SAPIENZA
UNIVERSITÀ DI ROMA

Documenting the Authenticity and Provenance of Digital Data

MARIA GUERCIO

Università degli studi di Roma "La Sapienza"

CINI – Consorzio Interuniversitario Nazionale per l'Informatica

maria.guercio@uniroma1.it

APA RSEN
Alliance Permanent Access to the Records of Science in Europe Network

cini consorzio interuniversitario nazionale per l'informatica

Content

1. A conceptual framework for authenticity management: a re-use approach
2. Authenticity management tools
3. Authenticity evidence record: the framework
4. Case studies
5. Conclusions and references
6. Exercise

1. A conceptual framework for authenticity management

a re-use approach

Trust, authenticity and digital preservation

- Trust in the (digital) world operates **within history;** it is related to concepts of **place** and **responsibility**; it implies trustful **relationships** between the entities (organizations and individuals) involved and **mechanisms** and **services** by which it can be established, implemented, promoted, verified

- These mechanisms cannot be limited to simple means of **authentication** commonly implemented because of their capacity of declaring the authenticity of a bitstream at one specific moment in time

Trust cannot be blind or feel like and act of faith

- In the networked society questions related to trust are more challenging than in the past because of the distributed archives in the cloud
- In the digital environment the assumption for trusted preservation implies (is based on) the capacity of ensuring and documenting:
  - **data accuracy**: a question of truthfulness, exactness, precision or completeness): it has to be **assessed per se** to govern the risks related to the transmission across space (between persons and/or systems) and time (between digital systems when upgraded or in case of migration)
  - **reliability of content information and provenance/context information when created**: a responsibility of the producer
  - **authenticity**: the digital identity and integrity are inferred from the circumstances of their maintenance and preservation thanks to “an unbroken chain of responsible and legitimate custody” which shifts from the producer to the **trusted** custodian (L. Duranti)

Trust in digital environment is based on movable responsibilities and trusted relationships

- Accuracy and authenticity are **shifting responsibilities** that move over time from the producer/data keeper to the trusted custodian/repository
- Because of the dynamic nature of the digital environment, these responsibilities must have an **institutional** nature and a complex and well defined structure: they need **trusted relationships** based on solid **business principles**, **formalized agreements** and **accreditation processes**
- Responsibilities and frameworks have to be evaluated periodically on formal basis, according to well stated recommendations by **recognized auditors** (to ensure impartiality and offer comparable and solid evidence)

Core elements for establishing trusted responsibilities for preservation

- ***reputation***, based on the assessment of the trustee's past actions and conduct;
- ***performance***, which is the relationship between the trustee's present actions and the conduct required to fulfill his or her current responsibilities as specified by the truster;
- ***competence***, which consists of having the knowledge, skills, talents, and traits required to be able to perform a task to any given standard;
- ***confidence***, which is an assurance of expectation of action and conduct the truster has in the trustee
  - Piotr Sztompka, *Trust.* Cambridge: Cambridge University Press, 1999, 6)

Trust and authenticity (Luciana Duranti)

- "In the digital environment authenticity is an inference based on foundation evidence and, in some measure, on **confidence** in the **performance** and **competence** of the keeper of the material, based on its **reputation**.
- The **level of trust** required is proportional to the **sensitivity of the material to be trusted as authentic** and the adverse consequences of its lack or loss of trustworthiness.
- To guarantee the authenticity of **digital records** [**content information**] requires intentional action or intervention by trusted entities imbued with accountability, but also an **adequate framework of policies, procedures**, and technologies. This has always been the case […]
- We can no longer determine authenticity on the object - record, which is composite (stored + manifested) and permanently new (re-production), but **must make an inference of authenticity from its environment of creation, maintenance & use and preservation**"

authenticity evidence for digital preservation: a demanding task and a complex approach

- Against the tendency of underestimating the role and the complexity of authenticity, InterPARES, CASPAR and APARSEN projects have recognized
  - the centrality of a **conceptual framework for ensuring and presuming authenticity** as part of the chain of custody for any kind of digital heritage
  - the **meaningfulness of standard developed by the documentary disciplines, mainly of the archival and recordkeeping concepts** when defining functions and requirements in this area and
  - the essential need of a **transdisciplinary cooperation** to cope with it

The transdisciplinarity approach to the authenticity evidence: actors profile

- The main actors involved in the main projects related to the authenticity in the digital environment have included:
  - **archivists** (senior and junior academic scholars) with competence on authenticity of digital records and direct experience of InterPARES project (participants of the IP1 Authenticity task force)
  - **experts for conceptual modeling and business workflows**
  - IT developers and IT engineers with experience of **orchestration systems** for digital curation and preservation
  - **scholars responsible for definition of the OAIS model** and its following revision (in 2012)
  - experts in domains and contents which require new concepts and tools for supporting authenticity (**digital music, e-science, performing arts**)
  - professionals responsible for managing **digital repositories** and **auditors involved in certification processes**

Conceptual and methodological framework: a re-use approach - 1

- **OAIS as a reference model** to be implemented as the basic architecture to manage workflows and responsibilities
- **InterPARES as the conceptual framework** for interrelating principles, policies and procedures to compare and assess quality and consistency of the digital practices for authenticity
- **CASPAR** as a methodological approach for a **standardized set of tools** (only partially developed**)** able to integrate and document the main events and functionally collect the information relevant for supporting authenticity.
- **APARSEN – WP 2400** to develop tools for providing **authenticity evidence** in the preservation processes and **case studies**

Conceptual and methodological framework: a re-use approach - 2

- **SCIDIP-ES: models and toolkit for supporting the authenticity** evidence record and automating the collection and the management of relevant information (with specific attention for provenance)
- **PREMIS as dictionary for supporting the interoperability** for managing authenticity evidence records
- **ISO 15489** and **ISO 23081** for defining crucial phases and steps when changes of custody and/or preservation are involved
- **ISO 16363 – Certification of TDR**: specific attention to the identification of **measures relevant for qualifying the preservation activities and presuming authenticity**

General assumptions - 1

- It is not possible (feasible) to preserve electronic resources in the form of original unchanged content information: we have only the ability to reproduce them in the form of **authentic copies** thanks to the preservation of valid 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/content information (not necessarily the same objects as those originally deposited) and the procedure of validating them.

General assumptions - 2

- Not only the digital preservation is 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.

Documenting the chain of custody in compliance with OAIS

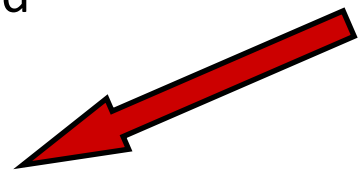
- The core issue concerns the capacity of developing of a multilayer approach able to support integrity and authenticity assessment according to interoperable processes
- Authenticity and integrity could be evaluated **as inference on the basis of the trustworthiness of the document/information system** in which the documents/information exist (in the creation and in the preservation environments)
- The document/information systems can ensure inference if compliant with standardized open models: **OAIS** and its **components for Preservation Description Information** are relevant for sustaining the inference process required to evaluate the authenticity evidence

Authenticity in OAIS and InterPARES

- **OAIS** (Magenta Book – July 2012):

  "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****"*

- **InterPARES clarified how the evidence has to be collected**: (both before and after preservation begins)
  - The authenticity has no degree in itself, BUT
  - the *presumption of the authenticity* is graduated
  - This assessment is supported by:
    - the preservation system but also by
    - the evidence collected in the business process as part of the information content management



The key issue is collecting as soon as possible the appropriate evidence for all the events that may affect authenticity

