AUTHENTICITY AND OAIS. THE CASPAR MODEL AND THE INTERPARES PRINCIPLES & OUTPUTS

Mariella Guercio

Delos Summer School
Tirrenia, 11 June 2008

1. CASPAR and InterPARES. The relevance of cooperation and continuity among projects
2. State of the art: the authenticity concepts
3. Critical issues
4. Managing authenticity: tools to develop for Caspar framework
5. Managing authenticity: use cases
6. Elements of an authenticity conceptual model for preservation

Information Society and Media

1. CASPAR AND INTERPARES. THE RELEVANCE FOR COOPERATION AND CONTINUITY AMONG PROJECTS

Information Society and Media

The CASPAR Consortium

CCLRC

metaware

acs
ADVANCED COMPUTER SYSTEMS

UNESCO

ircam
Centre
Pompidou

UNIVERSITY of GLASGOW

IBM®

CIANT

@semantics
worldwide information integration

CNRS
CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

ina

FORTH
ICS

STUDIOR UNIVERSITAS PVBLICA VRBINATEN

ENGINEERING
INGEGNERIA
INFORMATICA

Interdisciplinary Centre for Scientific Research in Music
University of Leeds
www.iccsrim.org.uk

UNIVERSITY OF LEEDS

esa

THE CASPAR PROJECT: THE ASSUPMTIONS

- Need to preserve information & **knowledge** – not just “the bits”
- Need to manage knowledge to keep archives **alive** through time
- Preservation is a **process**
  - **transforming content (migration, emulation, etc.) to adapt it to new constraints of rendition and playability**
  - **enriching content to preserve its intelligibility and (re)usability (not just rendering)**
  - **ensuring the integrity and authenticity of the digital resources**
- OAIS provides a **general framework**:
  - **current implementations deal more with format than the interpretation of data**
  - **CASPAR proposes a richer implementation for dealing with content interpretation and authentic objects preservation**

THE CASPAR PROJECT: THE OBJECTIVES

- **CASPAR methodology:** to lay the foundation for all future preservation activities
- **CASPAR components :** to create key advanced components to use in all the preservation activities
- **CASPAR framework** : to create the long-term autonomous system to support all the preservation activities
- **CASPAR testbeds**: to demonstrate the validity of the CASPAR framework with heterogeneous data and a variety of innovative applications
- **CASPAR open system: user oriented** able to interoperate with as many different systems as possible and to be operated and re-implemented in the framework of existing preservation solutions.

CASPAR TESTBEDS

- Three testbeds
  - **Cultural: UNESCO**
  - **Performing Arts: INA , IRCAM, CIANT**
  - **Scientific: European Spatial Agency (with CCLRC now SSCF)**

- Specific requirements on preservation (technical, delivery, legal)
- Specific research issues
- Specific user communities
- Common infrastructure elements

THE COMPLEXITY OF THE PROJECT

- The testbed institutions are not traditional and dedicated memory institutions: they have the problem and some degree of awareness, they run a digital keeping environmen; they have (had) not an accumulated knowledge in preservation
- Because of the nature of the resources (scientific, performing arts, artistic activity) the crucial goal concerns not only and mainly the contents, but the process, the behaviour, the *performance of the preserved bits* also with reference to their future use at high degree of granularity and trustworthiness (with respect of authenticity, accuracy and reliability of the resources and their metadata)
- The project ambition is to create real and open and implementable software compliant with OAIS details (not a Bible to leave in the drawer but a daily prayer book)

THE INTERPARES PROJECTS

- **INTERPARES 1 GOAL**: To develop the **theoretical and methodological knowledge** essential to the **permanent preservation** of **authentic records** generated and/or maintained electronically, and, on the basis of this knowledge, to formulate model policies, strategies and standards capable of ensuring that preservation (1999-2002)
- **INTERPARES 2 GOAL**: To ensure that the society's recorded memory digitally produced in **dynamic, experiential and interactive systems** in the course of artistic, scientific and e-government activities **can be created in accurate and reliable form and maintained and preserved in authentic form**, both in the long and the short term, for the use of those who created it and of society at large, regardless of digital technology obsolescence and media fragility (2003-2006)
- **INTERPARES 2 GOAL**: To enable **small and medium sized public and private archival organizations and records programs**, which are responsible for the digital records resulting from government, business, research, art and entertainment, social and/or community activities at **national level in different jurisdictions**, to **preserve over the long term authentic records that satisfy the requirements of their stakeholders and society's needs** for an adequate record of its past (2007-2011)

