. In practice this may be broken down into multiple SIP transfers. We assume that this AIP is the result of a set of previous transformations and that it complies with the submission agreement.

The transfer may indeed be considered as the sequence of two distinct steps: i) preparing the AIP in the origin system for shipping and delivering it to the destination system; ii) accepting the AIP as one or more SIPs in the destination system. In fact two distinct sets of authenticity evidence items must be gathered and preserved in the origin and the destination systems.

- **Agents:**
  - origin system administrator: the administrator, or the person who is responsible in the origin LTDP system, on behalf of the administrator, of performing the transfer.
  - destination system administrator: the administrator, or the person who is responsible in the destination system, on behalf of the administrator, of performing the transfer.
- **Input:** an AIP in the origin system
- **Output:**
  - a new version of the AIP in the origin system
  - SIPs received by the destination system
- **Authenticity evidence record:**

There are two separate authenticity evidence record. In the origin system, the evidence that the AIP has been transferred to the destination system needs to be incorporated in its PDI, thus generating a new version of the AIP. In the destination system, the evidence that an SIP has been received from the origin system must be gathered and will be later incorporated in the PDI of the newly generated AIP during the ingestion process.

Origin system

- Event type: transfer-out
- Identification data of the origin system
- Identification data of the destination system
- Date and time the AIP has been delivered to the destination system
- Reference to the submission agreement according to which the AIP has been transferred
- Identification and authentication data of the origin system administrator
- Evidence that the AIP was received and accepted by the destination system
- Digest of the of the AIP as updated after the transfer and authenticated (digitally signed) by the origin keeping system administrator

Destination system

- Event type: transfer-in
- Identification data of the origin keeping system
- Identification data of the destination keeping system
- Date and time the SIP has been received from the origin system
- Identification and authentication data of the destination system administrator
- Assessment by the destination keeping system administrator on the delivery of the SIP by the origin keeping system and the subsequent controls:
  - Assessment on the identification and authentication of the origin keeping system

- Assessment on the trustworthiness of the channel used of the data channel used for the transfer
- Assessment on the integrity check performed on the digest produced by the origin system administrator

- Digest of the SIP (including all the evidence items listed above) authenticated by the destination system administrator

5 GUIDELINES FOR THE MANAGEMENT OF AUTHENTICITY EVIDENCE

5.1 FROM THE MODEL TO THE OPERATIONAL GUIDELINES

Translating the model that we have discussed in the previous section into practical guidelines which could be actually adopted in a specific real-life environment requires dealing with two kinds of problems:

i. in our discussion we have restricted our analysis to a core set of events, which we have carefully selected in order to get the most important ones, as well as the ones which are most likely to occur; however, in a given environment additional events may need to be considered that are specific to that environment;
ii. we have detailed the evidence information to be gathered at a level that we may consider as reasonably complete and essential, but that in some environments it may be considered too complex to deal with, or, on the other hand, insufficiently detailed; this is especially likely to happen for the keeping phase, since keeping systems often do not comply with precise standards and requirements and may lack some capabilities, or may have to deal with problems we could not account for.

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

We will provide in the following subsections operational guidelines to deal with the problem in a systematic way, but for a deeper understanding of the matter and to have a feeling of the problems one has to face in the practical implementation of the guidelines, the reader should refer to the companion deliverable D24.2, where a case study analysis is documented that has been performed on several test environments provided by APARSEN partners. This section has been actually finalized only after having carried out the case studies, and therefore the guidelines we present have been refined according to the experience gained during their practical implementation, and some comments about our experience are included as well.

5.2 AUTHENTICITY AND THE DESIGNATED COMMUNITY

The concept of *Designated Community (DC)* ("an identified group of potential Consumers who should be able to understand a particular set of information") is indeed central to the OAIS reference model according to which "the primary goal of an OAIS is to preserve information for a designated community over an indefinite period of time".

Therefore, as a first step, one should understand what authenticity means to the DC, that is:

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

Answers to these questions may be quite different in different situations and may depend on several factors, as for instance the relationship and the level of trust between consumers and producers, legal compliance requirements and the specific nature of the DRs that are preserved.

On the other hand, increasing the level of detail in authenticity evidence adds cost and complexity to the management of the LTDP repository therefore it should be pursued only to the extent that is deemed necessary and reasonable by the DC.

When dealing with an existing LTDP repository, that is analysed to assess the adequacy of the current practices or to suggest improvements, the starting point may be understanding what kind of authenticity evidence is currently preserved and investigating if the DC actually deems it as sufficient for its purposes. The answer to this question is not at all evident, since in most cases the DC was not given any choice about which evidence had to be preserved, and just had to accept it.

Moreover, quite frequently, as our case study experiences have proved, the DC appears not to have a sufficient understanding of the whole DR lifecycle and of the consequent threats to the authenticity connected to transformations and changes of custody. Therefore it is very important to actively involve the users in the DC in the analysis, and to make sure that a proper level of understanding has been achieved and that they request what they actually need.

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

See for example the discussion of the DC in the case studies discussed in D24.2 [4].

5.3 IDENTIFYING AND ANALYZING RELEVANT LIFECYCLE EVENTS

The next step is to analyse the workflow of the DRs that are to be preserved in the repository, from their creation on, to identify the lifecycle events that are relevant to the management of the authenticity. As we have already pointed out several times, the *whole* lifecycle must be considered, including the pre-ingestion phase during which the DR may undergo important transformations and changes of custody that may affect its authenticity and integrity. Moreover, in some cases (as we experienced at least once in our case study analysis) we may identify, for a given repository, several DR types and several workflows. In these situations the lifecycle analysis and the following steps need be repeated for each workflow.

Once the relevant lifecycle events have been identified, they must be compared and fitted into the *core set events* that we have discussed in sections 4.2 and 4.3 and that provide a reference and a template on the way authenticity evidence should be gathered and managed.

According to our case study experience the core set that we have proposed has proved to be quite a robust choice, in the sense that all the relevant events we have identified could fit well in one of the core set events. However, it is still possible that in a given environment additional events may need to be considered that are specific of that environment. The additional event should be analysed and represented according to the same criteria we have used for the core set events, but additional descriptive information should be added to the authenticity evidence in order to make it *self-contained*. That means that it should be possible to understand the meaning of that information even for somebody who is not necessarily aware of all the details connected to this kind of event, which, being non-standard cannot be taken for granted. In practical terms, one should include in the authenticity evidence proper description that allow a third party to understand it.

Then, for each lifecycle event we have identified as relevant for the management of the authenticity, definitions of the corresponding core set event we have given in sections 4.2.and 4.3 should be considered as a template (a sort of checklist) to formalize the definition of that specific lifecycle event, to identify responsibilities and to understand which authenticity evidence should be gathered and which controls should be performed.

As a matter of fact, our case study experience has shown that our definition of the core set events is quite detailed, i.e. that the checklist is usually a quite comprehensive one. Therefore, sometimes (quite often) part of the authenticity evidence that the templates mandate to collect was not actually collected in the current practices that we had been analysing. This does not necessarily mean that there is something wrong, nor that the current practices are inadequate: one should instead carefully consider and make an assessment of, every single missing item of evidence, taking into account the specific needs of the designated community and a number of further details, for instance the systems involved and their ownership. This does not means either that the templates are wrong, since we are convinced that, due to the nature of the matter we are dealing with, it is still better to have a checklist that may be deemed as too large that to risk missing some important item.

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

- the item is intended to document a control that is actually performed but not recorded in the AER by a system under the ownership of an organization which is trusted by the DC;

- the item is intended to prove that the integrity of the item has not been affected by the transfer between two systems that are under the ownership of the same organization which is trusted by the DC;
- the item relates to some provenance information which is of no interest to the designated community.

Anyway, besides a few general criteria as above, it is difficult, probably impossible, to give an exhaustive list of specific criteria for deciding whether a given authenticity evidence item should be recommended or not, mostly due to the variety of situations and the complexity of systems. To get a more precise idea of the way to proceed the reader should refer to the case study analysis presented in deliverable D24.2 [4], and most specifically to sections 3.4 and 3.5.

5.4 DEFINING THE AUTHENTICITY MANAGEMENT POLICY AND THE AUTHENTICITY EVIDENCE RECORDS

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

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

In all three cases the conclusions should be explicitly and clearly documented. More specifically, in case a) one must specify where and when the information is actually collected in the current practices; in case b) a convincing proof should be given and the problem should be made clear to qualified representatives of DC who should give their explicit consensus; finally, in case c) an improvement of the current practices should be recommended and the information to be collected should be clearly specified, along with the procedure to collect it.

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

i. A general statement about the meaning of authenticity to the DC, specifying the kind of authenticity evidence the DC is interested in and its purpose in collecting and preserving it. The statement should be accompanied by a clear delimitation of the DC and by the explanation of how the opinion of the DC was actually gathered.
ii. The specification of the lifecycle, and more precisely of the event in the lifecycle that have been identified as relevant to the management of authenticity.
iii. For every relevant event in the lifecycle the definition of the controls corresponding to that event that must be performed and of the AER, that is the list of all the authenticity evidence items that must be collected, together with the specification of the procedures that should be followed to collect them.
iv. For every authenticity evidence item recommended by an event template in the model which is not part of the corresponding AER in the authenticity management policy, a clear explanation of why it is considered acceptable that that item is not part of the AER.

5.5 FORMALIZING AUTHENTICITY PROTOCOLS

A further step consists of the operational implementation of the authenticity management policy defined in the previous subsection, and more specifically the formal definition of the controls that must be performed in connection with each event and the procedures that must be followed to collect the AER (sees step iii in the definition of the policy). We propose an implementation strategy which is based on the concept of *Authenticity Protocol* that has been introduced within the CASPAR project.

According to the original definition in CASPAR [21] 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 analysed in every AS, and possibly the outcome of the execution.

According to the definition in CASPAR [21] 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 (see sect. 4.1), 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

For a practical example of this definitions the reader may refer to section 3.6 in the companion deliverable D24.2 [4], where the implementation of a specific authenticity protocol is discussed and developed.

6 SECURE LOGGING MECHANISMS

Digital objects traverse different transformations during their life cycle. The transformations trigger events that have been outlined earlier in Sections 4.2 and 4.3. The occurrence of the events has to be recorded in a proper way so that their impact on the authenticity and provenance of a DR can be traced, assessed and proved. The recoding of the events that can be seen as system events is denoted as logging. The log files are generated by the system in an automated way and they are fundamental to the analysis and audit of the events that occurred during the life cycle of an object. Standards as ISO 27002 (ISO 17799) (Information technology — Security techniques — Code of practice for information security management) [39] and ISO 27001 (Information Security Management Systems) [40] clearly identify log files, their monitoring and audits as well as their security as a fundamental property to system security. The data stored in log files is highly sensitive and has to be protected from unauthorized access and from tampering. The following section describes the requirements and properties of log file systems and how the content of log files can be protected.

6.1 LOG FILES AND LOGGING SYSTEMS

The data describing authenticity and provenance information of electronic records has to be stored in order to be analysable, understandable, reproducible and preservable for later use. Whenever there is some interaction with resources stored within an archival system the system has to document every event that has happened and even every attempt of interacting with the data stored within the archive. The capturing of these attempts and events is called logging, which is a crucial feature of every complex system exposed to interaction with other systems or with users in general. Log files are essential in order to understand and trace back every interaction with the system at a later point in time and in order to find explanations and evidence for certain incidents that need clarification.

The granularity of such logs can be very different and it is highly dependent on the requirements of the setting. Each event that occurs within an archive would lead to an entry in the log file if this event was considered important. Whether an event is considered important or not should hence be defined in a logging policy that describes the requirements to the log system. When thinking of provenance data obviously every event and incident that altered records in any way has to be documented at a high level of detail [41]. Scalability is also an issue. The larger an archive grows, the higher is the amount of user interactions. Recording all the events can become a serious challenge. Storing too many entries of low significance produces a lot of noise, which hinders efficient analysis for auditors. This in turn facilitates attackers to hide the traces of their malicious actions within a tangled mass of logged events.

6.1.1 Architecture, Types and Phases of Logging

A logging system consists of different components [1]: devices, relays and collectors. These components interact with each other. Whenever there is a relevant state change within the archive, this event has to be announced from the sender (device) to a facility in charge of producing the auditable log file (collector). The relay might be used to forward log messages. Which of these events are considered relevant depends on the purpose of the archive. This decision is made with the help of policies. In general all events that alter a digital object should be recorded. Archives tend to contain a large number of objects that are subject to frequent interactions. This means that log files can grow rapidly and become very large. In general there are two different types of log files. The first type is sequential logs, which are a sequence of all events. Such logs are never truncated and keep growing possibly indefinitely. The second type is cyclic logs that are rotated after a given amount of contained log events. In the area of digital preservation and archives, only sequential logs are of interest. The reason is that all information should be preserved for the long term and therefore never be truncated.

There are two phases when it comes to logging events [1]. Each phase is associated with certain requirements that have to be met in order to provide secure logging facilities. The first phase is called transmission phase and it consists of the following requirements:

- Origin authentication
- Message confidentiality

- Message integrity
- Message uniqueness
- Reliable delivery

Events are stored by the collector which is also responsible for generating the log files (audit trails). This central instance ensures the authenticity of the sending devices and the event information. In order to be stored, these events have to be transmitted via possibly uncertain channels. Methods have to be applied, that guarantee that the sending device is actually the one it is purporting to be. The provenance information to be logged itself is in many cases highly confidential. It contains for example information about the way people have been engaged in the evolution of a digital resource. Logs have to be safeguarded against unauthorized access. As the service that performs the logging is usually not on the same device that reports events, the messages need to be protected before they are transmitted. This means that the sending device and the receiving collector have to agree on a shared cryptographic procedure. The message integrity demands that the content of event logs cannot be altered, neither during transmission, nor once it was stored by the collector. This is a very fundamental requirement, as otherwise the integrity of the whole archive could be threatened by manipulating existing event logs. Depending on the amount of stored objects and the number of interactions, the total amount of log events can be- come enormous. This can become a problem for scalability and for identifying relevant information.

The second phase is the storage phase. This stage covers the generation of the audit log itself and it consists of three requirements, listed below:

- Entry accountability
- Entry integrity
- Entry confidentiality

As already identified by the requirements in the transmission phase, the events sent by the devices have to be verified before they can be appended to the audit log. Each of these log records must contain information about the sender and the receiver of the log event. Entry integrity refers to the no-alterations paradigm necessary in audit logs. When an event message is added to the log, it must not be possible to change its content or to delete entries from the audit trail. It must also be ensured that no fake entries can be added to the audit log. Details how these requirements can be accomplished can be found in Section 6.3. The confidentiality property may demand that the audit trail itself has to be encrypted.

6.2 CRYPTOGRAPHY AND PUBLIC KEY INFRASTRUCTURES IN DIGITAL PRESERVATION - SECURE LOGGING

Digital archives need to be trustworthy, which implies that the users of such an archive can rely on the authenticity of the records contained therein. This basis of trust is established by several factors and it cannot be mapped onto a single property. Trust on one hand is a social phenomenon, relying on the reputation of persons and institutions in charge. This type of trust is hard to measure. On the other hand, trust can be supported by technical methods from the field of encryption. With the availability of high performance network connections, distributed software architectures and the increasing amount of sensitive data being transferred, the threat of data manipulations, forgery and fraud is also increasing dramatically. Mechanisms for protecting data and for detecting attempts of manipulation are required.