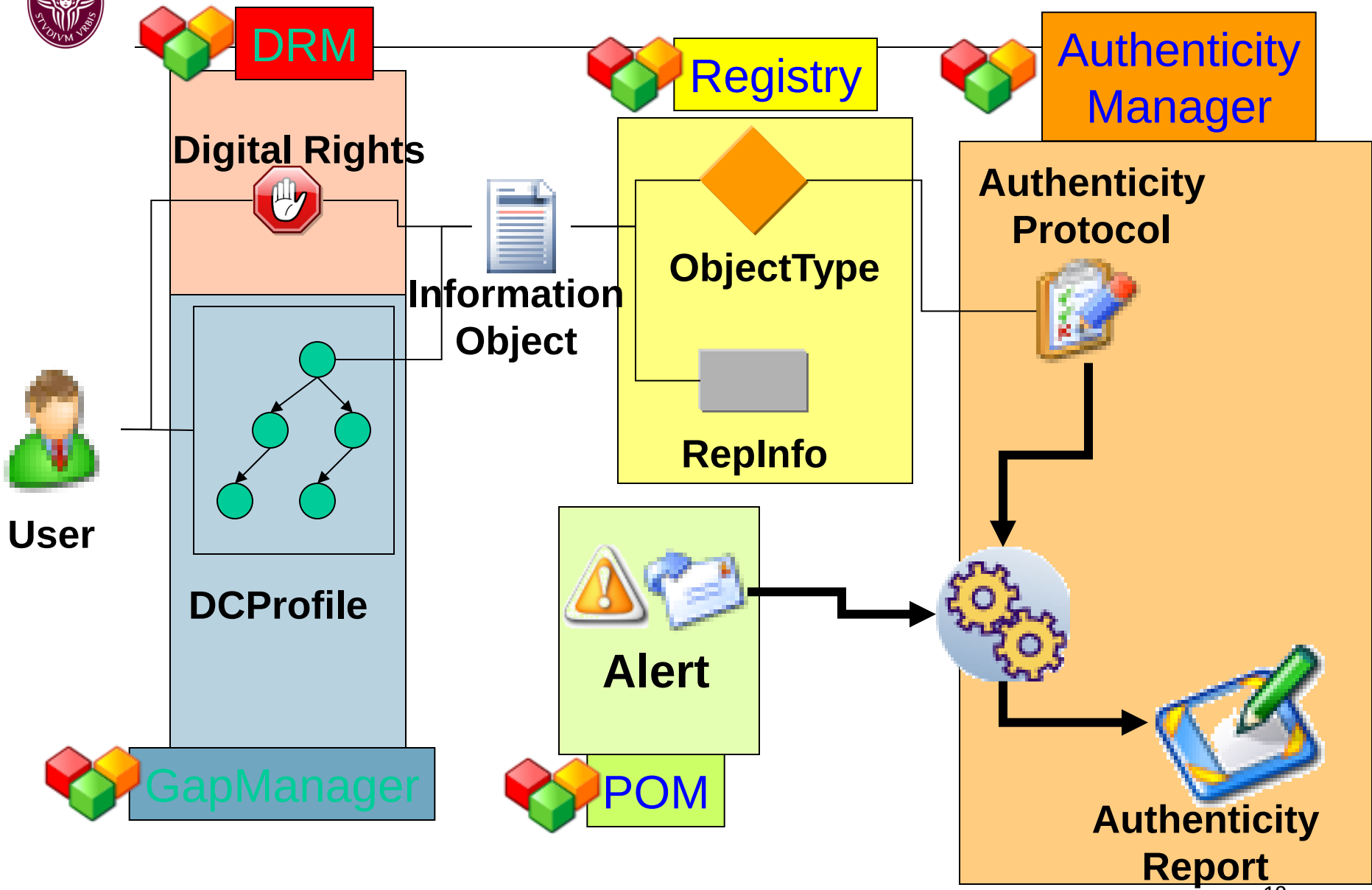
Authenticity management according to CASPAR conceptual model

- The process for protecting and assessing the authenticity needs the definition of **procedures** managed according to a model called **authenticity protocol** (AP) able to control the processes, the agents and events and collect relevant information according to a well designed and documented conceptual model
- An authenticity protocol is based on
  - a series of **steps** applied to a class of objects or to a class of events and related to the kind of PDI (reference, provenance, context, fixity, access rights information)
  - a workflow of events, which can be **automatic** or **manual**
  - the information related to the step **execution** (agent, time, place, context of execution)

Authenticity evidence according to CASPAR conceptual model

- The evidence includes the documentation of each execution of the procedures relevant for the authenticity (that is relevant for the identity and the integrity of the digital object) in the form of
  - an **authenticity report** for each step or for a series of steps,
  - the eventual **evaluation** of the execution
  - the **authenticity protocol history** developed as a synthesis of the outcomes from the application of the procedures and from he execution of the steps

Authenticity within the CASPAR framework



PDI in the OAIS framework: crucial types of information for assessing the authenticity of the digital content

- ***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 (why it was created, how it relates to other Content Information objects, etc.)*
- ***Provenance Information**: the history of the Content Information (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:** the information that identifies the access restrictions pertaining to the Content Information, including the legal framework, licensing terms, and access control*

The CASPAR and APARSEN contribution: authenticity management tools and definition of events

- The CASPAR project has identified the need for an Authenticity Management Tool with the capacity of **monitoring and managing protocols and procedures across the custody chain** in order to deliver the benefits of authenticity into information systems, from the creation to the preservation phase with specific reference to the definition of standardized **events** related to the **transformations** crucial for preservation and **authenticity tracking**
- These events can be usually categorized as part of PDI

The CASPAR and APARSEN contribution: operational guidelines and case studies

Alliance Permanent Access to the Records of Science in Europe Network

On the basis of and in continuation with CASPAR conceptual framework on authenticity

APARSEN proposes a *methodology for the management of the authenticity of Content Information (CI)* which includes:

- ***Formal authenticity model***: to represent the Content Information lifecycle/ business process and the management of related authenticity evidence based on a controlled list of transformational events
- ***Operational guidelines***: to guide a normalized process of instantiating the model in a specific environment
- ***Case studies***: carried out to improve the methodology and test its effectiveness in a set of heterogeneous environments

The CASPAR and APARSEN contribution: authenticity and the CI lifecycle/business process

Authenticity is affected by transformations and changes of custody the CI undergoes during its lifecycle. To assess authenticity we need to collect and preserve appropriate **evidence** *in order to be able, at a later time, to trace back these transformations*

- Transformations relevant to CI authenticity may occur (normally occur) before preservation begins
- The authenticity management process must encompass the whole CI lifespan since its creation and must consider with specific attention
  - changes of custody and
  - authenticity evidence needs to identify whose information have to be exchanged between different keeping and preservation systems
- Interoperability becomes a crucial requirement to be supported and defined by the authenticity management policy

Basic Definitions – 1/2

Agent — *Is the **actor** (human, machine, or software) associated with a given transformation of a CI, and **who bears the responsibility** of it.*

Representation — *Is a **set of digital objects** required to display, play, or otherwise make useable to a human a **given version of a CI**.*

Transformation — *Is **a change** that intervenes in conjunction with an event in the CI lifecycle, and produces **a new representation of the CI**, thus potentially **affecting its authenticity**.*

Basic Definitions – 2/2

Authenticity Evidence Record (AER)