THE MAIN INTERPARES FINDINGS

- A common and consistent **terminology**:
- **Data**: the smallest meaningful piece of information
- **Information**: a message intended for communication across time and space
- **Document**: recorded information (i.e., information affixed to a medium in an objectified and syntactic form)
- **Record**: any document created (i.e., made or received and set aside for action or reference) by a physical or juridical person in the course of activity as an instrument and by-product of it

THE MAIN INTERPARES FINDINGS - 1

- A more nuanced [open] concept of record, although still consistent with and within the boundaries of the traditional definition
- A more nuanced [open] concept of trustworthiness, encompassing reliability, accuracy, authenticity and authentication

TRUSTWORTHINESS

Reliability

The trustworthiness of a resource as a statement of fact,

based on:

- the competence of its author
- the controls on its creation

Accuracy

The correctness and precision of a resource's content

based on:

- the competence of its author
- the controls on content recording and transmission

Authenticity

The trustworthiness of a resource to be what it purports to be, untampered with and uncorrupted

based on:

- identity
- integrity

THE MAIN INTERPARES FINDINGS - 2

- **Metadata**: the attributes of the records [resources] that demonstrate its identity and integrity (authenticity)

- **Digital Components**: entities that either contain one or more records [resources] or are contained in the record [resource] and require a specific preservation measure

THE MAIN INTERPARES FINDINGS - 3

- **Stored record**: the digital component(s) used in re-producing one or more than one record, which include the data to be processed in order to manifest the record (content data and form [structure+appearance] data) and the rules for processing the data, including those enabling variations (composition data)
- **Manifested record**: the visualization or materialization of the record in a form suitable for presentation to a person or system. Sometimes, it does not have a corresponding stored record, but is re-created from fixed content data when a user's action associates them with specific form data and composition data (e.g. a record produced from a relational database)

THE MAIN INTERPARES FINDINGS – 4. TYPES OF RECORDS

Static: They do not provide possibilities for changing their manifest content or form beyond opening, closing and navigating: e-mail, reports, sound recordings, motion video, snapshots of web pages

Interactive: They present variable content, form, or both, and the rules governing the content and form of presentation may be either fixed or variable (more possibilities for presentation)

Not-dynamic: the rules governing the presentation of content and form do not vary, and the content presented each time is selected from a fixed store of data. Ex. Interactive web pages, online catalogs, records enabling performances—they are records

Dynamic: the rules governing the presentation of content and form may vary—they are potential records

THE LIMIT OF THE PRESERVER CAPACITY IN INTERPARES

- The preserver can only preserve what it receives from the creator by making an authentic copy of it, and has no right to stabilize it or alter its documentary form (structure, context, provenance) — only its digital presentation, or format

2. STATE OF THE ART: THE AUTHENTICITY CONCEPTS

The analysis of the projects output
from the University of Urbino team

CASPAR AUTHENTICITY POSITION PAPER -1

- Goal: to define how and on which basis authenticity has to be managed in the digital preservation process to ensure the trustworthiness of digital resources

- The paper has also tried to define the conceptual basis of authenticity for the CASPAR project in terms of a common glossary

- The glossary and the analysis of the key components of authenticity are based on the main results of international community projects, specifically InterPARES, and focused on the interconnections (and critical remarks) between these results and the OAIS conceptual model.