Fortunately, such technologies exist and therefore can be used in order to protect data from sabotage. There are two main methods available which can be used in order to support and prove authenticity and provenance. The first one is the concept of digital signatures; the second one is encryption of digital objects. Both of these methods stem from the field of cryptography [53] and can be used in order to enhance authenticity and the trustworthiness of archives. Digital signatures are used in order to verify that the content of a digital object did not change. The purpose of this technique is to sign the provenance information and therefore enable the detection of potential intruders. Although digital signatures are very valuable to digital preservation, they do not solve this problem completely. As the

signature itself is also data it faces the same threats regarding its preservation as do all other data. It has to be protected from unauthorized access, because otherwise an attacker could try to replace the stored signature with a newly calculated one. It is important to adapt the signing algorithms to current standards in the field of cryptography in order to guarantee the safety of the method. Although the above mentioned technologies seem to solve some crucial problems securing provenance data, this is not to say that absolute security can ever exist. Whenever the validity and integrity is based upon the correctness of hash values, it has to be clear that these values are also just data themselves. These could be manipulated as well, although the tamper resistance had been increased dramatically. There is always an overhead when it comes to the calculation and generation of signatures. Encryption and decryption consume considerable amounts of computing power as well.

Cryptographic mechanisms are of particular interest within the area of digital preservation. Encryption and digital signatures have to secure sensitive content from being accessed today, but they also have to be decipherable and verifiable in the long term.

As technological developments advance, so do the tools and methods for breaking the cryptographic methods that have once been considered safe. This improvements on both sides lead to the demand for constantly updated keys, signatures and ciphers, whenever the cryptographic tools become obsolete as well. This issue adds another layer of complexity to the generally challenging task of digital preservation.

Encryption is also a special threat for digital long term preservation. It is another factor that has to be taken considerable care of. If the decryption keys are lost the information might also be lost forever. On the other hand encryption methods also bear the risk of becoming obsolete. This entails that not only the digital objects themselves have to be migrated to new formats, but also the cryptographic protocols have to be adapted to the latest technological standards. As security in general is the race between the developers of secure protocols and code breakers this topic demands close attention.

After this general introduction of digital signatures and encryption, the next section provides an overview of the technologies introduced and their application in digital archives.

6.3 SCENARIOS OF APPLICATION

Provenance data is a highly sensitive source of information, the fundamental problems regarding provenance are described in [26]. A special security model of provenance data is needed, as outlined in [8]. It contains who contributed what kind of data at a given time and how all electronic resources in an archive are related with each other. This information is needed to prove the custody of some document or file, and to demonstrate what processes interfered with the resource. Provenance data can also be used in order to control the quality of data and it can be used to differentiate original documents from copies. All these functions can then only be used and applied properly if the provenance data is stored in a secure way. This includes not only the resistance against modifications, but also the prohibition of the deletion of records in the provenance chain or also their unjustified addition. All these possible threats have to be anticipated or at least detected. If the digital object itself contains sensitive data it should be encrypted as well. This prevents the information from being read without the permission required to do so. The required level of security depends on actual system and its requirements. There is no general rule that can be applied to all archives. As provenance data maps relations between different versions and contributors to a given resource, one possible representation of provenance models is a DAG [61]. In contrast to provenance chains, graph models can express parallelism. Such a graph consists of nodes and edges, where the nodes represent artefacts and the edges a relation between two artefacts. The graph itself is directed. This direction determines the sequence of events that occurred. For the same reason there also exist no cycles as it is impossible for one artefact, to return to completely the same state again, after some process interfered with the object. Such a provenance graph can include highly sensitive data. Aldeco et. al identified in [50] four fundamental security requirements have to be fulfilled as listed below:

- confidentiality
- authentication

- non-repudiation
- integrity

These requirements can be met by various approaches from the field of cryptography. One fundamental aspect is to limit access to the provenance graph itself, and only granting permission to read and write information to the provenance store to those persons who are in charge. This includes mainly two groups of people: contributors and auditors. The first group is directly involved in the data production processes, and all of their actions need to be recorded by the provenance system. Once an event is added, it can neither be deleted (integrity) nor can it be denied (non-repudiation, authentication). There are at least two different user groups, with different permissions and access rights, which have to be enforced by the system. A contributor must only be able to read his or her own contributions to a process (confidentiality). An auditor must be able to read and analyse the provenance paths that are under his current supervision. To prevent such evidence from being manipulated, write once media could be used.

6.3.1 Append-Only Signatures

Fixity is a key aspect of provenance data. Once it has been recorded it should not be changed. Digital signatures can be used in order to verify the integrity of the data. Signatures can be produced by human beings manually or automatically by services. Contributors use their digital signature in order to sign the event that was recorded by the system. Such a signature uniquely identifies the creator of some digital information and it seals the content of a log message. In order to be useful in the audit process, this signature must not be alterable without being detectable by the auditor. It has to be ensured that no information can be deleted from the graph. Also it has to be ensured that no fake entries can be added in the middle of a sequence. In order to record new provenance data, the graph must be extensible, but only at the end of a path within this graph; no intermediate events are allowed to be entered ex post. In order to achieve this requirement, the system should make use of the methods described in Section 6.2, by combining these in a special way. Append-only signatures aggregate the single signatures from records to an overall signature. By doing so it can be detected if an item within a path was changed. The last node within a provenance graph requires special attention. It is the only item in a chain that does not have a successor. Such a successor is needed in order to verify the integrity of this particular node. If an attacker gains read access to the provenance entry, (s)he could extract the changes that lead to the provenance node and undo these changes within the document. Then he could remove this last entry and thus reverting the history of the document. (S)He could then proceed with this method again and alter a complete provenance path [27]. Provenance chains have to be secured against the deletion or addition of intermediate records to that chain, but also the very last item has to be reliable. Details on the technical and theoretical background for systems implementing a Log forward-secure and Append-Only Signature system is for instance described by Yavuz et.al. in [65].

6.3.2 Untrusted Loggers

A secure logging system should also be protected against attacks from the inside. This entails that a system should also not rely on the personnel operating it. It must be guaranteed that all interactions with the logging system itself are monitored and that there is no possibility of changing provenance or authenticity data without evidence for doing so. After a system has been compromised the records added after the attack cannot be trusted. Nevertheless it is possible to ensure the validity of the log records that have been recorded before the intrusion [57]. The concept behind this technique is to ensure that the untrusted logging device has to answer audit questions to a trusted auditor on a regular basis. The fundamental aspects and suggestions for these kinds of problems are illustrated in [12].

6.3.3 Summary

To sum up, most of the models outlining an architecture for secure provenance records include the following fundamental properties [27]:

- Provenance data itself has to be encrypted to prevent eavesdropping

- Integrity of provenance data is protected by signatures
- The provenance graph in its complete form is protected by a signature

These three criteria have to be fulfilled in order to ensure completeness, validity and integrity as well as confidentiality of provenance data. Although there exist different theoretical models that describe secure mechanisms how to protect sensitive data, these models are often not implemented or used.

6.4 SECURE LOGGING PROTOCOLS

Different protocols that enable the storage of provenance and authenticity metadata in a secure way exist. All of these protocols utilize the techniques introduced in Section 6.2 and their applications as described in section 6.3. An overview of current secure logging protocols can be found in [2] by Accorsi et.al. The work by Accorsi contains not only an introduction about the fundamental concepts of secure logging, but also a description of attack models of the logging phases described in section 6.1.1. He provided a classification system of existing approaches and their properties, where he introduced the three classes: syslog, Schneier/Kelsey and Waters et.al. The report from Accorsi et. al. provides an overview of the features the corresponding logging protocol provides. The bandwidth of the mentioned approaches spans from concrete implementations that are widely used (syslog, syslog-ng) via prototypes (BBox, Waters) to concepts (Ohtaki). The only protocol that fulfils all criteria is the system proposed by the author himself. Further work related to secure authenticity and provenance metadata storage, protocol designs and identification of current challenges can be found for instance in [1, 8, 26, 27, 61].

6.4.1 Preservation and Encryption

Digital archives need to be safeguarded from unauthorized access and the digital objects need to be protected from unauthorized manipulations. While encryption is widely used, it is also a major threat to digital preservation. Encryption procedures become obsolete over time which entails two consequences. The first threat is that the encryption method becomes insecure and is then useless. The second issue is that the keys for decryption are unavailable and cannot be used anymore, which means that the content cannot be retrieved again. Both scenarios are serious issues in the area of digital preservation. Similar arguments are valid for digital signatures as well. Methods are needed that enable auditors to verify signatures over the long term and to use flexible encryption protocols that can adapt to current developments in the field. An example for an extensible signature standard is XML Advanced Electronic Signatures (XAdES)[7] . This standard adds extensions to the XML Signature standard XMLDSIG[8] . One of these extensions is called XML Advanced Electronic Signature with eXtended validation data incorporated for the long term (XAdES-X-L), which allows one to verify the integrity of signed data even if the source is lost. This is achieved by embedding the needed certificates as a list to the signature itself. Each new encryption layer can be seen as an envelope that protects the content it encloses. These layers can be successively removed by decrypting them one by one. This matryoshka (Russian nesting doll) principle ensures that encryption can stand the long term and solves the problem of the encryption standard obsolescence.

6.5 REPOSITORY AND LOG FILE AUDITS

Log files are used in order to verify the status of a system and in order to analyse interactions with it. The process of investigating the log files of a machine is denoted as audit. Provenance and authenticity data can only develop their full potential if the corresponding data is checked for integrity and validity on a regular basis. The analysis of recorded metadata is especially necessary if there are reasons to ensure that no manipulation of the provenance and authenticity data has occurred. Audits cover the detailed investigation of the captured data and the detection of abnormalities and attempts at fraud. The role of an auditor is critical and only trustworthy professionals should be entrusted with the task of

[7] http://www.w3.org/TR/XAdES/

[8] http://www.w3.org/TR/2002/REC-xmldsig-core-20020212/

an audit. The way an audit is structured is usually defined within certain policies. These policies regulate which data is relevant for auditors and which data they should investigate. It is also possible to use different encryption keys for different audit paths in a DAG. Syalim et.al. describe the details on the path-based access control in [61]. One fundamental property of audit logs of provenance data is that the records it contains are plausible. The history of the evolution of a document has to be traceable and reproducible. This is especially in regard to sequences of modifications that allow one to verify if these changes occurred in a believable and authentic manner. References to other documents that did not exist when the reference was made are for instance clear evidence that either there was an error or that there has been an attempt of tampering; the first incident being an indicator for the poor quality of data, the second one being an alert for a possible intrusion. It is in many cases not a trivial task to differentiate between an error and a manipulation. This is why in most cases audits involve human investigators although there exist means of detecting abnormalities automatically. Both scenarios are serious and should trigger concern within the organization responsible for an archive. Information about tamper detection in audit logs can be found in the correspondent paper [60] by Snodgrass et.al. Another example on how to model a secure provenance storage can be found in [50] by Perez et.al.. The work from Perez also contains descriptions of the message flows between the components of an audit system. A higher level checklist for auditing archives, as described above is the ISO/DIS 16363:2011: Space Data and Information Transfer Systems – Requirements for Audit and Certification of Trustworthy Digital Repositories (2011) [36]. It covers the security in terms of organizational structures as well as infrastructure and personnel. Log files in particular are mentioned as necessary evidence.

6.5.1 Current Examples of Open Source Archives

Different open source archiving solutions exist, which make use of different provenance models. These systems also produce event logs that keep track of changes within the archive. DSPACE for instance logs events in its log and it also makes use of provenance data (called history), but the log is not explicitly secured. The general architecture of DSPACE does implement a security model[9], but not especially for the generated history files. Nevertheless the discussions on the project Web sites show that in the future secure logging facilities are planned to be implemented[10]. The Fedora Commons Repository Software[11] project follows a similar approach. It also contains a history (called versioning) system that allows one to trace all the changes a digital object has undergone during its lifetime[12]. Fedora not only stores earlier versions of the objects contained in an archive but it also maintains audit trails of the events. This enables auditors to monitor all changes that have been applied to objects from within the Fedora environment. For detecting manipulations that occurred from the outside of the Fedora system, checksums are used. In combination with the versioning system this mechanism allows one to perform audits and to detect internal and external manipulations at object level. Fedora does also not implement a specialized security model for logging.

6.6 SECURE STORAGE

On one hand log files need to be stored in a secure way as they can contain sensitive information about processes and the people involved in their creation and transformations. On the other hand the contained data must be accessible by qualified personnel, as it is important to perform audits when needed. Such storage has to fulfil several requirements. It has to ensure the data integrity and confidentiality and it has to guarantee the accessibility of the data for the long term. Especially long term preservation of provenance and authenticity data is still a challenge. Many of the concepts use cryptographic keys, which need to be kept active and decryptable during the whole life cycle. Examples for specialized provenance stores are the Preservation Data Storage (PDS) and the

[9] 3 https://wiki.duraspace.org/display/DSPACE/SecuringDspace

[10] https://wiki.duraspace.org/display/DSPACE/HistoryDiscussion

[11] http://www.fedora-commons.org/

[12] https://wiki.duraspace.org/display/FEDORA35/Versioning

Provenance-Aware Storage System (PASS) project. PDS is a provenance and authenticity aware storage system developed by IBM [20]. The design is based on the authenticity protocol defined in OAIS [11] and consists of several layers, which support the life cycle of preservation and authenticity metadata. PDS moves high level methods for dealing with authenticity and provenance to the low level storage layer. By this the authors claim to reduce data exposure and thereby increase the security of the stored metadata. PASS as described in [48] by Muniswamy-Reddy et. al. is a modified Linux kernel that transparently keeps track of file access and modifications by building a DAG of provenance data. PASS does not yet implement security features.

Another way of securing log files is the usage of WORM (Write Once Read Many) systems. This system circumvent deletion or manipulation of data by not supporting the necessary file system operations of overwriting, moving or deleting data. This can be achieved either by software or hardware solutions. The industry provides different products. An example for a software solution is the product KOMworx from KOMnetworks[13] . It is a software module that enables the conversion of conventional hard disks into so called eWORM disks that do not allow deletion or manipulation. Systems that prevent data alteration or deletion by specialized hardware are even more secure, as they are very hard to undermine. These disk controllers do not allow one to delete or manipulate data once it has been written, they do not even have the instruction set for the necessary operations implemented. An example for hardware WORM systems is the product SilentCube from FAST LTA[14]. Such specialized storage solutions are very suitable for storing log files and other data in a secure way, as they cannot be modified once they have been written. The safety of data can often be enhanced by using redundant disks, as it is the case with RAID systems. Such a RAID can mirror disks and therefore dramatically increase the reliability of the storage against data loss.

6.7 OUTLOOK AND CONCLUSION

Digital archives and repositories store our knowledge within them and provide us with a huge source of information. This knowledge was generated often with large efforts and in many cases it contains valuable insights, which cannot be reproduced easily anymore. For this reason data repositories must safeguard their precious content from undesired alterations and manipulations. Systems have to implement logging facilities, which allow one to map the complete chain of custody digitally and allow auditors to draw conclusions on the degree of (presumption of) authenticity of data. Metadata demand specialized data structures, which allow one to model the properties of provenance and authenticity. These metadata can sometimes be more sensitive than the data they describe. For instance clinical trials, peer review processes and health or eScience data are often dependent on not displaying the full information of their contributors. Therefore, the captured metadata has to be protected from being read, which is done by using cryptographic methods. It is highly important to be able to demonstrate authorship in a way it can be verified whenever there is a dispute about intellectual property rights between parties. All these cases have to be considered in the area of digital repositories, which makes logging of authenticity and provenance data one of the most fundamental properties of an archive. Without proper knowledge about the evolution of some resource, assumptions about its authenticity are hardly possible. This is also the reason why authenticity and provenance are highly interconnected. This section gave an overview about current logging systems and their security considerations and provided a survey about existing methodologies, approaches and concepts in the same area. Although there exist different proposed solutions to the topic, securing sensitive provenance and authenticity still remains a challenging research topic.