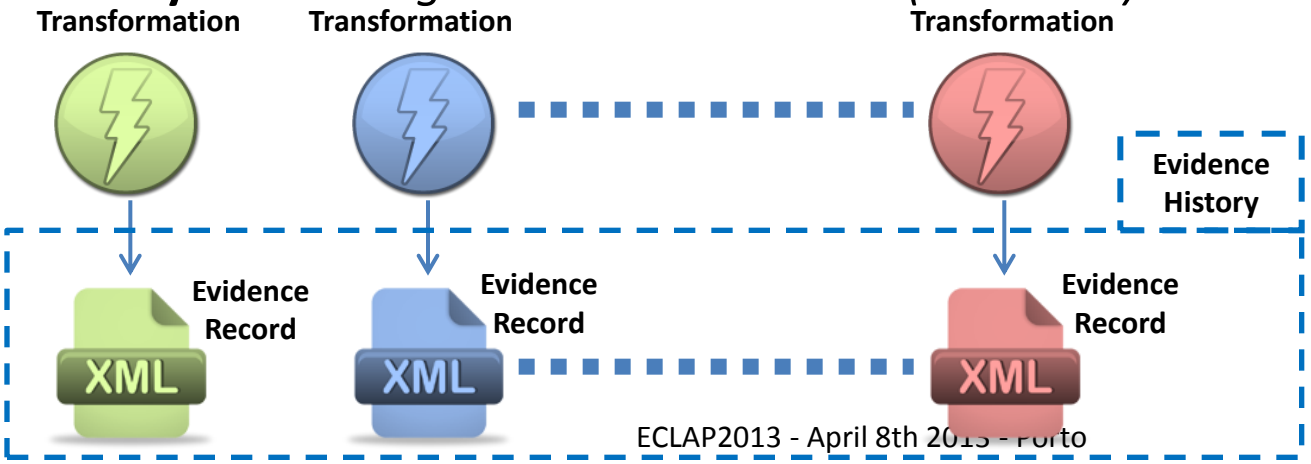
*Is the information that is **gathered and preserved in conjunction with a transformation**, to allow to assess, at a later time, the impact of that transformation on the IE authenticity, provenance and integrity (APARSEN)*

Authenticity Evidence Item

*Is an **individual element of the AER**. Typical AEI are specification of time, tools, etc., reference to involved digital CI and record of actions and controls performed by the agent during the transformation (APARSEN)*

Authenticity Evidence History

*Is the **set of the AERs for all the transformations** a CI has undergone since its creation (APARSEN)*



Authenticity Model: data dictionary - 1

Reference Step

- An Authenticity Step devoted to gather information about the **identification** of the content information

Provenance Step

- An Authenticity Step devoted to gather information about the **history** of the content information

Fixity Step

- An Authenticity Step devoted to gather information about the bit **integrity** of the content information

Context Step

- An Authenticity Step devoted to gather information about the **relationships** of the content informtion to its environment

Authenticity Model: data dictionary - 2

Authenticity Protocol History

- A report providing evidence of any changes of the *Authenticity Protocols*

Actor Type

Actor Occurrence

Manual Actor

Automatic Actor

Authenticity Recommendations

Experience

Best Practice

….

Applied To

- Association representing application

Based Upon

- Association representing control

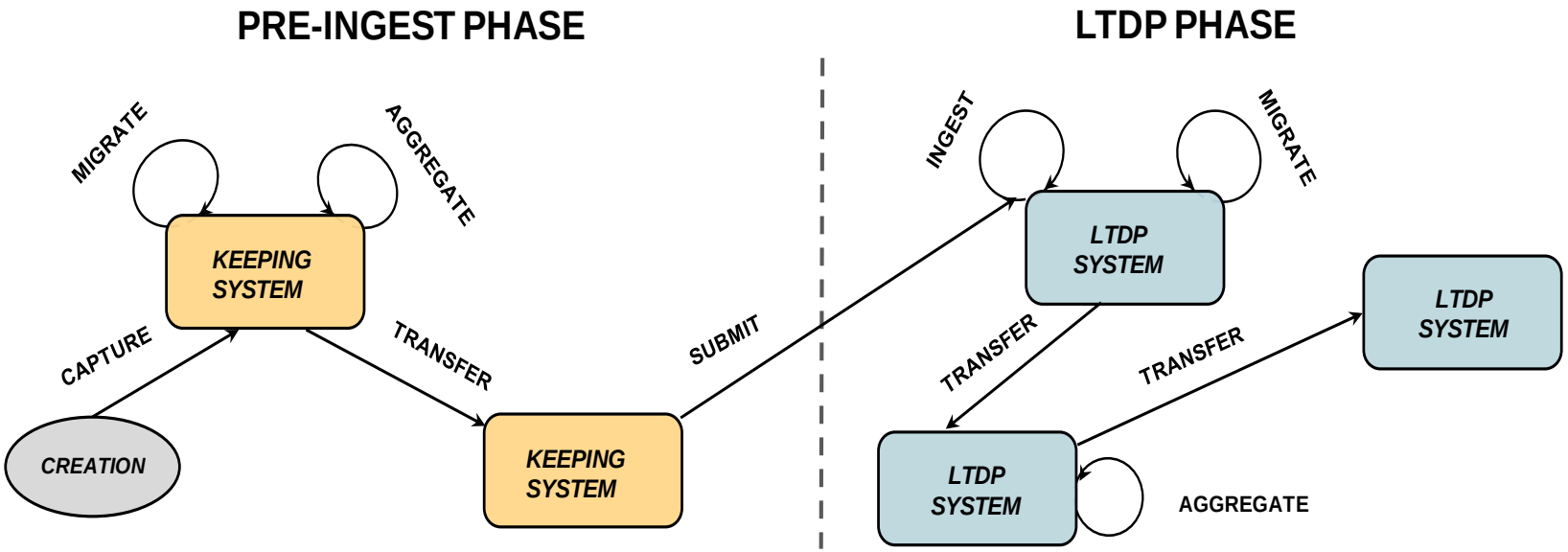
Documented By

- Association representing documentation

APARSEN contribution: state of the art (projects and standards)

- Analysis of the **outputs of the main research projects**
  - InterPARES and CASPAR (main reference on authenticity)
  - PLANETS, InSPECT, PROTAGE, SHAMAN, PARSE.Insight, LIWA, KEEP, PersID, PrestoPRIME, Wf4Ever, SCAPE, TIMBUS, ENSURE, SCIDIP-ES, ARCOMEM
- **Standards and recommendations** on management and the certification of **ERM** and **LTDP** systems
  - OAIS, PREMIS
  - MoReq2 and MoReq2010, ISO 15489-1:2001, ISO 23081-1:2006 (creation and management of digital resources)
  - UN/CEFACT – BRS. Transfer of Digital Records
  - TRAC, ISO 16363, ISO/DIS 16919 (certification of digital repositories)

Modeling the information content lifecycle/business process



- **PRE-INGEST PHASE**: from the creation of the CI to the beginning of the Long Term Digital Preservation (LTDP) process
- **LTDP PHASE**: encompasses all the transformations and the changes of custody the CI goes though along the LTDP process

Modeling the content information lifecycle/business process

- **Transformations** connected to lifecycle events **may affect authenticity**
- To **assess authenticity** it is necessary to **trace back all relevant transformations** in the form of **events**
- **Authenticity evidence** must be collected **in connection with lifecycle/business processes events**
- The model identifies a set of relevant events: **core set events**
- The model specifies for each event **the authenticity evidence to be collected**
- **Achieving interoperability is a crucial requirement for sustainability**

2. Authenticity management framework and tools

CASPAR authenticity protocols: a question of workflows

- **Authenticity protocol (AP)**: formal procedure that defines controls and actions to be performed in connection with transformations of digital resources during their preservation
- An authenticity protocol gives an **operational guideline to perform controls and to collect authenticity evidence**, and is based on
  - a **workflow** which can be automatic or manual
  - a **series of steps (relevant for authenticity)** applied to a class of objects or to a class of events and related to one or more components of the PDI
  - the **information related to the step execution** (actor, information, time, place, context of execution)