CASPAR AUTHENTICITY POSITION PAPER -2

- The main concepts used in this position paper are based on an international research project (InterPARES), and have been developed in the last 8 years and verified in 15 countries with the support of the S. Diego Super Computer Centre (SDSC)

- These concepts have also been implemented in the US by the National Archives and Records Administration (NARA), but also explicitly used in the recent rules (under definition) by the Italian government for digital repositories and in the most important projects and implementations at international level in the archival sector (see the definitions used in the last two days)

3. CRITICAL ISSUES

MANAGING AUTHENTICITY: A KEY CONCEPT

- Authenticity is never limited to the resource itself, but is extended to the information/document/record system, and thus to the concept of **reliability**.
- Authenticity is concerned with **control over the information/document/record creation process and custody**.
- The **verification** of the authenticity of a resource is related to the **reliability** of the system/resource, and this reliability should prove that it is fully documented with reference both to the **creation process and to the chain of preservation**.
- The authenticity as a result of process has to be documented (if possibly, automatically and on a modular basis) and ensure a high level of quality in the results by maintaining evidence of trustworthiness of the resources and their metadata.

Integrity

- 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

- While **the maintenance of the bit flow is not always necessary**, the completeness of the ‘intellectual form’ is required, especially with respect to the original ability to convey meaning e.g. maintenance of colours in a map, columns in a spreadsheet, etc.

INTEGRITY ACCORDING TO THE INSPECT PROJECT

- The physical integrity of a resource i.e. **the original bit stream can be compromised**, but the **content structure** and the **essential components must remain the same**
- Any successful preservation strategy must reconcile the requirement to maintain the fixity/integrity of that logical information object, with the inevitable transformation of the technical environment in which the object resides.

Identity

- A crucial point is that *identity* must be intended in a very wide meaning: the identity of a resource refers not only to its unique designation and/or identification (it is not a question of persistent identifier)

- **Identity** 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 **contexts (administrative, legal, documentary, technological, social)**
- This broad concept is relevant for major types of digital resources (not necessarily of archival nature): more complex are the digital environments, more complex is the identity to ensure (see the digital components within a web site)

Need to cope with authenticity

- Need to develop **tools** and **methods** that ensure authenticity of objects information (specifically the possibility for their verification and presumption) in the course of the creation and the preservation process
- Any environment involved in the preservation process has to cope with authenticity issues (see R. Ross and M. Hedstrom, Preservation research and sustainable digital libraries", *International Journal on Digital Libraries*, 5, n. 4(2005): 317-324, http://www.dijournal.org.

THE OAIS WEAKNESS ACCORDING TO THE INTERPARES RESEARCH - 1

- The OAIS seems not always to focus on the right issue, specifically when many of its efforts are dedicated to distinguish information and records.
- The real challenge has to do with the process of preservation and the presumption of authenticity any archives is able (and arguably required) to make when ingesting information packages for storage and/or dissemination:
  - **Regardless of whether the packages in question are records, their preservation as digital objects, and the designated communities abilities to reuse the objects, rely on some measure of authenticity**

THE OAIS WEAKNESS ACCORDING TO THE INTERPARES RESEARCH - 2

- Specifically, the OAIS definition ***does not explicitly*** remark upon the authenticity concept.
- OAIS offers a definition of **fixity** (as “a protective shield that protects the Content information from undocumented alternation”)
- This is not a definition of authenticity, but only of a technical tool or mechanism used to **authenticate what is stored (AIP) against what was received (SIP) and possible against what is disseminated (DIP)**
- In this sense fixity is “but one characteristic among many for helping verify authenticity”: it does not allow any comparison of what is stored (AIP) against what was created and used before ingest (SIP)

THE OAIS WEAKNESS ACCORDING TO THE INTERPARES RESEARCH - 3

- Fixity in OAIS is associated with **authentication**.
- The authenticity is the **status of being authentic**, while the authentication is the activity that provides means for further demonstration that something is authentic (a means of declaring the authenticity of a record at one particular moment in time -- possibly without regard to other evidence of identity and integrity).

THE OAIS WEAKNESS ACCORDING TO THE INTERPARES RESEARCH - 4

- OAIS should focus more on the **relationships between the authenticity and the ingestion** with reference to the interaction with the Producer, the negotiation for and acceptance of information and the function of ingest to allow or require that content information include an assertion of authenticity (according to the principles of different responsibilities in the management phases of the digital resources).

THE PROPOSALS OF INTERPARES AT THE OAIS REVIEW

- Include in the glossary new definitions: **authenticity**, **long term preservation**
- Change the definition of **provenance information**
- **Define the responsibility of the OAIS Archive with reference to the producer tasks,** that is to recognize that responsibilities and organizational aspects have to be considered as different management phases (production/preservation/dissemination) even if all would be planned as early as possibile and within a global and integrated approach.