[13] www.komnetworks.com

[14] www.fast-lta.de/en/

7 PROVENANCE INTEROPERABILITY AND REASONING

7.1 PROVENANCE AND INTEROPERABILITY

This section will summarize the contribution given by FORTH, related to the activities of Task 2430-Provenance Interoperability and Reasoning, that have already been published in the internal deliverable ID2401-Report on provenance interoperability and mappings.

7.1.1 Motivation

There are several models for representing **provenance**. The availability of mappings between these models is crucial for the interoperability required to allow us to follow the chain of provenance for any Digital Object, recorded in a variety of ways in a variety of systems over time, as described in section 4. Specifically, the availability of mappings allows the building of tools and systems for exchanging and integrating provenance information; they can be exploited for implementing a materialized integration (warehouse), or a virtual integration (mediator) approach (more in the internal deliverable ID2401).

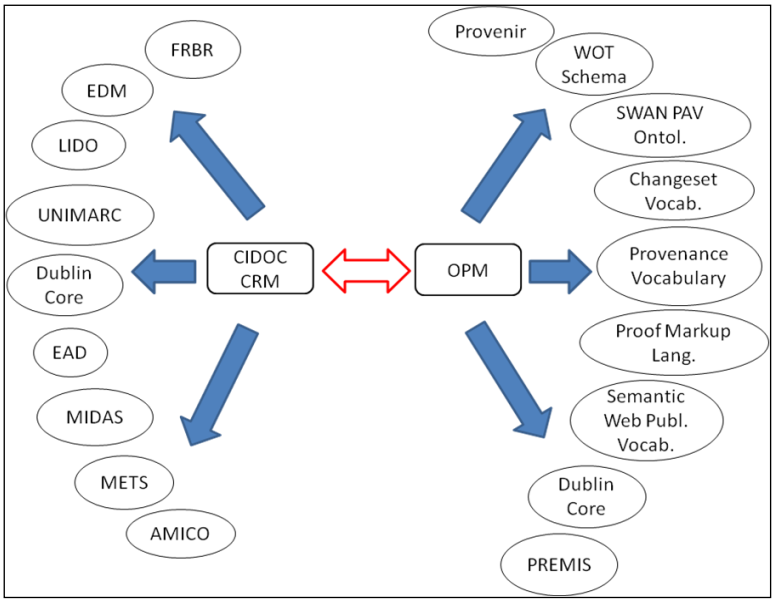

Figure 7.1 - CIDOC CRM's and OPM's mappings

Several mappings have been defined by the W3C Provenance Incubator group: ***Provenir ontology, Provenance Vocabulary, Proof Markup Language, Dublin Core, PREMIS, WOT Schema, SWAN Provenance Ontology, Semantic Web Publishing Vocabulary, Changeset Vocabulary, OPM (Open Provenance Model). OPM*** is used as the reference model.

Another good "hub" is CIDOC CRM (and its extension CRMdig). In brief, CIDOC CRM (ISO 21127:2006) is a core ontology describing the underlying semantics of data schemata and structures from all museum disciplines and archives. It is result of long-term interdisciplinary work and agreement and it has been derived by integrating (in a bottom-up manner) hundreds of metadata schemas. In essence, it is a generic model of recording of "what has happened" in human scale. It can generate huge, meaningful networks of knowledge by a simple abstraction: history as meetings of people, things and information. CRMdig (Theodoridou et al., 2010) is an extension of the CIDOC CRM ontology (ISO 21127:2006) for capturing the requirements of digital objects. Various mappings of CIDOC CRM are available at the following address http://www.cidoc-crm.org/crm_mappings.html.

We have seen that Open Provenance Model (OPM) and CIDOC CRM*dig* are good hubs for provenance models. This justifies the need for establishing mappings between them. OPM is quite

minimal: it comprises only 3 classes (Artifact, Process, Agent) and five associations among them (used, wasGeneratedBy, wasControlledBy, wasTriggeredBy, wasDerivedFrom). On the other hand CIDOC CRM contains **82** classes and **146** properties, while its extension CRMdig currently contains **31** classes and **70** properties.

7.1.2 Results

We have compared Open Provenance Model (OPM) and CIDOC CRMdig and have established mappings between these two models. We should note that the ontology assumed by OPM does not explicitly model the concept of *Event*, a concept that is of prominent importance, not only because events allow the tracing of the history of an object but also because they enable the integration of several pieces of information about an object. Without the notion of event and also of physical objects that are carriers (devices) it is not possible for example, to adequately describe the conditions under which a photograph was taken. Nevertheless, we should say that the way OPM treats *Processes* resembles events (however the corresponding ontological structure of OPM is not rich). In the internal deliverable we have specified **mappings** between OPM to CRMdig and CRMdig to OPM.

7.2 PROVENANCE-BASED INFERENCE RULES

7.2.1 Summary

The capturing, integration and management of provenance information is useful in various domains (and for various reasons), and current workflow systems can produce very large amounts of provenance information. In the context of this task we have introduced provenance-based inference rules as a means to (a) reduce the amount of provenance information that has to be recorded, and (b) to ease their update and quality control. We have motivated this kind of inference and we have identified a number of basic inference rules over CRMdig. In particular, the basic inference rules concern the interplay between (i) actors and carried out activities, (ii) activities and devices that were used in, and (iii) participation of information objects and physical objects to events. However, since a knowledge base is not static but it changes over time for various reasons, a rising question is how we can satisfy update requests while still supporting and respecting the aforementioned provenance-based inference rules. To tackle this problem we have proposed (and explained by examples) two *sets of basic change operations* and detail their application assuming the proposed inference rules.

7.2.2 Motivation and Context

In a naive view, the provenance of a digital object can be seen as a record specific to it of the events and their contexts that have causally contributed or had significant influence on its content. However, digital objects do not undergo "changes" as material items, which sum up to an accumulative effect. Each modification leaves behind the original version, which may or may not be reused in another context. Hence any realistic creation process of digital content gives rise to a set of digital items – temporary or permanent, connected by metadata forming a complex DAG (Directed Acyclic Graph) via the individual processes contributing to it. The provenance "history" of a single item is the DAG of all "upstream" events until the ultimate empirical capture (measurement), simulation S/W run or human creative process. In a production environment, often controlled by a workflow system, there are no clear a priori rules which data item will be permanent. Interactive processes of inspection of intermediate results and manual interventions or changes of processing steps may corrupt any preconceived order of events.

Therefore the only chance to capture reliably the complete provenance is by monitoring the metadata of each step individually, and then concatenating these elementary metadata into the complete provenance history of an item by use of shared URIs. In cases as in empirical 3D model generation, where ten thousands of intermediate files and processes of hundreds of individual manual actions are no rarity, it is prohibitive to register for each item its complete history because of the immense repetition of facts between the files: On one side, the storage space needed would be blown up by

several orders of magnitude, and on the other side any correction of erroneous input would require tracing the huge proliferation graph of this input.

The above notion of redundancy is yet formally not well understood and may even not be strictly logical. For instance, it is a question of convention, if we regard that persons carrying out a process carry out all of its subprocesses. Even if we make this convention, it is impractical for the monitoring system to expand the persons to all subprocesses. Therefore the question is rather, which system of propagation and exception rules would minimize redundancy for the typical statistics of processes under consideration. If such a system has been established, we may distinguish three epistemological situations:

The registered facts can reliably and completely be registered by a monitoring system, such as a workflow shell.

There are facts which users need to input manually to the monitoring system and may be lazy to do so.

Facts come from different monitoring systems or uncontrolled human input.

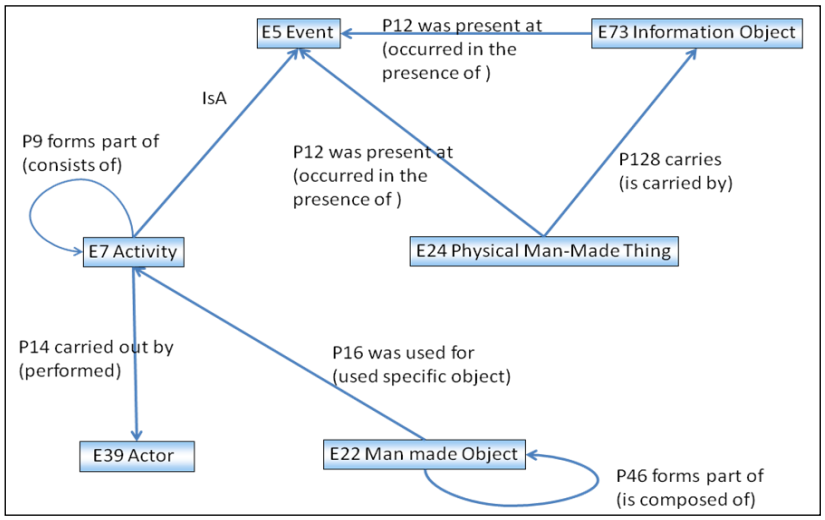

Figure 7.2 - Part of CRM Dig

The reasoning forms we consider in this work aim at the *dynamic completion* (deduction) of facts from original input by resolving transitive closures and propagating the properties.

In case a), we encounter a "Brave" Closed World of knowledge management. We may harmonize the monitoring system with a rule system that describes with minimal input any possible situation and exception. Then, all deductions from these rules will provide results that conform to reality.

In case b), we may also harmonize the monitoring system with a rule system, but include in our reasoning the possible absence of facts.

In case c), we may assume some default behaviour of users and systems, but our deductions will have a probabilistic character.

As an application example, in the context of the IP 3D-COFORM (3D-COFORM (Tools and expertise for 3D collection formation)), CRMdig which was developed in the CASPAR IP (CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval)) has been enhanced to describe in any required detail the very complex data acquisition and data processing processes both on an atomic - processing step by processing step - and on an integrated level - from acquisition to data ready for publishing. Note that a single acquisition process may create thousands of images and some terabytes of data. The complex processes yielding massive intermediate data and multiple versions of final products, reprocessing with improved methods or corrected input, give raise to a need for complex generic reasoning over provenance data in order to solve digital preservation tasks, such as: *propagation of properties from super to sub processes, propagation along processing steps,*

merging metadata of intermediate steps, relevance assessment, obsolescence control, "garbage collection" and appraisal, and others.

In this work we consider CRM Dig as the conceptual model for representing provenance and over this model we identify inference rules. The identified inference rules concern the interplay between (i) actors and carried out activities, (ii) activities and devices that were used in, and (iii) participation of information objects to events and participation of physical objects. We focus on these three rules as they occur frequently in practice. Of course, one could extend this set according to the details and conventions of the application at hand. Figure 7.2 shows one part of the model, specifically the part involved in the inference rules which are introduced at Section 7.3.

Rule Num	Rule Name	Rule Description	Brief Example
R1	ActorCarriedActivity	If a performer has carried out one activity, then he has carried out all of its sub activities.	”STARC-The Cyprus Institute” is the performer of a Laser scanning acquisition activity but also the performer of the detailed sequence of shots which are actually sub activities of the scanning acquisition activity.
R2	UsedPartofObject	If an object was used for an activity, then all parts of the object were used in that activity too	If a multidome camera was used for an event then a lens of it was also used for that event.
R3	InformationObjectPresence	If an information object was present at an event, then a physical object (that carries the information object) was present at that event. (note that an information object must have a physical object that carries that, otherwise it cannot exist).	If we know that a person X1 read a poem Y1 during an event E1, then certainly there was a carrier Z for that poem in that event. If a device was present at an event and uses a piece of software then this software was also present at that event.

Table 7.1

Below we will elaborate on a number of such rules. The inference rules are summarized in the following table. Subsequently we provide examples for each one of them.

Synopsizing these rules concern two binary and one ternary relation:

- carriedOut(Actor, Activity),
- wasUsedFor(Activity,Device), and
- wasPresentAt(InformationObject,Event, PhysicalObject).

Below we discuss each one by an example.

Rule 1: carriedOut(Actor, Activity)

Example: Scientists often use 3D laser scanning in order to construct digital 3D models. These processes involve taking many photographs of the desired model. One example from our data is the following process: **"Laser scanning acquisition of Canoe-shaped vase from Archaeological Museum of Nicosia"** which was carried out by the **"STARC-The Cyprus Institute”**. The process has its own sub processes. For example detailed information of each captured photo is represented by events. The above process can be analysed to: **"Detailed Sequence of shots - Canoe-shaped vase**

from Archaeological Museum of Nicosia". And that process of shots can have the following sub processes:

- "Capture 1_8 for Boat"
- "Capture 1_7 for Boat"
- "Capture photo DSC_0792 for Boat"
- "Capture photo DSC_0791 for Boat"

All the above sub processes have different recorded metadata. On the other hand, the information that the initial actor was the **STARC Institute** is desired to be preserved following the path. Figure 7.3 shows the edges (represented by dotted lines) which are inferred by the rule.

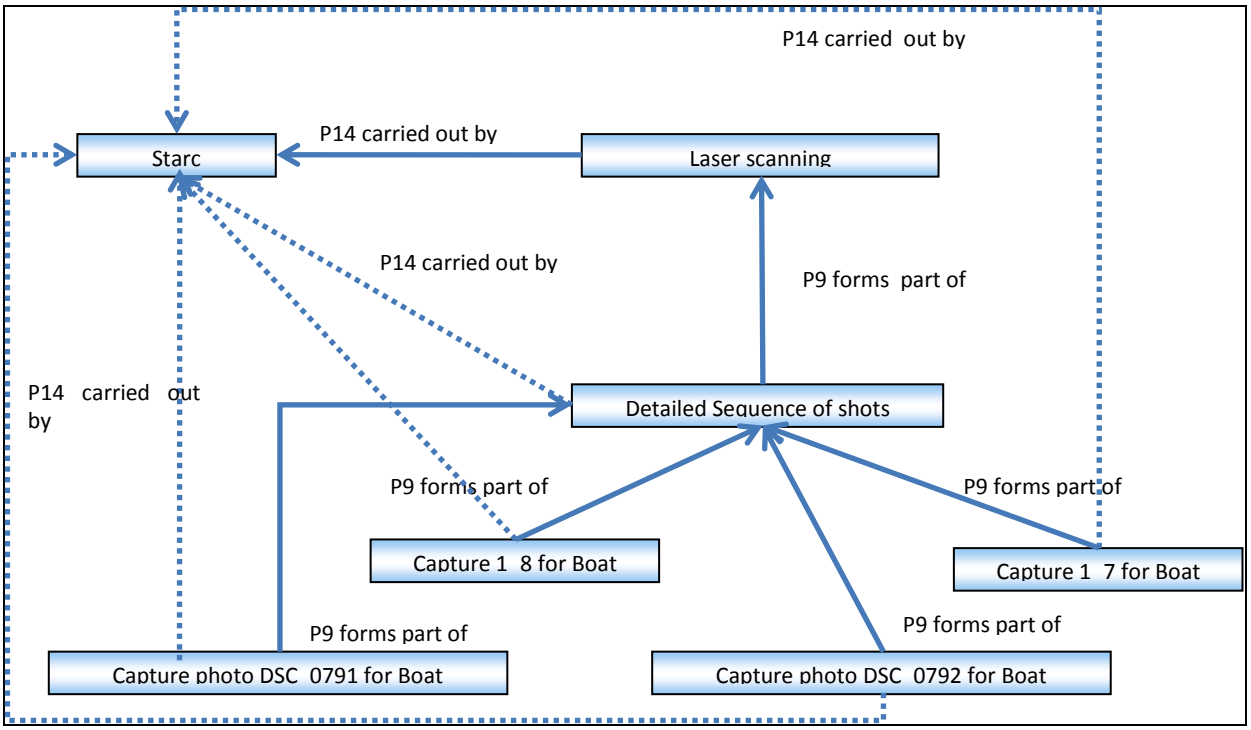

Figure 7.3 - Example of rule R1

Rule 2: wasUsedFor(Activity,Device)

Example: In 3D modelling devices with many cameras, are called *multiviewdome devices*, are usually used. These devices have as parts other cameras or just multiple lighting devices. Figure 7.4 illustrates an indicative modelling of such a setting.