The APARSEN proposal

- The state of the art testifies that significant scientific contributions have been given
- A good level of theoretical and methodological formalization has been achieved
- **A large gap** still divides the theoretical results from the actual practices carried on in most repositories
- **CASPAR** and **InterPARES** have provided **a solid basic framework**, but further contribution was still needed with reference to the development of general **detailed guidelines** at concrete and operational level for the management of authenticity evidence with specific reference to:
  - the definition of a **core set of events**: when the evidence should be collected
  - the specification of the **evidence** to collect and to its structure

Authenticity protocols according to the APARSEN re-interpretation

Authenticity Protocol (AP): *formal 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*

- **APARSEN** has extended and brought to concrete implementation a concept originally introduced by **CASPAR**
- An **AP** has to be defined for each **event** in the lifecycle model
- The **AP** is organized as a sequence of **Authenticity steps (AS)**.
- Each **AS** is a set of elementary actions meant to:
  - perform a specific **control**
  - and/or collect one or more **Authenticity Evidence Items (AEI)**
- The execution of the **AP** for a given lifecycle event generates the **Authenticity Evidence Record (AER)** for that event

Events in the pre-ingest phase

- The *core set* of events (derived from ISO RM 15489 on RM):

  (Includes the most important and the most likely to occur)

  - **CAPTURE**: the CI is delivered by its author/producer to a keeping system;
  - **INTEGRATE**: new information is added or associated to a CI;
  - **AGGREGATE**: several CI are aggregated to form a new CI;
  - **DELETE**: a CI is deleted according to a stated policy;
  - **MIGRATE**: one or several components of the CI are converted to a new format;
  - **TRANSFER**: a CI is transferred to another keeping system;
  - **SUBMIT**: a CI stored in a keeping system is delivered to a LTDP

  (*Specific environments may require the definition of additional events*)

Events in the LTDP phase

- **LTDP-INGEST**: a CI delivered in a SIP is ingested and stored as an AIP
- **LTDP-AGGREGATE**: several CI stored in different AIPs, are aggregated in a single AIC;
- **LTDP-EXTRACT**: CIs are extracted from an AIC to form individual AIPs;
- **LTDP-MIGRATE**: one or several components of an AIP are converted to a new format;
- **LTDP-DELETE**: a CI stored in an AIP is deleted when its preservation time expires;
- **LTDP-TRANSFER**: a CI stored in an AIP is transferred to another LTDP system;

Lifecycle Transformations/Events definition – 1/2

- **CAPTURE**.
  - At the end of the creation process, the author delivers to a repository the original representation of the CI.

- **MIGRATION**.
  - Within the repository that holds it, the current representation of the CI is converted to a different format, with the intention to preserve its intellectual content.

- **CHANGE OF CUSTODY**.
  - The custody of a CI is transferred to a new repository, by handing to it the representation held by the current custodian.

Lifecycle Transformations/Events definition – 2/2

- **AGGREGATION.**
  - The representation of two or more CI are aggregated to form a new CI.

- **EXTRACTION.**
  - A component of a CI is extracted from its current representation to form the representation of a new CI.

- **INGESTION.**
  - The custody of a CI is transferred to a preservation repository. A new representation of the CI is generated, as an *Archival Information Package (AIP)*, conforming with the prescription of the OAIS Reference Model

Characterizing the events

- **Description**: circumstances and actions
- **Agents**: the person(s) who take the responsibility
- **Input**: the CI which are the object of the transformation
- **Output**: the CI which are the result of the transformation
- **Controls**: which controls are performed and by whom

The Authenticity Evidence Record is the synthesis of the characterization and should include:

- Identity and authentication data of agents and systems involved
- Date and time
- Specification of the actions performed
- Results of controls performed
- Other…

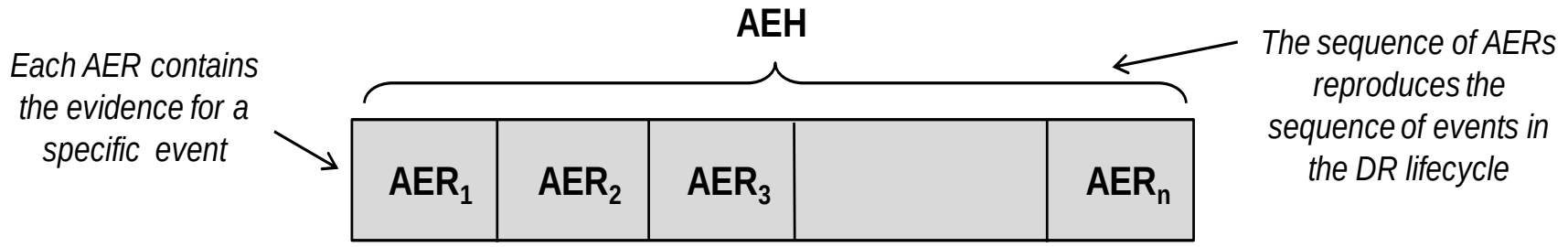
From theory to practice: the operational guidelines

Adapting the model to the needs of specific environments

1. **Analyze the needs of the Designated Community**
   - What does authenticity mean to the Designated Community?
   - Which kind of evidence is to be preserved?
2. **Identify relevant lifecycle events**
   - Events that may affect the authenticity and the integrity of the CI
3. **Define the Authenticity Evidence Records**
   - Which authenticity evidence items should actually be collected?
4. **Formalize** Authenticity Protocols to consistently perform the preservation action
   - Define the documentation and an AP for each relevant lifecycle event

3. Authenticity evidence record: the framework

Authenticity Evidence Records



- **Authenticity Evidence Record (AER):** structure containing the evidence collected in connection with a specific event relevant for preservation
- **Authenticity Evidence History (AEH):** incremental structure of AERs
- The **AEH** collected during the **PRE-INGEST** phase provides crucial information to generate the **PDI** in the **SIP**
- During the **LTDP** phase the new **AERs** contribute to the **PDI** of the **AIP**

AERs and interoperability

*Proper definition and standardization of **AERs** are crucial steps towards interoperability among keeping and preservation systems*

- Content Information go though several **changes of custody** through their lifecycle
- Keeping and preservation systems have **to interpret evidence collected by other systems**
- The way authenticity evidence is **collected**, **organized** and **exchanged** is the hearth of the problem
- **Standardization** is the **final goal**, but preservation is dynamic and along process and requires time and consensus
- Defining a standardized framework of transformational events and providing **operational guidelines** that could be **reasonably implemented** is an important preliminary step

The Operational Guidelines

Procedure to be followed when instantiating the model in a specific environment to get to the definition of an appropriate authenticity management policy

STEP 1. *Understanding the needs of the Designated Community*

- What authenticity means to the DC? Which evidence is needed?

STEP 2. *Identifying lifecycle or LTDP events relevant for preservation*

- Map specific events into the core set.
- Add context to specific events, if needed.

STEP 3. *Defining the policy and the Authenticity Evidence Records*

- Specify which evidence should actually be collected

STEP 4. *Formalizing Authenticity Protocols*

- Specify operational procedures to be followed for capturing complete and accurate information for preserving content information

Event templates

- **Events:** correspond in the lifecycle to relevant transformations and changes of custody that affect the Content Information (and its authenticity)
- **Core set of events:** defined by the model under quite general assumptions
- **Event templates:** provided for all core set events to specify:
  - **Agents** who are responsible for the transformation
  - **Input and output** digital Content Information
  - **Controls** that should be performed
  - **Authenticity evidence** to be collected and preserved

Event templates are comprehensive checklists of controls to be performed and evidence to be collected. They are meant to ensure completeness and accuracy, and provide a common ground for interoperability.