Information Society and Media

4. MANAGING AUTHENTICITY: TOOLS TO DEVELOP FOR CASPAR FRAMEWORK

REQUIREMENTS - 1

- Authenticity Management Tools have to **monitor and manage protocols and procedures across the custody chain** to ensure that the elements required to evaluate the authenticity could be captured as metadata at any time from the creation to the preservation phase and provided (as detailed as required) in the course of the dissemination action.

REQUIREMENTS - 2

- Even if a digital resource is authentic or not**, authenticity cannot be evaluated by means of a boolean flag** telling us whether a document is authentic or not.
- **there are degrees in the capability to verify and presume authenticity**: the certainty about authenticity is a complex goal.
- We have to design all the mechanisms and tools keeping in mind that we could have alteration, corruption, lack of significant data and so on, and we need tools, mechanisms and 'weights' to understand their relevance and their impact on authenticity

- The consequence is that ensuring authenticity means
  - **providing a proper** set of attributes **related to content, context and processes/procedures, and**
  - verifying/checking **(possibly against a metrics and documented processes) the completeness or the alteration of this set (and its nature)**

REQUIREMENTS - 3

- Authenticity Management Tools have to identify mechanisms for ensuring the maintenance and verification of the authenticity in terms of identity and integrity of the digital objects.
- These tools have to provide **content and contextual information relevant to authenticity**, i.e. to the identity and integrity profile, **all along the whole preservation process** (including information related to the preservation process)

REQUIREMENTS - 4

- The **most critical issues** are the right attribution of **authorship**, the identification of **provenance** in the life cycle of digital resources, the insurance of **content integrity** of the digital components and their relevant contextual relationships, and the provision of mechanisms to allow future users to **verify the authenticity** of the preserved objects or, at least, to provide the capability of evaluating their reliability as part of the authenticity presumption process

REQUIREMENTS - 5

So these requirements imply working on:

- Authorship attribution mechanisms and provenance control
- Content and contextual relationships
- Integrity control mechanisms
- Annotation process

- **Every aspect has to be described and documented (**in the form of metadata, that have to be organized to be correctly available in the preservation system) **at every stage in the management processes** so to have, any time is needed, a sort of 'Authenticity Card' for any object in the repository

SOLUTIONS

- **Identify a set of attributes** (metadata) in order to catch relevant information for the authenticity as it can be collected along the management phases of objects belonging to different domains. This means analysing the main and most promising metadata schemas and their basic components

- **Develop a conceptual model** to describe the dynamic profile of authenticity i.e. **to describe it as process** aimed at gathering, protecting and/or evaluating information mainly about identity and integrity

METHODOLOGY -1

- **Authenticity Team started mapping ISAD onto OAIS** just to have a very general idea of some fundamental information elements which are to be preserved for ‘authenticity purposes’

- This is assumed **as a starting point** to find some more elements by taking into account other resources (i.e. ISAAR, EAD, EAC, PREMIS, InterPARES).
- CIDOC CRM is going to be used and evaluated both as a suitable means of expressing concepts and as a resource giving us clues about relevant aspects needed for consideration, especially about dynamic aspects (temporal entities)

METHODOLOGY - 2

Problems:

- **level of granularity**. Essential authenticity requirements must be clearly identified in order to have neither overload nor exiguity of information

- **variety of domains**. Authenticity methodology and concepts are cross-domain but their deployment is strongly dependent on specific environment. For example:
  - the Reference Information for a book could be ISBN, very specific and not suitable for other typologies
  - the authorship concept is quite ‘easy’ for a book but what about the author of a movie?