In provenance metadata, there could be a fact stating that a multiviewdome device was used for a particular activity (e.g. in a 3D modelling activity). With R2 we can infer that the constituent devices, in our example the camera Nicon D90, were also used in that activity. Indeed, the multiviewdome device cannot be used without essentially using its parts.

Rule 3: wasPresentAt(InformationObject,Event, PhysicalObject)

Example: 3D reconstruction from images is a common process used in archaeology in order to document, digitize and model archaeological exhibits such as statues. Consider the exhibit shown in Figure 7.5 which is part of a column of Ramesses II located in the Egyptian museum garden in Cairo and is used in the aforementioned process.

The 3D reconstruction process could be modelled as an *event* and the exhibit as a *physical man-made thing.* Moreover that physical thing contains information which is represented by the carved hieroglyphics. That information could be modelled as an *information object.* According to rule R3 if

that part was present in an event then that information was also present in that event. The above modelling is illustrated in Figure 7.6.

Rule R3 infers the presence of information in hieroglyphics at the event of a 3D reconstruction because the part of Ramesses II was also present at that event.

The inference is reasonable because the information was carved in hieroglyphics when the column was built, thus the information in hieroglyphics *coexists* with the part of a column which *carries* it and this coexistence implies their presence at events. As a result, if there is a *carries* relationship between an information object and a physical thing, rule R3 infers the presence of the former in all the events that the latter *was present at*.

Also note that the inference rule can be viewed as a constraint rule, i.e. every "Information Object" that was present at an "Event" must be connected to a "Physical Thing".

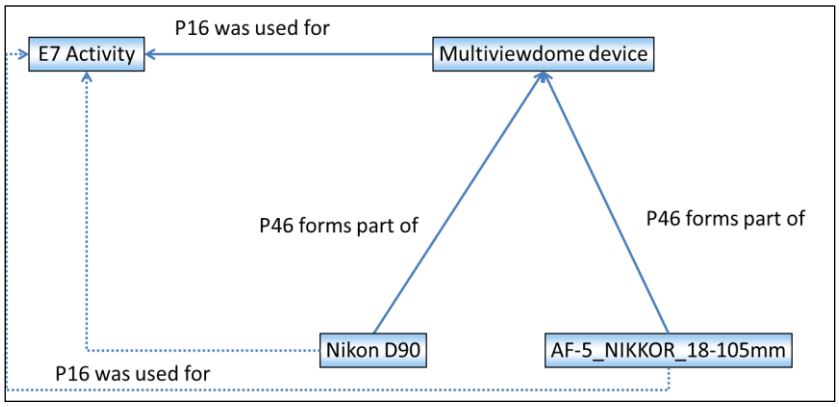

Figure 7.4 - Example of rule R2

Figure 7.5 - Part of a column of Ramesses II

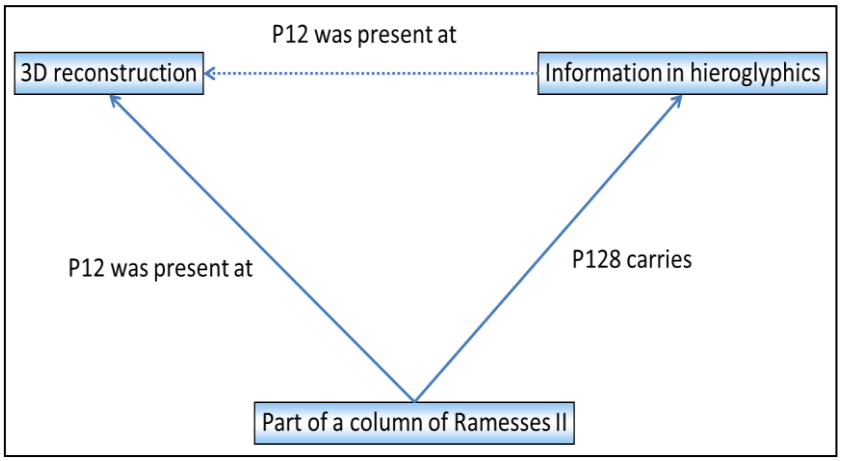

Figure 7.6 - Example of rule R3

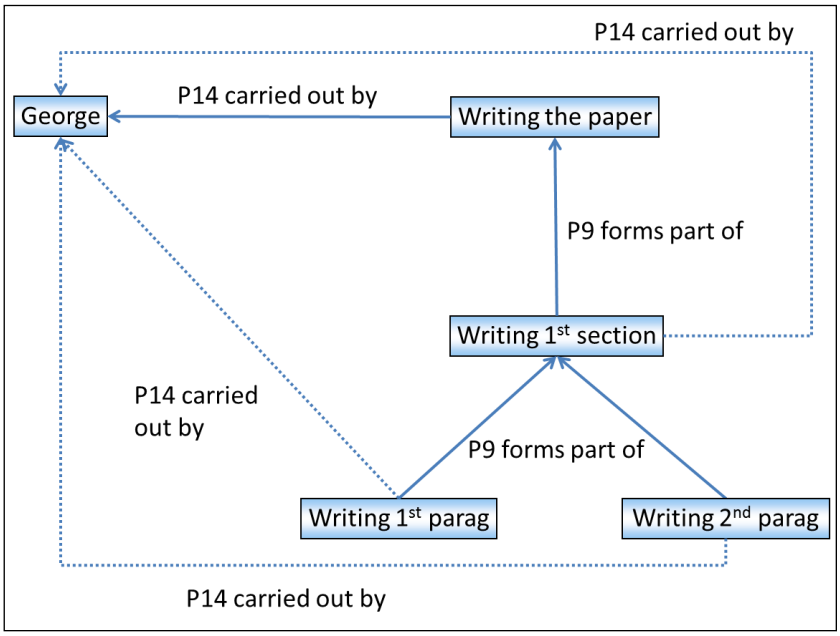

Figure 7.7 - Initial state of the KB

7.3 PROVENANCE INFERENCE RULES AND KNOWLEDGE EVOLUTION

A Knowledge Base (KB) changes over time, i.e. we may have requests for adding, deleting or replacing facts. The question arising is how can we satisfy such requests while still supporting the aforementioned provenance-based inference rules. Below we describe various cases that require special attention and treatment, using a running example. For each update we describe the KB's states (through figures) and explain any inconsistencies that may occur due to the application of the inference rules.

Consider a KB that contains an activity *a1* of writing a paper and several sub activities such as writing the 1st paragraph of section one, and so on. *George* is the performer responsible for *writing the paper* and this association has been "propagated" (due to rule R1) to all sub activities of *a1*. This means that each one of them is related to *George* through the inferred "P14 carried out by" association. The initial state of the KB is demonstrated by Figure 7.7 where inferred associations are illustrated by dotted lines.

Over this example below we shall see examples of four update operations: *addition, disassociation, contraction and replacement.*

7.3.1 Performer creation

Suppose a request for adding a new performer to the sub activity *writing 1st section*, e.g. that *Michael* is also responsible for *writing 1st section*. The update on the DB of Figure 7.7 is shown in Figure 7.8.

We observe that new performer *Michael* has been created and associated with *writing 1st section.* Due to rule R1, he has also been associated with *writing 1st parag* and *writing 2nd parag* which is reasonable. There is not any ambiguous situation after this update request.

7.3.2 Performer disassociation and performer contraction

Now consider an update request saying that *George* is not responsible for *writing the 1st paragraph*. The rising question is whether *George* is not responsible only for *writing 1st parag, or* also for other activities. For this request we can distinguish two cases:

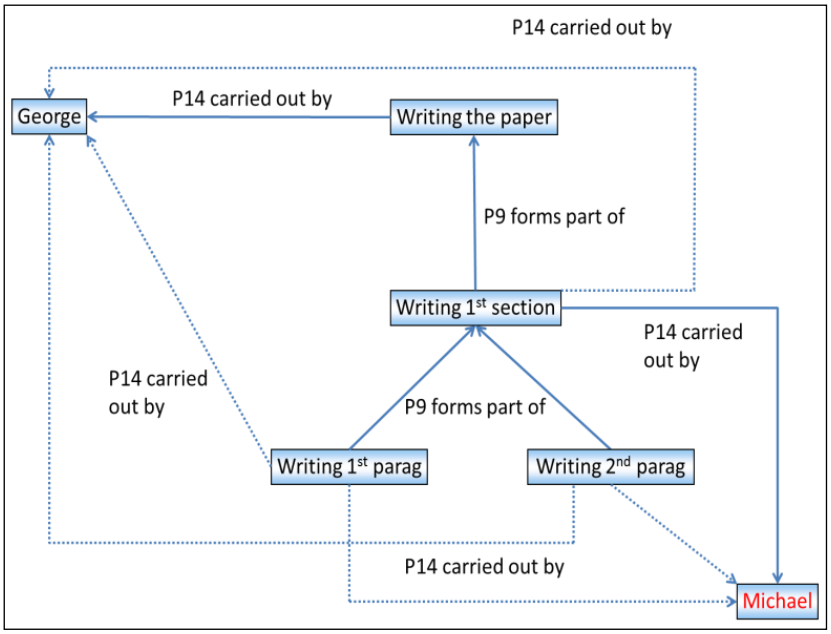

Figure 7.8 - State of the KB after the addition

Disassociation. The non-responsibility of *George* about the *writing of 1st paragraph* implies some uncertainty about his responsibility for other related activities (e.g. was he responsible for the 2nd paragraph?). Since we cannot explicitly represent facts which are not necessarily true, all such associations must also be deleted, i.e. we should delete the following:

(Writing 1st parag, carried out by, George) // as request

(Writing 2nd parag, carried out by, George)

(Writing 1st section, carried out by, George)

(Writing the paper, carried out by, George)

Contraction. There is a high degree of certainty that the non-responsibility of *George* is only for *writing 1st parag*. However, there might be other activities which are still associated with *George*, such as *writing 2nd parag*. In this case, these associations must be preserved. In this case we have to delete only the following:

(Writing 1st parag, carried out by, George) // as requested

(Writing 1st section, carried out by, George)

(Writing the paper, carried out by, George)

We will be referring to the above cases as *performer disassociation* and *performer contraction* respectively.

Notice that in comparison to disassociation, the contraction preserves the association (Writing 2nd parag, carried out by, George).

After each such operation (either disassociation or contraction), the inference rules should be applied (in our case rule R1). Notice that both approaches delete the association (Writing the paper, carried out by, George). If that association were not deleted, then rule R1 would again create the association (Writing 1st parag, carried out by, George) as illustrated in Figures 7.9 and 7.10. This justifies the behaviour of the above operations.

Also note that even if we did not apply R1 after a disassociation/contraction operation, the association (Writing 1st section, carried out by, George) would remain (due to past applications of R1). This means that it not the application of R1 that causes the problem.

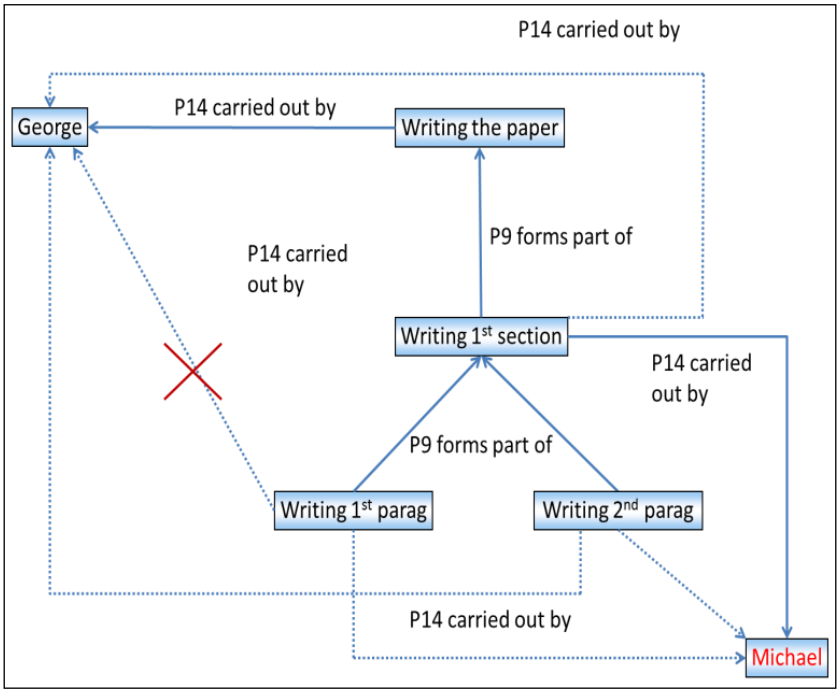

Figure 7.9 - Deletion of an association

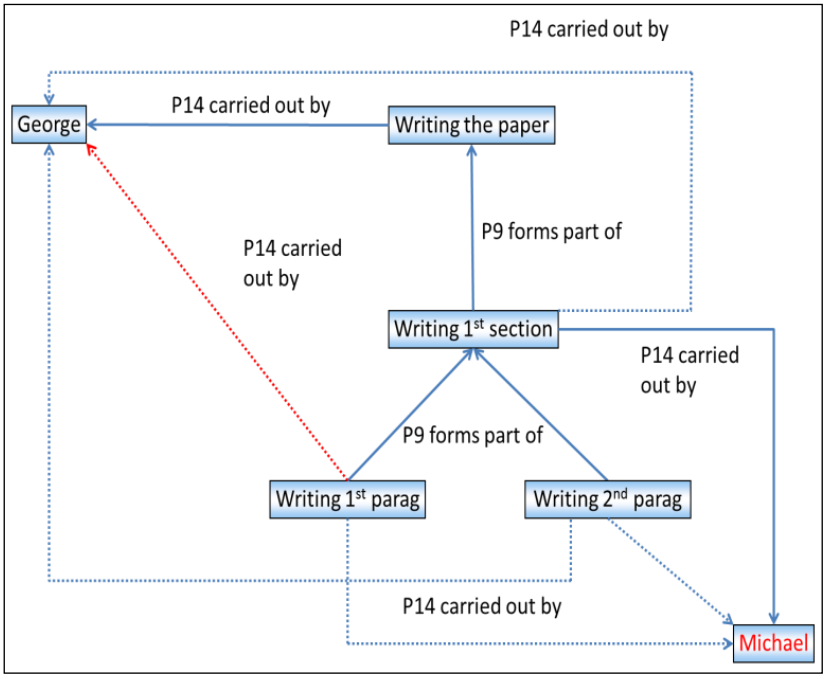

Figure 7.10 - State of the KB after the deletion

7.3.3 Performer replacement

Suppose that we acquire the information that *John* instead of *George* is responsible for *writing 1st parag*, meaning that *George* should be replaced by *John*.