4. Case studies

Testing the methodology: case study analysis

- Case study analysis performed to check the validity of the model
- Test environments selected among the APARSEN partners
- Three main case study in different domains:
  - Health care data (ULSS Vicenza)
  - High Energy Physics data (CERN)
  - Social Science data (UK Archive)

Testing the methodology: the phases

- **Phase 1**: analyze current practices
  - What does authenticity mean to the community of users?
  - How is authenticity currently managed?
  - Which are the relevant events of the lifecycle?
  - Which evidence is currently collected?
- **Phase 2**: propose improvements
  - Identify deficiencies (relevant events not properly handled)
  - Specify actions and controls to be performed
  - **Define content and structure of the Authenticity Evidence Records**

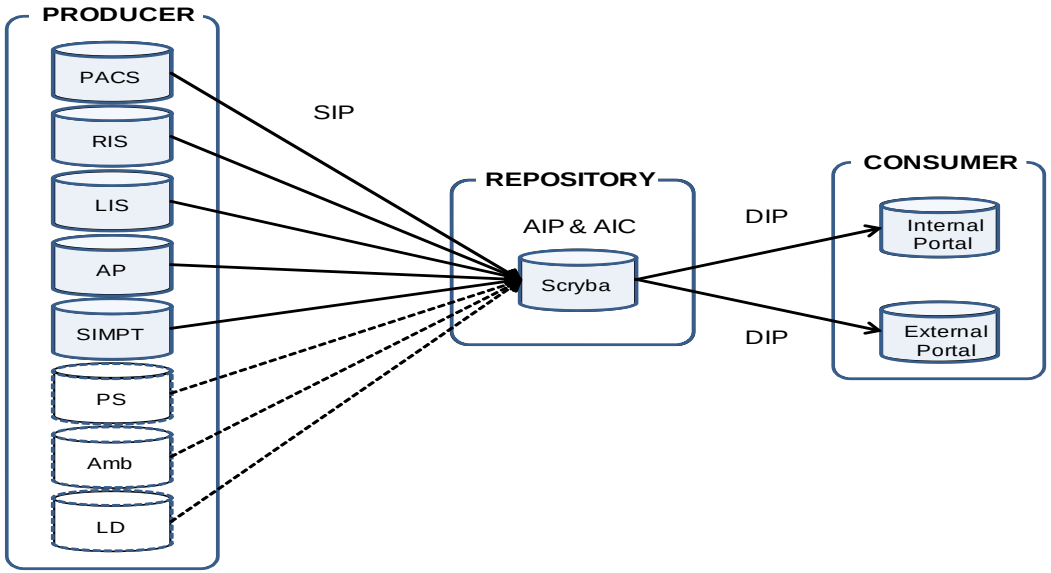
4. Case studies
4.1. The e-health records

Case study analysis

- **Repository of Vicenza Public Health Care system:**
  - Regional repository of the Italian Public Health Care System
  - Manages several kinds of digital information content: test results, diagnostic images, medical reports.
  - Complies with complex Italian regulations on LTDP

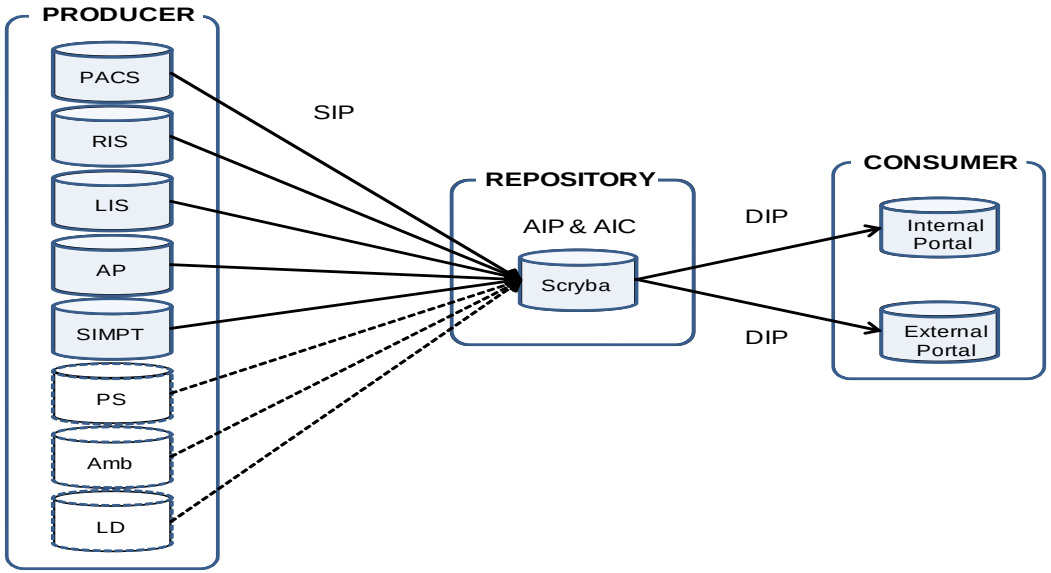
Authenticity is a crucial requirement in the Health Care environment, both because of the scientific relevance of the data and for the attribution of legal responsibilities

The preservation infrastructure - 1



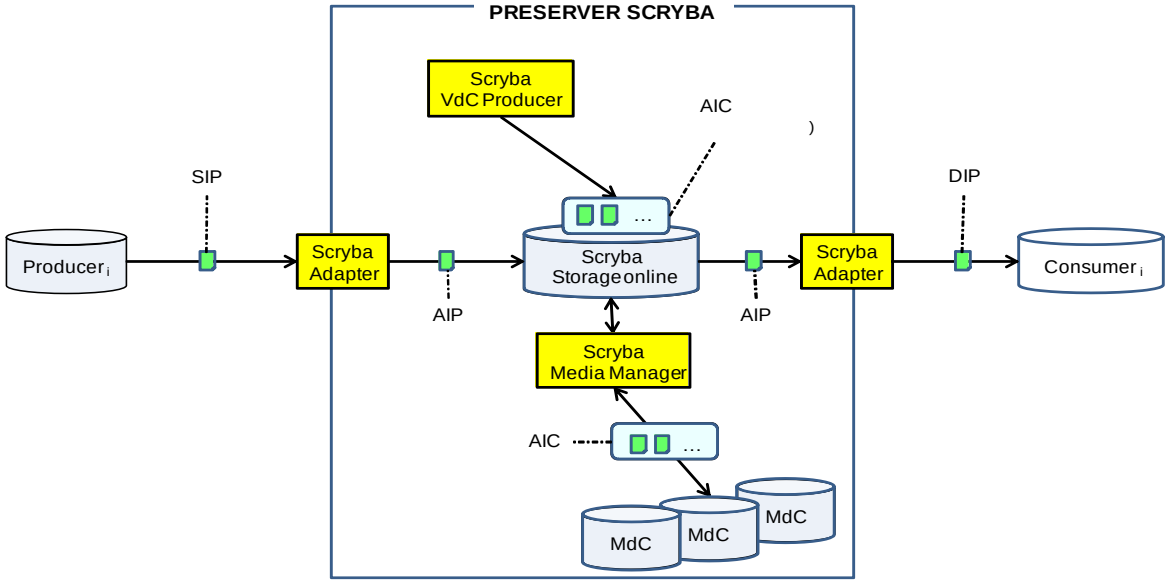
- The repository interfaces with a variety of producers
- Digital records are delivered by several departmental systems that collect them from satellite systems (imaging devices etc.) and include several types and separate workflows: **test results** (files in DICOM format) and **medical reports** (digitally signed by physicians)
- Digital records are delivered to LTDP repository shortly after creation
- The system includes distinct interfaces for internal personnel and patients