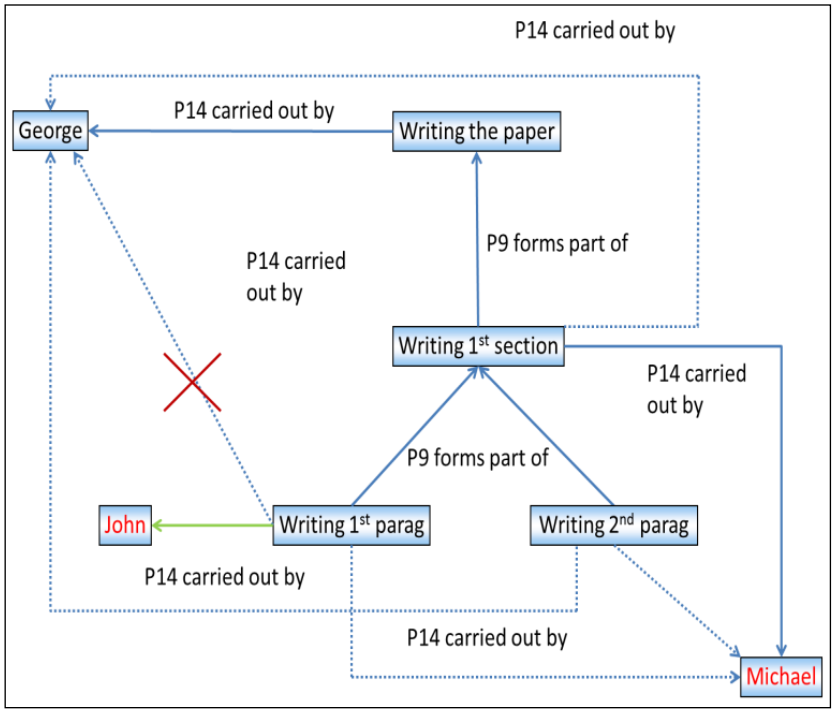

Figure 7.11 - Performer replacement

A performer, who is not responsible for an activity, is also not responsible for its super activities. Therefore, the replacement of a performer implies the deletion of "P14 carried out by" not only from the requested activity but also from its super activities (in our case *writing 1st section, writing the*

paper). However, some associations from other activities with the replaced performer must be preserved, in our case the association (Writing 2nd parag, carried out by, George).

The default plan of actions for this replacement would be:

a) deletion of the association between *writing 1st parag*,
b) creation of a new performer *John*
c) addition of a new association between *John* and the referred activity and
d) the application of rule R1.

One problem of this default action plan is that the deleted association between *writing 1st parag* and *George* will be created again by rule R1. It will be inferred because there is still an existing one between *writing the paper* and *George*. As a consequence, this "propagation" will cancel the deletion. The situation is illustrated in Figures 7.11 and 7.12.

What would be reasonable to do, would be the deletion of the following associations:

(Writing 1st parag, carried out by, George) // as requested

(Writing 1st section, carried out by, George)

(Writing the paper, carried out by, George)

and the addition of the following association:

(Writing 1st parag, carried out by, John) // as requested

We can observe that this would be the result of an addition and a contraction (as demonstrated earlier). As we shall see below this holds in general, i.e. **a *replacement* operation can be simulated by an *addition* and a *contraction.***

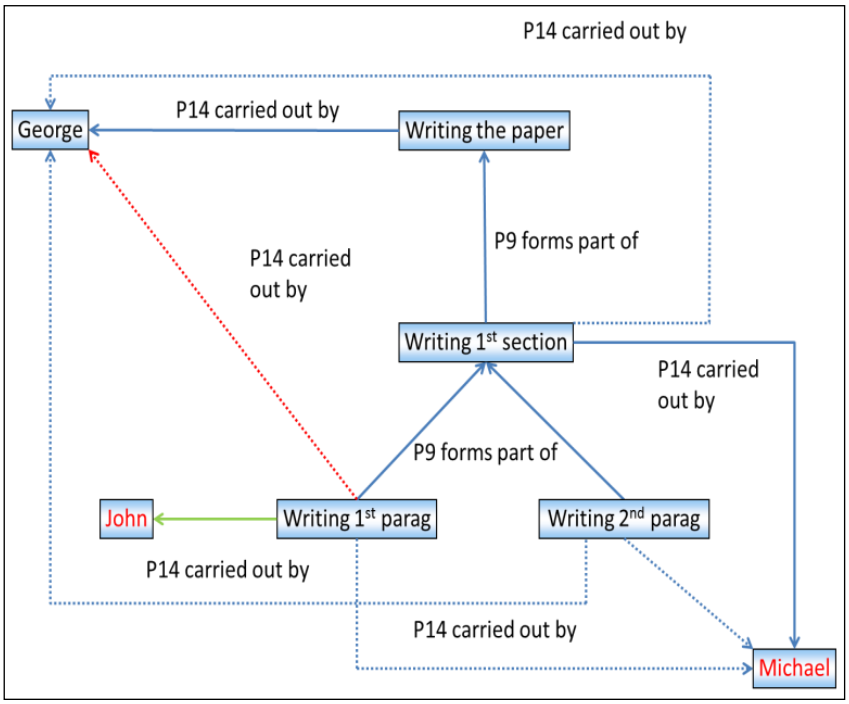

Figure 7.12 - State of the KB after the replacement

7.3.4 Basic Sets of Change Operations

It follows from the above that we can distinguish *two* sets of basic change operations:

Set A: Add, Disassociate, Replace

Set B: Add, Disassociate, Contract

If one of these two sets is supported, then the change requests can be tackled. More details, as well as a case-based definition of the required change operations, are given in the internal deliverable ID2401.

7.3.5 Conclusion

We motivated the need for provenance-based inference rules (for reducing the amount of provenance information that has to be stored, and to ease quality control), and we identified three basic rules accompanied by real world examples. These rules involve classes that are found in almost any provenance model. However the use of inference rules introduces difficulties with respect to the evolution of knowledge. We elaborated on these difficulties and described how we can address this problem. We identified two ways to deal with deletions in this context, based on the philosophical stance against explicit (ingested) knowledge and implicit (inferred) ones (foundational and coherence semantics). Based on these ideas, we specified a number of update operations that allow knowledge updating under said inference rules. Although we confined ourselves to CRMdig, and to three specific inference rules, the general ideas behind our work (including the discrimination between foundational and coherence semantics of deletion) can be applied to other models and/or sets of inference rules. Finally we described implementation policies of inference rules and change operations over the various existing technologies (RDF triple stores, rule engines and query languages (they are provided in ID2401).

8 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; c) ability to exchange and integrate provenance information by exploiting mappings; d) novel techniques that employ inference for reducing the amount of provenance information that have to be kept stored and for making their correction easier.
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) - ISO standards for trusted digital repositories (ISO 16363) - Premis as common dictionary - CIDOC CRM which is an ISO standard, and to OPM which is promoted by W3C
WP22 Identifiers and Citability	a) 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. b) The entities that participate in provenance graphs (information carriers and objects, actors, events, activities, etc.) certainly need an identification mechanism. This issue becomes indispensable in distributed settings (distributed production processes).
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 artifacts. The models here discussed can be adopted for that need.
WP31 Digital Rights and Access Management (M27-M38)	Provenance and authenticity are a crucial aspect of digital rights; therefore the models discussed and their mappings are strongly related to this.
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.

9 INTEGRATION AND OUTREACH

In planning and carrying out the activities of WP 24, that are documented in this deliverable and in the companion deliverable D24.2 *Implementation and testing of an authenticity protocol on a specific domain*, 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 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 takes place 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 to successfully transfer 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 items 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 different meanings to this concept. 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 D24.2, 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, STFC, 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 to incorporate, in cooperation with SCIDIP-ES project which is part of the worldwide Earth Science Long Term Data Preservation program, our methodology into the SCIDIP-ES Authenticity Toolkit, which is part of the services and tools that the project proposes to implement preservation for all types of digitally encoded information, with specific testing for Earth Science data.

10 CONCLUSIONS

In this deliverable we have reported the main results of the activities carried out within tasks 2410 Review of authenticity systems, 2420 Evaluation of authenticity evidence and 2430 Provenance interoperability and reasoning.

As regards **authenticity**, the report contains both a quite detailed state of the art and the proposal of *operational guidelines* to be used in gathering, managing and preserving authenticity evidence through the *Digital Resource (DR)* lifecycle, in order to allow the interoperability required to support changes in data holders and processing workflows.

We have analysed the results of the main international projects in the field, which are detailed in the appendix, as well as the recommendations, standards and guidelines for keeping and preserving DRs. This conceptual and methodological background has provided a solid basis and a good starting point to shift the investigation towards a more practical ground, aiming to fill the gap that still divides the mostly theoretical results of the scientific community from the actual practices carried out in most repositories.

As generally acknowledged by all the relevant international projects, the authenticity assessment cannot be limited to a final verification of the bit-stream integrity, but requires a series of interrelated controls to be carefully performed and documented on a systematic basis along the whole DR lifecycle. In other words, one must be able to trace back, along the whole extent of its lifecycle since its creation, all the transformations the DR has undergone and that may have affected its authenticity and provenance. For each of these transformations one needs to collect and preserve the appropriate evidence that would allow, at a later time, to make the assessment.

According to this principle we have proposed a model of the DR lifecycle based on identifying the events that impact on authenticity and provenance, and on defining for each of them the evidence that should be gathered and preserved in order to conveniently document the history of the DR. The model introduces the concepts of *Authenticity Evidence Record (AER)* that is the set of authenticity evidence items that must be gathered in connection with a given event, and of *Authenticity Evidence History (AEH)*, that is the incremental sequence of AERs that is collected when the DR progresses along its lifecycle. The model may also constitute an important step in the direction of interoperability, since, as a consequence of the changes of custody along the lifecycle, the authenticity evidence needs to be managed and interpreted by systems which may be different from the ones that gathered it, and therefore it needs to comply with a common standard, based on shared terminology and on a consistent cross-domain framework of actions and procedures.

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.

To formalize the policy we have resorted to the important concept of *Authenticity Protocol (AP)*, that has been introduced by the CASPAR project, and that we have extended and adapted to our purposes. In the CASPAR definition the *Authenticity Protocol Execution Evaluation (APEE)* is executed at a single given moment of the lifecycle and takes into account all the information that is available at that moment in order to perform an authenticity assessment. According to our philosophy, authenticity management which enables the APEE should be a *structured* and *distributed* process, based on the principle of performing controls and collecting authenticity evidence along the whole DR lifecycle. Therefore, in our definition, an AP is 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.

The model and the guidelines that we have proposed have been successfully put to test on several experimental environments provided by the APARSEN project partners. These case studies, which are documented in the companion deliverable D24.2, provided important feedback and have proved on the field the substantial robustness of our proposal.

Regarding **provenance** there are several models for representing it. The availability of *mappings* between these models is crucial for *interoperability* and allows building tools and systems for exchanging and integrating provenance information. We discussed such models, the mappings that exist between them, and defined a mapping between OPM and CIDOC CRM (and its extension CRMdig) since both are good hubs.

Subsequently, we introduced provenance-based inference rules for reducing the amount of provenance information that has to be stored, and to ease quality control. In particular we identified three basic inference rules accompanied by real world examples. These rules involve classes that are found in almost any provenance model. However the use of inference rules introduces difficulties with respect to the evolution of knowledge. We elaborated on these difficulties and described how we can address this problem. We identified two ways to deal with *deletions* in this context, based on the philosophical stance against explicit (ingested) knowledge and implicit (inferred) ones (*foundational* and *coherence* semantics).

Based on these ideas, we specified a number of update operations that allow knowledge updating under said inference rules. Although we confined ourselves to CRMdig, and to three specific inference rules, the general ideas behind our work (including the discrimination between foundational and coherence semantics of deletion) can be applied to other models and/or sets of inference rules. Finally we described implementation policies of inference rules and change operations over the various existing technologies (RDF/S triple stores, rule engines and query languages).

REFERENCES

1. Accorsi, R.: Log data as digital evidence: What secure logging protocols have to offer? In: Sheikh Iqbal A., Bertino, E., Chang, C.K., Getov, V., Liu, L., Ming, H., Subramanyan R., (eds) COMPSAC (2), pp 398–403. IEEE Computer Society (2009)

2. Accorsi, R.: Safekeeping digital evidence with secure logging protocols: State of the art and challenges. (2009)

3. Ambacher B.: Government Archives and the Digital Repository Audit Checklist. Journal of digital information, 8, 2 (2007)

http://journals.tdl.org/jodi/article/view/190 [visited Jan 2012]

4. APARSEN Project: Deliverable 24.2. Implementation and testing of an Authenticity Protocol on a Specific Domain. (2012)

5. Authenticity Task Force, Appendix 2: Requirements for Assessing and Maintaining the Authenticity of Electronic Records. In: The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project, Luciana Duranti, (ed.), San Miniato, Italy, Archilab, pp 204–219 (2005)

http://www.interpares.org/book/interpares_book_k_app02.pdf [visited Jan 2012]

6. Becker C., Rauber A.: Decision criteria in digital preservation: What to measure and how. Journal of the American Society for Information Science and Technology (JASIST), 62(6), pp. 1009-1028 (2011)

http://onlinelibrary.wiley.com/doi/10.1002/asi.21527/abstract [visited Jan 2012]

7. Becker C., Kulovits H., Guttenbrunner M., Strodl S., Rauber A., Hofman H.: Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. International Journal on Digital Libraries (IJDL), 1(4), (2009)

http://www.springerlink.com/content/801685k478136425/ [visited Jan 2012]

8. Braun, U., Shinnar, A., Seltzer, M.: Securing provenance. In: Proc. of the 3rd USENIX Workshop on Hot topics in Security (HotSec) (2008)

9. Bresslau H.: Handbuch der Urkundenlehre für Deutschland und Italie. Walter de Gruyter, Leipzig (1931)

10. CLIR: Authenticity in a Digital Environment, Council on Library and Information Resources, Pub92 (2000)

http://www.clir.org/pubs/abstract/pub92abst.html [visited Jan 2012]

11. CCSDS: Reference Model for an Archival Information System – OAIS. Draft Recommended Standard, 650.0-P-1.1 (Pink Book), Issue 1.1 (2009)

http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/CCSDSAgency.aspx [visited Jan 2012]

12. Crosby, S.A., Wallach, D.S.. Efficient data structures for tamper-evident logging. In: Proc. of the 18th conference on USENIX security symposium, SSYM'09, Berkeley, CA, USA, pp 317–334. USENIX Association (2009)

13. Dale, R.L., Ambacher, B.: Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist. OCLC Online Computer Library Center, (2007)

http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf [visited Jan 2012]

14. DCC, DPE, NESTOR, CRL: CORE Requirements for Digital Archives. Chicago (2007)

15. Doerr, M.: The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI Magazine, 24(3), 75-92 (2003)

16. Dollar C.: Authentic Electronic Records: Strategies for Long-Term Access. Cohasset Associates, Chicago (1999)

17. Duranti L., Preston R. (eds): International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records. Associazione nazionale archivistica italiana, Padova (2008)

http://www.interpares.org/ip2/book.cfm [visited Jan 2012]

18. Duranti L., Suderman J., Todd M.: InterPARES 2 Project, Policy Cross-domain Task Force. In: InterPARES 2 Project Book: Appendix 19, Duranti L., Preston R. (eds): International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records, Associazione nazionale archivistica italiana, Padova (2008)

http://www.interpares.org/ip2/book.cfm [visited Jan 2012]