The preservation infrastructure - 2



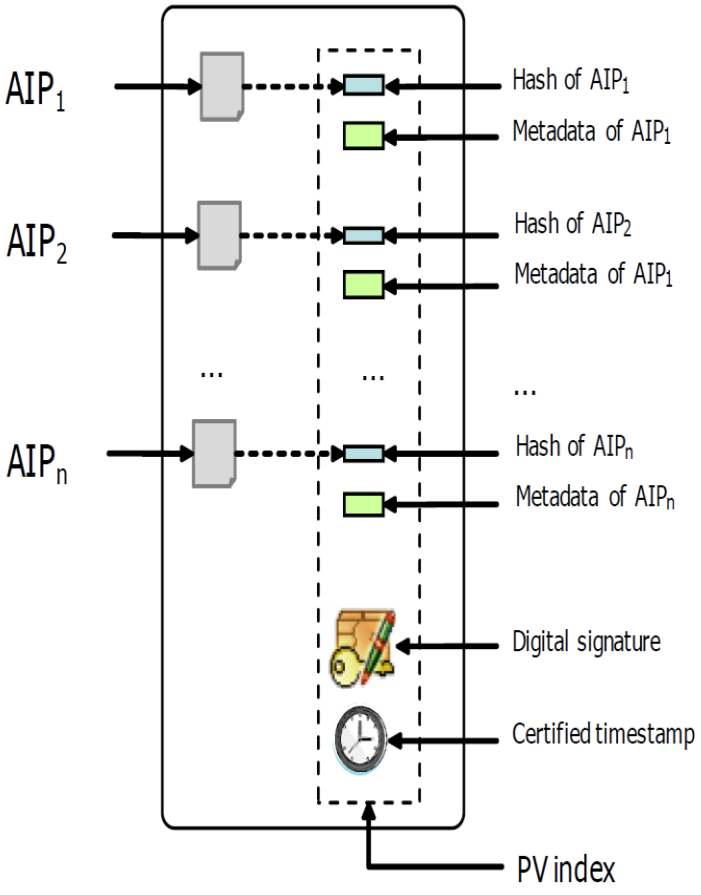
- Pre-ingest workflow involves several systems under different responsibilities
- Italian rules on LTDP are very specific, mostly centered on digital signatures, certified timestamps and aggregation of CI of various nature and provenance

The Scryba preservation system



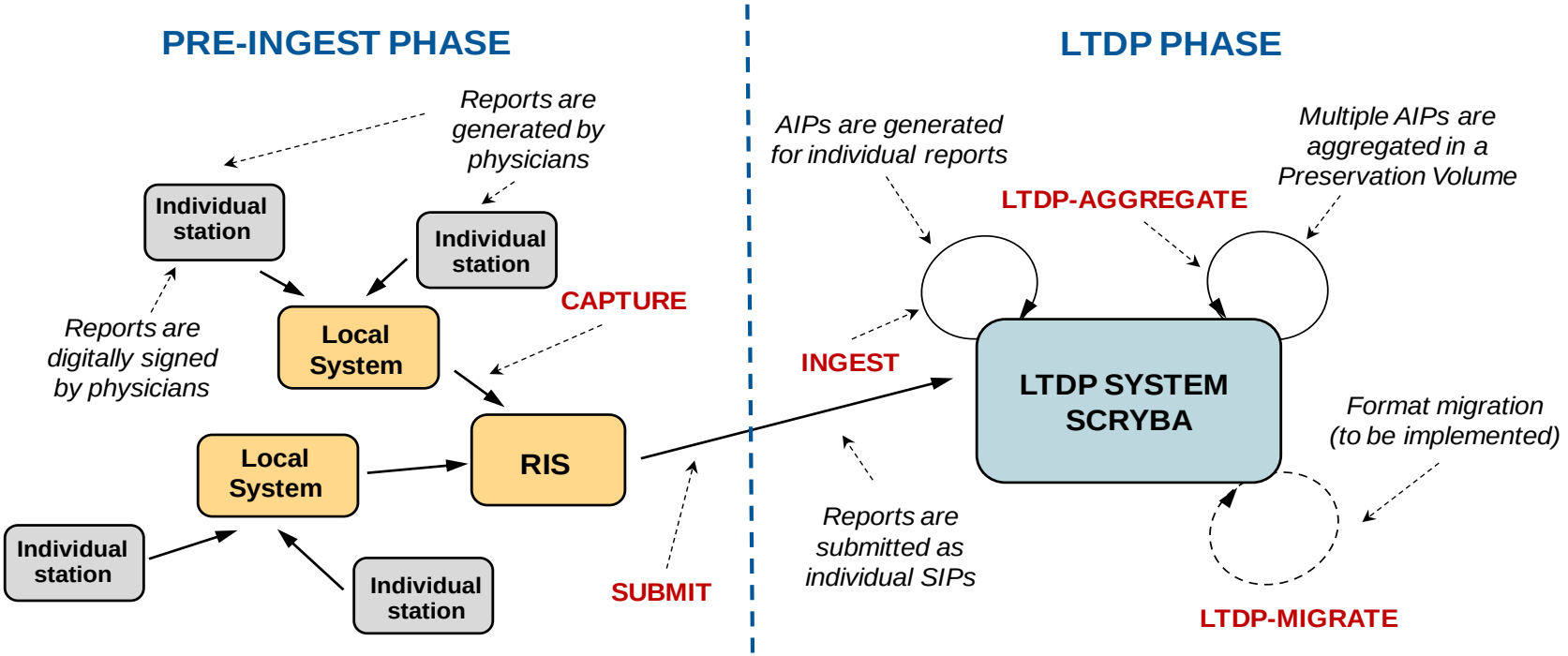
- Based on the **OAIS** principles, with additional features to guarantee the compliance with Italian **LTDP** regulations
- Modular structure based on **Adapters** modules tailored on the specific producers' and consumers' needs
- Core functions are: management of **AIPs**, the related transformations (aggregation and format migrations) and their secure storage

Italian regulations: Preservation Volumes



- CI is ingested as **AIP**
- **AIPs** are aggregated in large batches: **Preservation Volumes (PVs)**
- **PVs** are **AICs (Archival Information Collections)** and the actual object of the preservation process
- Besides the aggregated **AIPs**, the **PV** contains a **PV index**:
  - a hash file for each **AIP** in the **PV**
  - metadata for each **AIP**, in a format complying the national UNI SInCRO standard
- The **PV index** is digitally signed and time-stamped
- Several backup copies are preserved for each PV

The model for preserving radiology information



- Medical reports are written through a *Radiology Information System (RIS)*
- Reports go through changes of custody in the **PRE-INGEST** phase
- Five **events** corresponding to relevant transformations are represented in the model
- Digital signature provides crucial authenticity and provenance evidence

To define the repository policy for authenticity ~~according to the guidelines~~

- **Current practices** have been carefully identified
- **Templates** have been developed for the **relevant events**
- Event templates have included a **comprehensive list of controls** and have specified **authenticity evidence items**
- The **AEI** have been identified with reference to their capacity of proving the **identity** and **integrity** of content information transferred between systems under the ownership of **trusted organization** according to **PDI categories** and with reference to the **Designated Community**

Selecting appropriate Authenticity Evidence according to the designated community

- The AEI have been **carefully analyzed** and **agreed** with the **Designated Community**
- In the **Vicenza** case, **integrity checks** are not always performed and **AEIs** are not gathered for the **changes of custody of some specific workflows,** but this is accepted by the **Designated Community**, since:
  - All systems involved **are under the same trusted ownership**
  - **Adequate security policies** are enforced
  - **Access is restricted** to registered users

Anyway, in the final recommendations the gathering of some additional evidence has been proposed to allow tracing changes of custody which may be relevant from a legal point of view

Example: specific AER for INGESTING images into the LTDP repository: the items

- ***AEI-1. Event type***: INGEST
- ***AEI-2. Original identifier:*** identifier of the report.
- ***AEI-3. Identifier in the LTDP system:*** ID-DOC generated by Scryba
- ***AEI-4. Context information:*** DICOM identifier of the study dossier to which the report refers.
- ***AEI-5. Date and time of the ingestion:*** identified by the certified timestamp or (in other cases) by a registry system (more persistent than timestamp)
- ***AEI-6. Identification and authentication data of the LTDP system administrator:*** information generated by Scryba
- ***AEI-7. Assessment on the authenticity and provenance:*** outcome of controls on the digital signature and other relevant information collected in the pre-ingestion phase
- ***AEI-8. Digest of the AIP***: from the certified timestamp.