19. Duranti L., Thibodeau K.: The Concept of Record in Interactive, Experiential and Dynamic Environments: the View of InterPARES. Archival Science 6 (1) 13-68 (2006)

http://www.interpares.org/ip2/display_file.cfm?doc=ip2_book_appendix_02.pdf [visited Jan 2012]

20. Factor, M., Henis, E., Naor, D., Rabinovici-Cohen, S., Reshef, P., Ronen, S., Michetti, G., Guercio, M.: Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage. In: Proc. of the First Workshop on the Theory and Practice of Provenance, TaPP '09, San Francisco (2009)

http://www.usenix.org/event/tapp09/tech/full_papers/factor/factor.pdf [visited Jan 2012]

21. Giaretta, D.: Advanced Digital Preservation (specifically chapt. 13 and sect. 17.11). Springer-Verlag, Berlin-Heidelberg (2011)

22. Giaretta D., Matthews B., Bicarregui J., Lambert S., Guercio M., Michetti G., Sawyer D., Significant Properties, Authenticity, Provenance, Representation Information and OAIS, iPRES 2009, The Sixth International Conference on the Preservation of Digital Objects: Proceedings, California Digital Library, 67-73 (2009)

23. Guercio M.: Conceptual Framework and Chain of Custody for Sustaining the Digital trustworthiness. In: Perspectives on Metadata. Digital edition & preservation. Vienna (2009)

https://fedora.phaidra.univie.ac.at/fedora/get/o:45908/bdef:Asset/view [visited Jan 2012]

24. Guercio M., Barthelemy J., Bonard A.: Authenticity Issue in Performing Arts Using Live Electronics, (2008)

http://articles.ircam.fr/textes/Guercio07b/index.pdf [visited Jan 2012]

25. Guenther R., Wolfe R.: Integrating Metadata Standards to Support Long-Term Preservation of Digital Assets: Developing Best Practices for Expressing Preservation Metadata in a Container Format. In: Proc. of iPRES 2009, the Sixth International Conference on Preservation of Digital Objects (2009)

http://escholarship.org/uc/item/0s38n5w4#page-2 [visited Jan 2012]

26. Hasan, R., Sion, R., Winslett, M.: Introducing secure provenance: problems and challenges. In Proc. of the 2007 ACM workshop on Storage security and survivability, StorageSS '07, pp 13–18, New York, NY, USA. ACM Press (2007)

27. Hasan, R., Sion, R., Winslett, M.: Secure provenance: Protecting the genealogy of bits. ;login: the USENIX magazine, 3 (2009)

28. Hedstrom M., Lee C.: Significant Properties of Digital Objects: Definitions, Applications, Implications. Proceedings of the DLM-Forum, pp. 218-223 (2002)

http://www.ils.unc.edu/callee/sigprops_dlm2002.pdf [visited Jan 2012]

29. Interagency Science Working Group, National Archives and Records Administration (NARA): Establishing Trustworthy Digital Repositories: a Discussion Guide Based on the ISO Open Archival Information System (OAIS) Standard Reference Model (2011)

http://www.archives.gov/records-mgmt/toolkit/pdf/ID373.pdf [visited Jan 2012]

30. InterPARES: Creator Guidelines. Making and Maintaining digital Materials: Guidelines for Individuals; Preserver Guidelines. Preserving Digital Records: Guidelines for Organizations, Appendix 21. In: InterPARES 2, International Research on Permanent Authentic Records in Electronic Systems 2, Experiential, Interactive and Dynamic Records, Duranti L., Preston R. (eds.), pp. 685-698, 699-712 (2009)

http://www.interpares.org/ip2/display_file.cfm?doc=ip2_book_part_5_modeling_task_force.pdf [visited Jan 2012]

31. ISO 14721:2003 - Space Data and Information transfer Systems – Open Archival Information System – Reference Model (2003)
http://public.ccsds.org/publications/archive/650x0b1.PDF [visited Jan 2012]
At the time of writing the revised version is available at
http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf
or elsewhere on the CCSDS web site http://www.ccsds.org
32. ISO 15489-1: 2001 Information and Documentation – Records Management. Part 1: General (2001)
33. ISO 23081-1:2006 Information and Documentation – Records Management Processes – Metadata for records – Part 1: Principles (2006)
34. ISO 23081-2:2009 - Information and Documentation – Managing Metadata for Records – Part 2: Conceptual and Implementation Issues (2009)
35. ISO 23081-2:2011 - Information and Documentation – Managing Metadata for Records – Part 3: Self-Assessment Method (2011)
36. ISO/DIS 16363:2011: Space Data and Information Transfer Systems – Requirements for Audit and Certification of Trustworthy Digital Repositories (2011)
http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206521R1/Attachments/652x1r1.pdf [visited Jan 2012]
37. ISO/DIS 16919 Space data and information transfer systems - Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories (2011)
38. ISO/IEC 17021:2006: Conformity assessment – Requirements for Bodies Providing Audit and Certification of Management Systems (2006)
39. ISO/IEC 27002:2005 - Information technology, Security techniques - Code of practice for information security management (2005)
40. ISO/IEC 27001:2005 - Information technology, Security techniques - Information security management systems - Requirements (2005)
41. Kent, K., Souppaya, M.: Guide to computer security log management: recommendations of the National Institute of Standards and Technology. NIST special publication 800-92 edition (2006)
http://csrc.nist.gov/publications/nistpubs/800-92/SP800-92.pdf [visited Jan 2012]
42. Knight, G., Framework for the Definition of Significant Properties, (2008)
http://www.significantproperties.org.uk [visited Jan 2012]
43. McNeil H.: Trusting Records. Legal, Historical and Diplomatic Perspectives. Kluwer Academic Publishers, Dordrecht (2000)
44. Moreau, L. et al.: The Open Provenance Model core specification (v1.1). Future Generation Computer Systems, 27(6), 743-56 (2011)
45. Moreau, L., Missier, P.: The PROV Data Model and Abstract Syntax Notation (2011)
http://www.w3.org/TR/2011/WD-prov-dm-20111018/ [Visited Jan 2012].
46. MoReq2 – Model Requirements for the Management of Electronic Records (2008)
http://www.moreq2.eu [visited Jan 2012]
47. MoReq2010 - Modular Requirements for Records Systems (2011)
http://moreq2010.eu [visited Jan 2012]
48. Muniswamy-Reddy, K.K., Holland, D.A., Braun, U., Seltzer, M.: Provenance-aware storage systems. In: Proc. of USENIX '06 Annual Technical Conference, Berkeley, pp 4–4. USENIX Association (2006)
49. OCLC/RLG Working Group on preservation metadata: Preservation metadata and the OAIS Information Model: A metadata framework to support the preservation of digital objects, Dublin (Ohio, USA). OCLC Online Computer Library Inc. (2002)
http://www.oclc.org/research/activities/past/orprojects/pmwg/pm_framework.pdf [visited Jan 2012]
50. Perez, R., Moreau, L.: Securing Provenance-based Audits. LNCS 6378. Springer-Verlag, Berlin-Heidelberg (2010)

51. PREMIS Editorial Committee: PREMIS Data Dictionary Version 2.1 (2011)
http://www.loc.gov/standards/premis/v2/premis-report-2-1.pdf [visited Jan 2012]
52. PREMIS Working Group: Preservation Metadata: Implementation Strategies (PREMIS): Data Dictionary for Preservation Metadata: Final Report. On-Line Computer Library Center and Research Libraries Group, 237 (2005)
http://www.oclc.org/research/projects/pmwg/premis-final.pdf [visited Jan 2012]
53. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Comm. ACM, 21,120–126 (1978)
54. RLG-NARA Task Force on Digital Repository Certification: Audit Checklist for Certifying Digital Repositories (2004)
55. RLG/OCLC Working Group on Digital Archive Attributes: Trusted Digital Repositories. Attributes and Responsibilities. RLG-OCLC Report, RLG, Mountain View (2002)
www.rlg.org/en/pdfs/repositories.pdf. [visited Jan 2012]
56. Ross S., McHugh A.: Audit and Certification of Digital Repositories: Creating a Mandate for the Digital Curation Centre (DCC). RLG Diginews, 9, 5, (2005)
www.rlg.org/en/page.php?Page_ID=20793#article1 [visited Jan 2012]
57. Schneier, B., Kelsey, J.: Secure audit logs to support computer forensics. ACM Trans. Inf. Syst. Secur., 2, 159–176 (1999)
58. Sedona Conference Working group: The Sedona Principles. Best Practice Guidelines and Commentary for Managing Information and Records in the Electronic Age (2007)
http://www.thesedonaconference.org/dltForm?did=Guidelines.pdf [visited Jan 2012]
59. Sedona Conference Working group: The Sedona Principles. Best Practices, Recommendations and Principles for Addressing Electronic Document Discovery (2007)
http://www.thesedonaconference.org/dltForm?did=Guidelines.pdf [visited Jan 2012]
60. Snodgrass, R.T., Yao, S.S., Collberg, C.: Tamper detection in audit logs. In: Proc. of the Thirtieth international conference on Very Large Data Bases - Volume 30, VLDB '04, pages 504–515. VLDB Endowment (2004)
61. Syalim, A., Nishide, A., Sakurai, K.: Preserving integrity and confidentiality of a directed acyclic graph model of provenance. In: Proc. of the 24th annual IFIP WG 11.3 working conference on Data and applications security and privacy, DBSec'10, pp 311–318 Springer-Verlag, Berlin-Heidelberg (2010)
62. Theodoridou, M. et al., 2010: Modeling and querying provenance by extending CIDOC CRM. Distributed and Parallel Databases, 27(2), 169-210 (2010)
63. Thibodeau K. et al.: Part Three – Trusting to Time: Preserving Authentic Records in the Long Term: Preservation Task Force Report, The Long-term Preservation of Authentic Electronic Records. Findings of the InterPARES Project, Duranti, L. (ed), Archilab, San Miniato, Italy, pp. 99–116 (2005)
http://www.interpares.org/book/interpares_book_f_part3.pdf [visited Jan 2012]
64. UN/CEFACT, Business Requirements Specification BRS – Transfer of Digital Records (2008)
www.unece.org/cefact/brs/BRS_TransferOfDigitalRecords_V1.0.pdf [visited Jan 2012]
65. Yavuz, A., Ning, P., Reiter, M.: Efficient, compromise resilient and append-only cryptographic schemes for secure audit logging. Technical report (2011)
http://www4.ncsu.edu/~aayavuz/YavuzNingReiterLogFAS11.pdf [visited Jan 2012]

APPENDIX - RESEARCH PROJECTS

The research projects listed in this appendix have been analysed with specific attention to issues relevant for the deliverable 24.1. The presentation of projects' research goals, expected results and references is based on this perspective.

ARCOMEM

Project title: ARCOMEM (Collect-All ARchives to COmmunity MEMories)

Website: http://www.arcomem.eu

Research goals:

The project started in January 2011 with two declared aims:

- to help transforming archives into collective memories that are more tightly integrated with their community of users,
- to exploit Social Web and the wisdom of crowds to make Web archiving a more selective and meaning-based process.

Applications for use cases have been selected in two domains: TV (media) and political debate, according to the strong evidence (from tools like Icerocket, BlogPulse or Technorati) that these topics generate lots of conversations.

Project terms: 2011 - 2013

Main results:

- innovative models and tools for social web driven content appraisal and selection, and intelligent content acquisition,
- novel methods for social web analysis, web crawling and mining, event and topic detection and consolidation, and multimedia content mining,
- reusable components for archive enrichment and contextualization,
- two complementary example applications (media-related web archives and political archives),
- a standards-oriented ARCOMEM demonstration system.

Partners:

- University of Sheffield, UK
- Internet Memory Foundation
- University of Southampton, UK
- Athena Research and Innovation Center in ICKT
- Institut Télécom, Télécom ParisTech
- Deutsche Welle, Germany
- Südwestrundfunk, Germany
- Yahoo! Iberia, Spain
- L3S Research Center, Germany
- Hellenic Parliament
- Austrian Parliament
- Athens Technology Center SA

References: not available at the moment

CASPAR

Project title: CASPAR - Cultural, Artistic, and Scientific knowledge for Preservation, Access and Retrieval

Website: www.casparpreserves.eu

Research goals:

- to implement, extend, and validate the OAIS reference model (ISO:14721:2003),
- to enhance the techniques for capturing representation information and other preservation related information for content objects,
- to design virtualisation services supporting long term digital resource preservation, despite changes in the underlying computing (hardware and software) and storage systems, and the

designated communities,
- to integrate digital rights management, authentication, and accreditation as standard features of CASPAR,
- to investigate more sophisticated access to and use of preserved digital resources including intuitive query and browsing mechanisms,
- to develop case studies to validate the CASPAR approach to digital resource preservation across different user communities and assess the conditions for a successful replication,
- to actively contribute to the relevant standardisation activities in areas addressed by CASPAR,
- to raise awareness about the critical importance of digital preservation among the relevant user-communities and facilitate the emergence of a more diverse offer of systems and services for preservation of digital resources.

Project terms: 2006-2009

Main results:

- CASPAR conceptual model,
- Authenticity management protocol based on the position paper on authenticity,
- Overall component architecture and prototypes for services and toolkits.

Partners:

- Science and Technology Facilities Council, UK
- European Space Agency, ESRIN, Italy
- University of Glasgow, Humanities Adv. Technology and Information Institute, UK
- Università di Urbino, Istituto di studi per la tutela dei beni archivistici e librari, Italy
- UNESCO
- Advanced Computer Systems S.p.A., Italy
- Asemantics S.r.l., Italy
- IBM Haifa Research Laboratory, Israel
- Consiglio Nazionale delle Ricerche – Institute of Information Science and Technologies, Italy
- Metaware S.p.A., Italy
- Institut National de l'Audiovisuel, France
- University of Leeds, Interdisciplinary Centre for Scientific Research in Music, UK
- Engineering Ingegneria Informatica S.p.A., Italy
- Foundation for Research and Technology - Hellas, Greece
- Centre National de la Recherche Scientifique, France
- Institut de Recherche et Coordination Acoustique/Musique, France
- International Centre for Art and New Technologies - Czech Republic

References:

- CASPAR Conceptual model (CASPAR-D1201-TN-0101-1_0), 2007 (http://www.alliancepermanentaccess.org/index.php/practices/member-resources/documents-and-downloads/?did=18)
- CASPAR Position Paper on Authenticity, 2008 (http://aparsen.digitalpreservation.eu/bin/view/Main/ApanWp24)
- Overall Component Architecture and Component Model (CASPAR-D1301-TN-0101-1_1), 2007 (www.CASPARpreserves.eu/Members/cclrc/Deliverables/CASPAR-overall-component-architecture-and-component-model-1/at_download/CASPAR-D1301-TN-0101-1_1.pdf)
- Prototypes of Authenticity Tools and of OAIS-based Information Browsing (CASPAR-2303-RP-0101-1_0), 2009
- Guercio M., Barthelemy J., Bonard A., Authenticity Issue in Performing Arts Using Live Electronics, http://articles.ircam.fr/textes/Guercio07b/index.pdf , 2008

- Factor M., Henis E., Naor D., Rabinovici-Cohen S., Reshef P., Ronen S., Michetti G., Guercio M.: Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage. TaPP '09. First Workshop on the Theory and Practice of Provenance. San Francisco, (2009) http://www.usenix.org/event/tapp09/tech/full_papers/factor/factor.pdf
- Guercio M.: Conceptual Framework and Chain of Custody for Sustaining the Digital trustworthiness. Perspectives on Metadata. Digital edition & preservation. Vienna (12 November 2009), https://fedora.phaidra.univie.ac.at/fedora/get/o:45908/bdef:Asset/view
- Giaretta D., Matthews B., Bicarregui J., Lambert S., Guercio M., Michetti G., Sawyer D., Significant Properties, Authenticity, Provenance, Representation Information and OAIS, iPRES 2009, The Sixth International Conference on the Preservation of Digital Objects: Proceedings, California Digital Library, 67-73 (2009)
- Giaretta, D.: Advanced Digital Preservation, Springer (2011), specifically chapters 13 and 17.11

ENSURE

Project title: ENSURE - Enabling Knowledge Sustainability Usability and Recovery for Economic Value

Website: http://ensure-fp7-plone.fe.up.pt/site/

Research goals:

Drawing on motivation from use cases in health care, finance and clinical trials, the project intends to significantly extend the state-of-the-art in digital preservation which to-date has focused on relatively homogeneous cultural heritage data. Specifically, the use cases are intended to address the following digital preservation issues

- safely leveraging scalable pay-as-you-go infrastructure such as clouds,
- having businesses understand the economic implications of preservation,
- conforming to regulatory, contractual and legal requirements as part of a whole workflow,
- managing long term integrity and authenticity significant intellectual property or highly personal data,
- using off-the-shelf IT technologies for preservation to support different types of digital resources.