Example: Authenticity Protocol for INGEST

- **CI type**: *Radiology Information System - Digitally signed medical reports*
- **Event type**: *LTDP-INGEST*
- **Agent**: *Administrator of the Scryba system*
- **AER**: (as defined above)
- **AS sequence**:

  - **AS-1**: check provenance
  - **AS-2**: check integrity
  - **AS-3**: check context
  - **AS-4**: generate internal identifier
  - **AS-5**: generate timestamp
  - **AS-6**: generate **AEI**: *Original identifier*
  - **AS-7**: generate **AEI**: *Internal identifier*
  - **AS-8**: generate **AEI**: *Context information*
  - **AS-9**: generate **AEI**: *Date and time*
  - **AS-10**: generate **AEI**: *Administrator data*
  - **AS-11**: generate **AEI**: *Assessment of authenticity and provenance*
  - **AS-11**: generate **AEI:** *Digest of the AIP*

Sample Authenticity Step

- Each **AS** is structured as a sequence of **Elementary Actions**

AS-1 for INGEST

- **AS-1.1**: get the digital signature certificate from the pkcs#7 file
- **AS-1.2**: get the original digital certificate from the Certification Authority
- **AS-1.3**: check the certificate in the pkcs#7 file against the original certificate
- **AS-1.4**: check the expiration date in the digital certificate against the current date
- **AS-1.5**: get the revocation list from the Certification Authority and check it
- **AS-1.6**: if any of the checks in **AS-1.3**, **AS-1.4** and **AS-1.5** fails then abort ingestion

Feedback from the case study experience

- Case studies have been crucial in validating the model
  - The model displayed substantial robustness and flexibility
  - Practical experience has suggested improvements
- Better understanding of the problems
  - Central role of authenticity in the preservation process
  - How authenticity contributes to Trust
- Providing a systematic way to assess current practices
  - Should become part of the audit and certification process
- Results for individual case studies
  - Problems and deficiencies have become evident to the management of the repositories
  - A systematic way to fix them has been suggested

4. Case studies

4.2. Authenticity tools for assessing provenance

APARSEN SCIDIP-ES Cooperation

SCIDIP-ES
SCIENCE DATA INFRASTRUCTURE FOR PRESERVATION - EARTH SCIENCE

EU funded CP & CSA project that aims at providing an Infrastructure for the implementation of Long Term Data Preservation (LTDP), specifically for Earth Science domain

- Cooperation allowed to achieve:
  - Definition of Interoperable structures for the exchange of **authenticity evidence**
  - Implementing an **authenticity management service**

Main Contributions 1/2

- exploiting, adapting and extending the ***Open Provenance Model (OPM)*** formalism to model the CI provenance information (as a ***provenance graph***), and, for this purpose, adapting and extending it to meet our specific requirements;

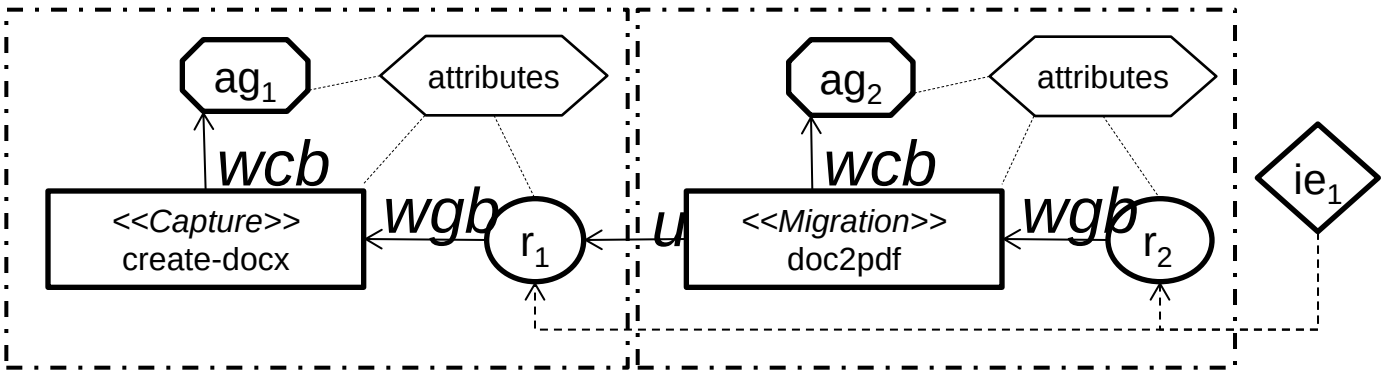
- defining a set of ***standardized XML-based structures***, to represent both the provenance graph and the authenticity evidence gathered and preserved in connection with CI transformations and changes of custody

Main Contributions – 2/2

- achieving **interoperability** among different repositories in managing the authenticity evidence, through the definition and the reference to a *common dictionary*, based on ***PREMIS*** *Data Dictionary for Preservation Metadata*, increasingly recognized and implemented in the digital preservation community;

- *implementing a prototype version of an* ***authenticity management service*** *that provides a set of basic functions (specifically based on provenance information) for the management of the authenticity evidence according to our model through an API interface.*

Augmented Provenance Graphs



Nodes

- r — Artifact/Representation
- Process/Transformation
- ag — Agent

Edges

- wcb – wasControlledBy
- u – used
- wgb - wasGeneratedBy

ECL AP2

Attributes of Graph Nodes

ENTITY	ATTRIBUTE	DESCRIPTION
AGENT	Identification	Personal identification and authentication data
	Role	Role within the repository and in the transformation
REPRESENTATION	Reference	Identifier within the repository
	Type/Structure	Internal structure of the representation, files composing it
	Format	File formats, version
	Fixity	Hash method, and hash file values
TRANSFORMATION	Type	Transformation type: creation, migration, transfer, etc.
	Timestamp	Day and time the transformation was performed
	Tools	Software application
	Controls	Report of controls performed by the agent

Interoperable Authenticity Management

- Structure of the Authenticity Evidence Record

SECTION	ELEMENTS	DESCRIPTION
HEADER	**IntellectualEntity**	Reference to the Intellectual Entity
	Sources	External AERs referenced in this AER
ENTITIES	**Agent**	Agent responsible for controlling the transformation
	Transformation	Transformation described in this AER
	Representation	Representation generated by the transformation
DESCRIPTION	**Report**	Additional evidence including report on controls