Expected results:

- to evaluate the cost and benefit of different quality solutions, enabling the selection of the most cost-effective solution,
- to build on lifecycle management approaches to manage the preservation lifecycle, ensuring regulatory compliance, allowing changes in the preservation approach to reflect environmental changes, addressing evolution of ontologies, and managing the quality of the digital objects over time,
- to ensure long-term data protection, addressing changes in personally identifiable information, new and evolving regulations, managing user identities over decades
- to evaluate the costs, risks and benefits, and demonstrate how to use emerging commonly available IT to enable scalable solutions for long term digital preservation, considering in particular cloud storage and virtual application image capture.

Project terms: 2011-2014

Partners:

- IBM Israel – Science and Technology ltd
- Universidade do Porto, Portugal
- Luleå University of Technology, Sweden
- Fraunhofer-Gesellschaft zur Foerderung der Angewandten Forschung E.V., Germany

- Atos, Spain
- Custodix NV, Belgium
- Cranfield University, UK
- Maccabi Healthcare Services, Israel
- Science and Technology Facilities Council, UK
- Philips Electronics Nederland B.V., Netherlands
- Centro superior de investigacio en el saud publica, Spain
- JRC Capital Management Consultancy Research GMBH, Germany
- Tessella PLC

References:

- D11.1 Global Architecture Document Version 1.0, (2011) http://ensure-fp7-plone.fe.up.pt/site/deliverables/d11.1-global-architecture-document-release-1.0/view
- D1.2.1 Requirements Document: ENSURE Requirements Specifications, Use Cases and Scenarios Descriptions, (2011) http://ensure-fp7-plone.fe.up.pt/site/deliverables/requirements-deliverable-d1.2.1/view

InSPECT

Project title: InSPECT - Investigating the Significant Properties of Electronic Content Over Time

Website: www.significantproperties.org.uk/

Research goals:

- the analysis of the whole concept of significant properties and the identification of those relevant for future representation of four types of digital objects (raster images, emails, structured text, digital audio): the development of a general methodology able to support the actions required for digital preservation.

Main results:

- to expand and articulate the concept of 'significant properties',
- to determine sets of significant properties for a specified group of digital object types,
- to evaluate methods for measuring these properties for a representative sample of representation formats,
- to investigate and test the mapping and comparison of these properties between different representation formats,
- to identify issues which will require further research.

Project terms: 2007-2009

Partners:

- Centre for e-Research, King's College London
- The National Archives (TNA), UK

References:

- InSPECT, Final report, (1 December 2009) http://www.significantproperties.org.uk/inspect-finalreport.pdf

InterPARES

Project title: InterPARES - International Research on Permanent Authentic Records in Electronic Systems

Website: www.interpares.org

Research goals:

- to develop the knowledge essential to the long-term preservation of authentic records created and/or maintained in digital form,
- to provide the basis for standards, policies, strategies and plans of action capable of ensuring the longevity of such material and the ability of its users to trust its authenticity. InterPARES has developed in three phases.

Project terms: 1999 – 2012. The project has been developed in three phases:

- InterPARES 1 (1999-2001) focused on the development of theory and methods ensuring the preservation of the authenticity of records created and/or maintained in databases and document management systems in the course of administrative activities. Its findings present the perspective of the records preserver,
- InterPARES 2 (2002-2007) continued to research issues of authenticity, and examined the issues of reliability and accuracy during the entire lifecycle of records, from creation to permanent preservation. It focused on records produced in dynamic and interactive digital environments in the course of artistic, scientific and governmental activities,
- InterPARES 3 (2007-2012) built upon the findings of InterPARES 1 and 2, as well as other digital preservation projects worldwide. It put theory into practice, working with archives and archival / records units within organizations of limited financial and / or human resources to implement sound records management and preservation programmes.

Main results:

- Requirements for Assessing and Maintaining the Authenticity of Electronic Records (2002) http://www.interpares.org/book/interpares_book_k_app02.pdf
- Chain of preservation (COP), (2009) http://www.interpares.org/ip2/ip2_models
- Creator Guidelines. Making and Maintaining Digital Materials: Guidelines for Individuals (2009) http://www.interpares.org/ip2/display_file.cfm?doc=ip2(pub)creator_guidelines_booklet.pdf
- Preserver Guidelines. Preserving Digital Records: Guidelines for Organizations (2009) http://www.interpares.org/ip2/display_file.cfm?doc=ip2(pub)preserver_guidelines_booklet.pdf

Partners:

[The actual set of partners is very large. This list is limited to the main partners involved as public institutions for custody].

- National Archives and Records Administration of United States – NARA
- National Archives of Canada
- National Archives of China
- National Archives of France
- National Archives of Italy
- National Archives of Norway
- National Archives of Sweden
- National Archives of The Netherlands
- National Archives of UK
- Public Records Office of Hong Kong

References:

- Authenticity Task Force, Appendix 2: Requirements for Assessing and Maintaining the Authenticity of Electronic Records. The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project, Luciana Duranti, ed., San Miniato, Italy, Archilab, 204–219, (2005), online reprint available at http://www.interpares.org/book/interpares_book_k_app02.pdf

- Thibodeau K. et al.: Part Three – Trusting to Time: Preserving Authentic Records in the Long Term: Preservation Task Force Report, The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project, Luciana Duranti, ed., Archilab, San Miniato, Italy, pp. 99–116, (2005), online reprint available at http://www.interpares.org/book/interpares_book_f_part3.pdf .
- Duranti L., Preston R. (eds): International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records, Associazione nazionale archivistica italiana, Padova, (2008), online reprint available at http://www.interpares.org/ip2/book.cfm.
- Duranti L., Suderman J., Todd M.: InterPARES 2 Project, Policy Cross-domain Task Force, in InterPARES 2 Project Book: Appendix 19. Duranti L., Preston R. (eds): International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records, Associazione nazionale archivistica italiana, Padova, (2008), online reprint available at http://www.interpares.org/ip2/book.cfm.
- Chain of preservation (COP), (2009) http://www.interpares.org/ip2/ip2_models
- Creator Guidelines. Making and Maintaining Digital Materials: Guidelines for Individuals, (2009) http://www.interpares.org/ip2/display_file.cfm?doc=ip2(pub)creator_guidelines_booklet.pdf,
- Preserver Guidelines. Preserving Digital Records: Guidelines for Organizations, (2009) http://www.interpares.org/ip2/display_file.cfm?doc=ip2(pub)preserver_guidelines_booklet.pdf

KEEP

Project title: KEEP - Keeping Emulation Environments Portable

Website: http://www.keep-project.eu/

Research goals:

- to help facilitate universal access to our cultural heritage by developing flexible tools for accessing and storing a wide range of digital objects,
- to develop an emulation access platform to enable accurate rendering of both static and dynamic digital objects: text, sound, and image files; multimedia documents, websites, databases, videogames etc.

Project terms: 2009-2011

Main results:

- KEEP Emulation Framework (EF), http://emuframework.sf.net

Partners:

- Bibliothèque nationale de France
- Joguin SAS, France
- Koninklijke Bibliotheek, Netherlands
- Computerspiele Museum, Germany
- University of Portsmouth, UK
- Deutsche Nationalbibliothek, Germany
- Cross Czech a.s., Czech Republic
- Tessella, UK
- European Games Developer Federation, Germany

References:

- D3.3 Final document analyzing and summarizing metadata standards and issues across Europe (2010) http://www.keep-project.eu/ezpub2/index.php?/eng/Products-Results/Public-

deliverables/D3.3-Final-document-analyzing-and-summarizing-metadata-standards-and-issues-across-Europe

LiWA

Project title: LiWA - Living Web Archive.

Website: http://www.liwa-project.eu

Research goals:

- to develop and demonstrate web archiving tools able to capture content from a wide variety of sources, to improve archive fidelity and authenticity and to ensure long term interpretability of web content. In particular, web archiving faces many of the same challenges as the emulation community, and synergies could be developed across these areas,
- to look beyond the pure "freezing" of web content snapshots for a long time, transforming pure snapshot storage into a "Living" Web Archive (i.e. long term interpretability as archives evolve, improved archive fidelity by filtering out irrelevant noise and considering a wide variety of content).

Project terms: 2008-2011

Main results:

- Release in open-source of the complete components and tools issued from the project, grouped under the "liwa-technologies" project on Google code (http://code.google.com/p/liwa-technologies):
- The Rich Media Capture Module-a plug-in dedicated to the capture of streaming video content http://code.google.com/p/liwa-technologies/downloads/detail?name=rich-media-capture-plugin-1.0.jar
- The Temporal Coherence Analyser - a plug-in dedicated to the analysis of the temporal coherence of the archived Web content http://code.google.com/p/liwa-technologies/source/browse/temporal-coherence
- The Spam Assessment Interface - a Web service that enables the quality assessment of the archived Web content http://code.google.com/p/liwa-technologies/source/browse/assessment-interface
- The Semantic Analizer - a component dedicated to the detection of terminology evolution: http://code.google.com/p/liwa-technologies/source/browse/SemanticAnalyser http://code.google.com/p/liwa-technologies/downloads/detail?name=SemanticAnalyser 1.0.zip
- The Web Archive UI Framework - a client-side framework that helps creating User Interface helpers for Web archive browsing: http://code.google.com/p/liwa-technologies/source/browse/web-archive-ui-framework

Partners:

- Universität Hannover, Learning Lab Lower Saxony, Germany
- European Archive Foundation (now Internet Memory foundation), Netherlands
- Max-Planck-Institut für Informatik, Germany
- Computer and Automation Research Institute, Hungarian Academy of Sciences, Hungary
- Stichting Nederlands Instituut voor Beeld en Geluid, Netherlands
- Hanzo Archives Limited, UK
- National Library of the Czech Republic, Czech Republic
- Moravian Library, Czech Republic

References:

- The Rich Media Capture Module - a plug-in dedicated to the capture of streaming video content (2011) http://code.google.com/p/liwa-technologies/downloads/detail?name=rich-media-capture-plugin-1.0.jar

- The Temporal Coherence Analyser - a plug-in dedicated to the analysis of the temporal coherence of the archived Web content, (2011) http://code.google.com/p/liwa-technologies/source/browse/temporal-coherence
- The Spam Assessment Interface - a Web service that enables the quality assessment of the archived Web content, (2011) http://code.google.com/p/liwa-technologies/source/browse/assessment-interface
- The Semantic Analizer - a component dedicated to the detection of terminology evolution (2011) http://code.google.com/p/liwa-technologies/source/browse/SemanticAnalyser, http://code.google.com/p/liwa-technologies/downloads/detail?name=SemanticAnalyser-1.0.zip
- The Web Archive UI Framework - a client-side framework that helps creating User Interface helpers for Web archive browsing (2011) http://code.google.com/p/liwa-technologies/source/browse/web-archive-ui-framework

PARSE.Insight

Project title: Parse.INSIGHT - Permanent Access to the Records of Science in Europe

Website: http://www.parse-insight.eu/index.php

Research goals:

- to develop a roadmap and recommendations for developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information in Europe, from primary data through analysis to the final publications resulting from the research,
- to highlight the longevity and vulnerability of digital research data and concentrate on the parts of the e-Science infrastructure needed to support persistence and understanding of digital EU research assets.

Project terms: 2008-2010

Outputs:

- Science data infrastructure roadmap, 2010

Partners:

- Science and Technology Facilities Council (STFC), UK
- Koninklijke Bibliotheek (KB), Netherlands
- Deutsche Nationalbibliothek (DNB), Germany
- Max Planck Gesellschaft (MPG), Germany
- International Association of Scientific, Technical and Medical Publishers (STM), Netherlands
- European Space Agency ESRIN (ESA), France
- FernUniversitat in Hagen (FUH), Germany
- European Organization for Nuclear Research (CERN), Switzerland
- Georg-August-Universitat Gottingen Stiftung Oeffentlichen Rechts (UGOE), Germany

References:

- Deliverable D2.2. Science Data Infrastructure Roadmap, (June 2010), http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf

PersID

Project title: PersID - Building a persistent identifier infrastructure

Website: http://www.persid.org

Research goals:

- to establish an infrastructure for persistent identifiers using the "Uniform Resource Names for National Bibliography Numbers" (URN:NBN),
- to provide persistent identifiers as well as a transparent policy and technical framework for

their use in internet,
- to provide an independent, flexible and trustworthy system of identifying resources and making reliable links to them through implementation of an international standard system, the National Bibliography Number (NBN), built upon proven technologies and standards already in wide use (IETF RFC3188).

Expected results:
- a global governance infrastructure to host services that make it easy to resolve names association, like NBN, to resources without link rot,
- a system based on open cooperation of research organizations, publishers, national libraries and others who benefit of persistent identifiers,
- trustworthy, independent, public, flexible governance of URNs,
- solid and secure infrastructure based on IETF standards and open source software,
- a network of up-to-date knowledge bases about URNs that offers easy use of URNs for managers of all kinds of digital contents, and easy use of URNs for users of digital contents.

Project terms: 2009-2011

Partners:
- Data Archiving and Networked Services (DANS), Netherlands
- Denmark's Electronic Research Library, Denmark
- Fondazione Rinascimento Digitale (FDR), Italy
- Knowledge Exchange, Netherlands
- SURFfoundation, Netherlands
- Consiglio nazionale delle ricerche (CNR, National Research Council), Italy
- National Library, Sweden
- National Library, Finland
- Royal Library, Denmark
- National Library, Germany

References:
- PersID Project report (2011) http://www.persid.org/documents.html

PLANETS

Project title: PLANETS- Preservation and Long-term Access to our Cultural and Scientific Heritage.

Website: http://www.planets-project.eu/

Research goals:
- The primary and general goal for Planets is to build practical services and tools to help ensure long-term access to our digital cultural and scientific assets. The specific objectives concern the development of:
- preservation planning services that empower organisations to define, evaluate, and execute preservation,
- methodologies, tools and services for the characterisation of digital objects,
- innovative solutions for preservation actions tools which will transform and emulate obsolete digital assets,
- an interoperability framework to seamlessly integrate tools and services in a distributed service network,
- a testbed to provide a consistent and coherent evidence-base for the objective evaluation of different protocols, tools, services and complete preservation plans,
- a comprehensive dissemination and take-up program to ensure vendor adoption and effective user training.

Main results:

- Supplier Vendor Briefing White Paper summarising the findings of interviews with 18 of the world's leading IT companies,
- recommendations and components to extract and evaluate digital object properties, namely the planning tool Plato, which may be used to ensure the authenticity of any object with regard to changes stemming from the application of a preservation action.
- cases studies related to: 1) integrating Planets and Fedora Commons, 2) analysis of the emulation processes for dynamic records at the National Archives of the Netherlands, 3) ingestion of digital material representing folklore for Denmark at the Royal Library in Copenhagen.

Project terms: 2006-2010

Partners:

- The British Library
- The National Library of the Netherlands
- Austrian National Library
- The Royal Library of Denmark
- State and University Library, Denmark
- The National Archives of the Netherlands
- The National Archives of England, Wales and the United Kingdom
- Swiss Federal Archives
- University of Cologne, Germany
- University of Freiburg
- HATII at the University of Glasgow, Uk
- Vienna University of Technology
- The Austrian Institute of Technology
- IBM Netherlands
- Microsoft Research Limited
- Tessella Plc, UK

References:

- Planets components for the extraction and evaluation of digital object properties (2010) http://www.planets-project.eu/docs/reports/Planets_PC3-D23B(DOPWGreport).pdf
- Becker C., Rauber A.: Decision criteria in digital preservation: What to measure and how. Journal of the American Society for Information Science and Technology (JASIST), 62(6), pp. 1009-1028 (2011) http://onlinelibrary.wiley.com/doi/10.1002/asi.21527/abstract
- Becker C., Kulovits H., Guttenbrunner M., Strodl S., Rauber A., Hofman H.: Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. International Journal on Digital Libraries (IJDL), 1(4), (2009) http://www.springerlink.com/content/801685k478136425/

PrestoPRIME

Project title: PrestoPRIME - Keeping audiovisual contents alive

Website: www.prestoprime.org

Research goals:

PrestoPRIME will pursue the research and development with four strands of activity, each of which is associated with one principal objective, against which progress and outcomes will be assessed:

- to research and develop means of ensuring the permanence of digital audiovisual content in archives, libraries, museums and other collections,
- to research and develop means of ensuring the long-term future access to audiovisual content in dynamically changing contexts,
- to integrate, evaluate and demonstrate tools and processes for audiovisual digital permanence and access,
- to establish a European networked Competence Centre to gather the knowledge created by PrestoPRIME and deliver advanced digital preservation advice and services in conjunction with the European Digital Library Foundation and other projects

Expected results:

- data model: a specification of the PREMIS-based digital preservation approach implemented by PrestoPRIME, including support for the special digital preservation requirements of audiovisual content,
- financial models and calculation mechanisms: a market analysis for use by archives and by vendors,
- strategy for use of preservation metadata for within a digital library with examples of use in audiovisual preservation,
- tools for modelling and simulating migration-based preservation,
- metadata models, interoperability gaps, and extensions to preservation metadata standards,
- media formats, identification methods and implementations for multivalent preservation,
- analysis of the threats to data integrity from use of large-scale data management environments,
- design and specification of the audiovisual preservation toolkit,
- European Digital Library implementation guidelines for audiovisual archives,
- preservation process modelling (including a review of semantic process modelling and workflow languages),
- definition of scenarios,
- glossary of rights.

Project terms: 2009-2012

Partners:

- Institut national de l'audiovisuel (Ina), France
- British Broadcasting Corporation (BBC), UK
- IT Innovation Centre, UK
- Joanneum Research (JRS), Austria
- Eurix, Italy
- Institute of Sound and Vision (B&G), Netherlands
- Rai Radiotelevisione italiana, Italy
- ExLibris LTD, Israel
- Oesterreichischer Rundfunk (ORF), Austria
- Doremi, France
- The European Foundation, Netherlands
- Vrije Universiteit Amsterdam (VUA), Netherlands
- University of Liverpool, Uk
- Universität Innsbruck (UIBK), Austria
- Technicolor, Netherlands

References:

- Wright R.: Deliverable D7.1.4 Audiovisual Digital Preservation Status Report 2 (2011)
- https://prestoprimews.ina.fr/public/deliverables/PP_WP7_D7.1.4_Annual_AV_Status_2_R0_v1.00.pdf

- Kashi N., Sherwinter N.: Deliverable D2.1.3 AV Data Model: Final Specification (2011) https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.1.3_AV_Data_Model_R0_v1.00.pdf
- Kashi N., Wright R.: Deliverable D2.2.3 Strategy for Use of Preservation Metadata for within a Digital Library with examples of use in audiovisual preservation (2011) https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.2.3_Strat_Pres_Metadata_R0_v1.00.pdf
- Addis M., Jacyno M.: Deliverable D2.1.2 Tools for modelling and simulating migration-based preservation (2010) https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.1.2_PreservationModellingTools_R0_v1.00.pdf
- Schreiber G.: Deliverable D2.2.2 Metadata Models, Interoperability Gaps and Extensions to Preservation Metadata Standards (2010) https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.2.2_MetadataModels_InteroperabilityGaps_v1.50.pdf
- Addis M.: Deliverable ID3.2.1 Threats to data integrity from use of large scale data management environments (2010) https://prestoprimews.ina.fr/public/deliverables/PP_WP3_ID3.2.1_ThreatsMassStorage_R0_v1.00.pdf
- Kashi N.: Internal Deliverable ID3.1.1 Specification and Design of a Preservatio Environment for Audiovisual Content (2010) https://prestoprimews.ina.fr/public/deliverables/PP_WP3_ID3.1.1_preservation_specification_R1_v2.01.pdf

PROTAGE

Project title: PROTAGE - Preservation Organizations Using Tools in AGent Environments

Website: http://www.protage.eu/index.html

Research goals:

- to research the potential of software agent ecosystems to support the automation of digital preservation tasks,
- to demonstrate the technical feasibility of such a system,
- to analyze implementation in different organizational environments,
- to explore possible integration with other digital preservation environments,
- to explore synergies with other RTD activities in digital preservation.

Main results:

- to allow content producers to create and publish in a preservation-compatible manner,
- to provide digital repositories with means of further automating the preservation processes,
- to facilitate seamless interoperation between content providers, libraries and archives, and end-users throughout Europe.

Project terms: 2007-2010

Partners:

- National Archives of Sweden
- LDB - Centre of competence for long-term digital preservation, Sweden
- Luleå University of Technology, Sweden
- National Archives of Estonia
- Fraunhofer-Gesellschaft, Germany
- Easy Innova, Catalunya
- University of Bradford, UK
- eXact learning solutions S.P.A., Italy

References:

- Briefing Paper: Value of Software Agents in Digital Preservation. Ver 1.0 (2010) http://www.protage.eu/files/Potential%20of%20agents%20in%20DP.pdf
- D 1.3 Methodology Handbook. The PROTAGE Approach to Digital Preservation. Version 1.0 (2009) http://www.protage.eu/files/D%201.3%20PROTAGE%20Methodology%20Handbook%20final%20version.pdf

SCAPE

Project title: SCAPE - SCAlable Preservation Environments
Website: http://www.scape-project.eu

Research goals:
To develop scalable services for planning and execution of institutional preservation strategies on an open source platform that orchestrates semi-automated workflows for large-scale, heterogeneous collections of complex digital objects. These services are intended to:

- identify the need to act to preserve all or parts of a repository through characterisation and trend analysis,
- define responses to those needs using formal descriptions of preservation policies and preservation plans,
- allow a high degree of automation, virtualization of tools, and scalable processing,
- monitor the quality of preservation processes.

Expected results:

- three large-scale testbeds from diverse application areas (digital repositories from the library community, web content from the web archiving community, and research data sets from the scientific community) to highlight digital preservation challenges;
- scalable services for planning and execution of institutional preservation strategies on an open source platform that orchestrates semi-automated workflows for large-scale, heterogeneous collections of complex digital objects

Project terms: 2011-2014
Partners:

- AIT Austrian Institute of Technology GmbH
- The British Library
- Open Planets Foundation, UK
- Internet Memory Foundation, UK
- ExLibris LTD , Israel
- Fachinformationszentrum Karlsruhe Gesellschaft für Wissenschaftlich-Technische Information GmbH, Germany
- Koninklijke Bibliotheek, Netherlands
- Keep Solutions NDA, Netherlands
- Microsoft Research Limited, UK
- Österreichische Nationalbibliothek, Austria
- Statsbiblioteket, Denmark
- Science and technologies Facilities Council, UK
- Technische Universität Berlin, Germany
- Technische Universität Vienna, Austria
- The University of Manchester, UK
- Université Pierre et Marie Curie, Paris 6, France

References:

- Report on decision factors and their influence on planning (2011) http://www.scape-project.eu/wp-content/uploads/2011/12/SCAPE_D14-1_TUW_V1.0.pdf

SCIDIP-ES

Project title: SCIDIP-ES – SCIence Data Infrastructure for Preservation – with a focus on Earth SCience

Website: http://www.scidip-es.eu

Goals:

To put in place long lived services which support repositories in undertaking long-term preservation of their digital holdings.

Expected results:

- Infrastructure services including registry/repository of representation information, gap analysis service, orchestration/brokerage service
- Toolkits which help to create the metadata used in these services
- A user community of critical mass

Project terms: 2011-2014

Partners:

- European Space Agency
- Science and Technology Facilities Council
- Stichting European Alliance for Permanent Access
- Advanced Computer Systems A.C.S. S.P.A.
- Foundation For Research and Technology Hellas
- Deutsches Zentrum Fuer Luft – Und Raumfahrt Ev
- Natural Environment Research Council
- Istituto Nazionale di Geofisica e Vulcanologia
- Engineering - Ingegneria Informatica S.P.A
- FTK Forschungsinstitut Für Telekommunikation E.V
- InConTec Gmbh
- Istituto Superiore per la Protezione e la Ricerca Ambientale
- Jacobs University Bremen Ggmbh
- Capgemini Italia S.P.A.
- Centre National d'Etudes Spatiales
- G.I.M Geographic Information Management NV
- Università degli Studi di Roma "Tor Vergata"

SHAMAN

Project title: SHAMAN- Sustaining Heritage Access through Multivalent ArchiviNg

Website: www. shaman-ip.eu

Research goals:

- to establish an open distributed resource management infrastructure framework enabling grid-based resource integration, reflecting, refining and extending the OAIS model and taking advantage of the latest state of the art in virtualization and distribution technologies from the fields of GRID computing, federated digital libraries, and persistent archives,
- to develop and integrate technologies to support contextual and multivalent archival and preservation processes which are adapted and significantly extended from the fields of content and document management and information systems,

- to develop and integrate technologies to support semantic constraint-based collection management to target one of the key challenges in automating one class of digital preservation core functions;
- to support the managing of future requirements by securing interoperability with future environments and maintaining essential properties of the preserved content.

Main results:

- three prototype application solutions built as a multilayer model on service oriented architecture in the domains of scientific publishing and parliamentary archives, industrial design and engineering, scientific applications

Project terms: 2007-2011

Partners:

- INMARK Estudios y Estrategias, Spain
- University of Liverpool. UK
- InConTec, DE
- Swedish School of Library and Information Science
- Xerox Research Centre Europe, FR
- FernUniversität in Hagen, DE
- Philips Innovation Lab, UK
- University of Strathclyde, DE
- Deutsche Nationalbibliothek, DE
- Georg-August-Universität Göttingen, DE
- Otto-von-Guericke Universität Magdeburg, DE
- Industrious Media, UK
- Globale Informationstechnik GmbH, DE
- HATII at University of Glasgow, UK
- INESC-ID, Portugal
- University of Illinois, US
- University of California, US

References:

- Dobreva M., Kim Y., Ross S.: Designing an Automated Prototype Tool for Preservation Quality Metadata Extraction for Ingest into Digital Repository. Collaboration and the Knowledge Economy: Issues, Applications, Case Studies. Cunningham P., Cunningham M. (eds), IOS Press, Amsterdam (2008) http://echallenges.org/e2008/outbox/eChallenges_ref_196_doc_4893.pdf
- D6.3 - Implementation of Templates to manage the ingest workflow (2010) www.shamanip.eu/shaman/sites/default/files/SHAMAN%20D6.3_Implementation%20of%20Templates%20to%20manage%20the%20ingest%20workflow_1.pdf
- Report on demonstration and evaluation activity in the domain "Memory institutions" www.shamanip.eu/shaman/sites/default/files/SHAMAN%20D14.2_Report%20on%20Demonstration%20and%20Evaluation%20activity%20in%20the%20domain%20on%20MI_0.pdf

TIMBUS

Project title: TIMBUS - Digital preservation for timeless business processes and services

Website: http://timbusproject.net/

Research goals:

The EU-cofunded TIMBUS project focuses on resilient business processes. It will make the execution context, within which data is processed, analysed, transformed and rendered, accessible over long periods. Furthermore, continued accessibility is often considered as a set of activities carried out in the

isolation of a single domain. TIMBUS, however, considers the dependencies on third-party services, information and capabilities that will be necessary to validate digital information in a future usage context.

Expected results:

Activities, processes and tools that ensure

- continued access to services and software
- the capacity to maintain the context within which information accessible, properly rendered, validated and transformed into knowledge

Project terms: 2011-2014

Partners:

- Caixa Magica Software (Portugal)
- Digital Preservation Coalition (UK)
- INESC – ID (Portugal)
- Intel (Ireland)
- iPharro Media (Germany)
- Karlsruhe Institute for Technology (Germany)
- Laboratorio de Intrumentacao e Fisica Experimental de Particulas (Portugal)
- Laboratorio Nacional de Engenharia Civil (Portugal)
- Muenster University (Germany)
- SAP – Lead partner (Germany)
- Secure Business Austria (Austria)
- Software Quality Systems (Germany)

References:

- TIMBUS Survey on Long-Term Aspects of Business Continuity (2011) https://intel.qualtrics.com/SE/?SID=SV_1LWzcs15h0TIiAk

Wf4ever

Project title: WF4ever - Advanced Workflow Preservation Technologies for Enhanced Science

Website: http://www.wf4ever-project.org

Research goals:

- to develop new models, techniques and tools for the preservation of scientific workflows, including the novel definition of a Research Object, which packages workflow descriptions, the provenance of their executions, and links to all the related resources upon which they depend,
- to propose methods and tools to proactively preserve and inspect workflow integrity and authenticity through the evaluation of workflow information quality, based on the provenance of workflows and their research objects and described in new vocabularies for the representation of the provenance of research objects in digital preservation systems.

Expected results:

- a software architecture for the design and implementation of scientific workflow preservation systems,
- a reference implementation instantiating the architecture and enabling the preservation and efficient retrieval of scientific workflows across a range of domains,
- a new Research Object model for the description of scientific workflows and related materials,
- new techniques and tools for workflow decay analysis, abstraction and comparison,

- new techniques and tools for Research Object evolution, personalized recommendations and collaboration between scientists,
- new techniques and tools for integrity and authenticity management based on provenance models of workflow-related Research Objects,
- the application of our results and technology to two workflow-intensive scientific use cases in the areas of astronomy and genomics.

Project terms: 2010-2013

Partners:

- The University of Manchester, UK
- Academisch Ziekenhuis Leiden - Leids Universitair Medisch Centrum, Netherlands
- The Chancellor, Masters and Scholars of the University of Oxford, UK
- Agencia Estatal Consejo Superior de Investigaciones Cientificas, Spain
- Universidad Politecnica de Madrid, Spain
- Instytut Chemii Bioorganicznej Pan, Poland

References:

- D4.1. Workflow Integrity and Authenticity Maintenance Initial Requirements (2011) http://repo.wf4ever-project.org/dlibra/doczip?id=18