XML Structure – 1/2

Header

```
<evidence-record label="AER-002">
    <intellectual-entity label="ie1">
        <identifier type="UniSapienza" value="UniSap-Salza-2013-021-ECLAP"/>
        <annotation value="Preserving authenticity ECLAP2013"/>
    </intellectual-entity>
    <sources>
        <source value="AER-001"/>
    </sources>
```

Entities

```
    <agent label="ag2">
        <identifier type="Italian Fiscal Code" value="SLZSLV48C05H501O"/>
        <type value="Person"/>
        <annotation value="Agent name: SILVIO SALZA"/>
    </agent>
    <representation label="r2">
        <identifier type="URI" value="https://archive.uniroma1.it/docs/SalzaECLAP13"/>
        <type value="file"/>
        <format value="pdf " version="7.1"/>
        <annotation value="PDF version of final draft"/>
    </representation>
    <transformation label="doc2pdf">
        <identifier type="DisEventId" value="E-2013-02-19-000119"/>
        <type value="Migration"/>
        <software  swtype="application" swname="Adobe Acrobat Pro" swversion ="9" />
        <annotation value="convert docx file into pdf; include fonts"/>
    </transformation>
```

XML Structure – 2/2

Description

```
<report>
    <datetime value="2013 February 21 18:00:12"/>
    <used value="AER-001:r1" />
    <fixity type="MD5" value="0f218e0e483cc7937bd81d354b520e7"/>
    <significant-properties>
        <significant-property type="page count" value="12 "  outcome="true"/>
        <significant-property type="page breaks correspond " value= "11" outcome="true"/>
    </significant-properties>
    <agent-assessment value="true"/>
    <annotation value="All fonts have been compared in the two versions and correspond"/>
</report>
</evidence-record>
```

APARSEN-SCIDIP-ES cooperation

- Model of the CI provenance as a ***provenance graph***,
  - by adopting the ***OPM*** formalism and by adapting and extending it in order to meet specific requirements;
- Definition of a set of ***standardized XML-based structures***,
  - to represent both the provenance graph and the authenticity evidence;
- Definition of a common terminology to support the **interoperability** of the authenticity evidence,
  - through the definition and the reference to a *common dictionary*, based on ***PREMIS***;
- Implementation of a prototype version of an ***authenticity management service***
  - that provides a set of basic functions for the management of the authenticity evidence according to our model through an **API interface**.

5. Conclusions and references

Conclusions

- **Authenticity of Digital Content Information** is affected by **relevant events**: **transformations** and **changes of custody**
- **Authenticity Evidence** should be systematically collected **along the whole CI lifecycle/business process**
- **APARSEN** proposes a **systematic methodology**:
  - **Formal model** based on a core set of **events**
  - **Event templates** to specify controls and evidence to be gathered
  - **Authenticity Evidence Records** to ensure interoperability
  - **Operational guidelines** to guide implementation of the model
- The methodology has been **tested on several case studies**, by successfully implementing it in a variety of environments

References: InterPARES and CASPAR

- **M. Guercio**, *Modeling authenticity in CASPAR (2009)*, http://www.casparpreserves.eu/training/advanced-digital-preservation-training-lectures/03.html
- **D. Giaretta, B. Matthews, J. Bicarregui, S. Lambert, M. Guercio, G. Michetti and D. Sawyer**, *Significant properties, authenticity, provenance, representation information and OAIS*, *iPRES 2009, The Sixth international conference on the preservation of digital objects: proceedings*, California Digital Library, 2009, pp. 67-73
- Mi. Factor, E. Henis, D. Naor, S. Rabinovici-Cohen, P. Reshef, S. Ronen, G. Michetti, M. Guercio, *Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage*, *TaPP '09. First Workshop on the Ttheory and Practice of Provenance. San Francisco, 23 February 2009* http://static.usenix.org/event/tapp09/tech/full_papers/factor/factor.pdf

References: APARSEN - 1

- **D24.1 - Report on authenticity and plan for interoperable authenticity evaluation system,** 2012**,** http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D24_1-01-2_3.pdf
  - Detailed analysis of the state of the art (projects and standard); proposal of a common view for capturing and evaluating authenticity evidence in a standardized way; development of a consistent methodology and of concrete guidelines to allow interoperability and support changes in data holders and processing workflows; analysis and discussion of secure logging mechanisms

.

References: APARSEN - 2

- **D24.2 - Implementation and testing of an authenticity protocol on a specific domain**, 2012, http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D24_2-01-2_2.pdf
  - Test the methodology and the guidelines to check how they specialize on specific environments, case study analysis in different environments, to explore the current practices and to propose improvements , proposal and implementation of authenticity protocols (according to the CASPAR methodology)

References: APARSEN - 3

- Silvio Salza, Mariella Guercio, *Authenticity management in long term digital preservation on medical records*, in *I-Pres. Proceedings of the 9th International Conference on preservation of digital objects. Toronto, October 1 – 5, 2012*, University of Toronto: 171-179, https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf
  - *Overview of test audits,* http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D33_1B-01-1_0.pdf

6. Open questions and exercise

Open questions to discuss

- Is the authenticity and its evidence a **crucial** question for **any digital repository**?
- The authenticity assessment made on the basis of evidence can be concretely **supported by a controlled environment and standardized dedicated workflows**?
- Which is the level of **sustainability** of the APARSEN approach? Which **tools and service** could make this effort more feasible?

Authenticity Evidence Record: identify a scenario

- identify a type of **content Information** to be preserved and the specific scenario for its preservation
  - if transferred from the creation to the preservation
  - if transferred from one repository to another
  - if a change occurs within the repository

- These scenarios of preservation implies different elements to be documented for ensuring or supporting authenticity evidence:
  - provide a list of AEI relevant for building an AER (according to the examples discussed)

Authenticity Evidence Record: define the relevant events and its items for AER: see the example

- ***AEI-1. Event type***: (INGEST, AGGREGATE, EXTRACT, MIGRATE, DELETE, TRANSFER)
- ***AEI-2. Original identifier***
- ***AEI-3. Identifier in the LTDP system***
- ***AEI-4. Context information***
- ***AEI-5. Date and time of the ingestion***
- ***AEI-6. Identification and authentication data of the LTDP system administrator***
- ***AEI-7. Assessment on the authenticity and provenance***

Authenticity Evidence Record: define the steps for an authenticity protocol: see the example

- **AS-1**: check provenance
- **AS-2**: check integrity
- **AS-3**: check context
- **AS-4**: generate internal identifier
- **AS-5**: generate timestamp
- **AS-6**: generate **AEI**: *Original identifier*
- **AS-7**: generate **AEI**: *Internal identifier*
- **AS-8**: generate **AEI**: *Context information*
- **AS-9**: generate **AEI**: *Date and time*
- **AS-10**: generate **AEI**: *Administrator data*
- **AS-11**: generate **AEI**: *Assessment of authenticity and provenance*
- **AS-11**: generate **AEI:** *Digest of the AIP*

Authenticity Evidence Record: define the sequence of each step for an authenticity protocol: see the example

- **AS-1.1**: get the digital signature certificate from the pkcs#7 file
- **AS-1.2**: get the original digital certificate from the Certification Authority
- **AS-1.3**: check the certificate in the pkcs#7 file against the original certificate
- **AS-1.4**: check the expiration date in the digital certificate against the current date
- **AS-1.5**: get the revocation list from the Certification Authority and check it
- **AS-1.6**: if any of the checks in **AS-1.3**, **AS-1.4** and **AS-1.5** fails then abort ingestion

SAPIENZA
UNIVERSITÀ DI ROMA

Thank you for the attention!

maria.guercio@uniroma1.it

APA RSEN
Alliance Permanent Access to the Records of Science in Europe Network

cini consorzio interuniversitario nazionale per l'informatica