METHODOLOGY - 3

Problems:

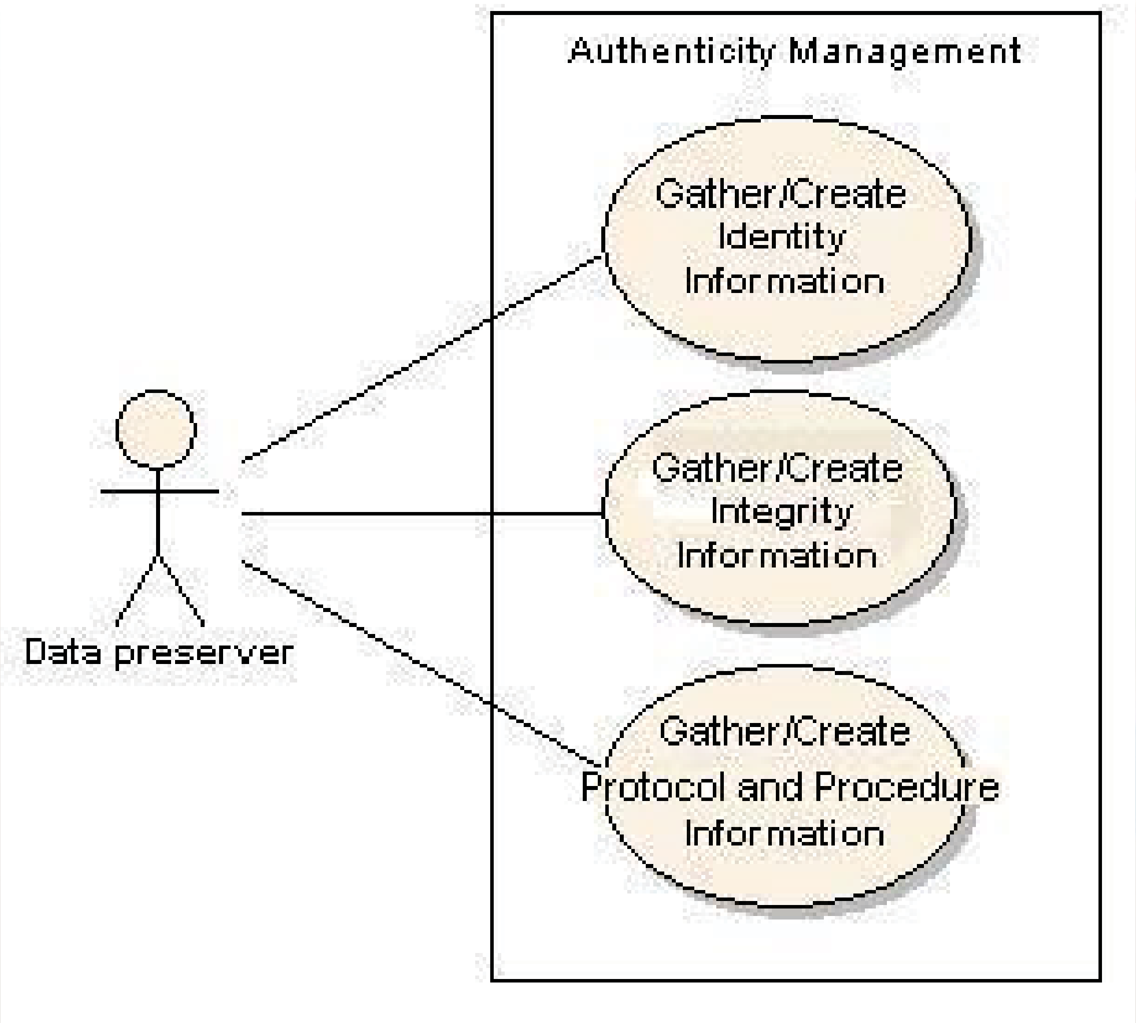
- **overlapping of concepts** coming from different schemas. It's not easy to decide whether an element has to be mapped onto OAIS conceptual elements (e.g. whether the ISAD element "System of arrangement" belongs to either OAIS Provenance or OAIS Context). Anyway, the Authenticity Team recognizes that its aim is to find a set of information elements and assign them to an OAIS category (on the basis of a formal convention)

5. MANAGING AUTHENTICITY: USE CASES

USE CASES MODEL

- The Authenticity Management includes the following use cases:
  - ***Gather identity information***
  - ***Gather integrity information***
  - ***Gather protocol and procedure information***

USE CASES MODEL


Authenticity Management
Gather/Create Identity Information
Gather/Create Integrity Information
Gather/Create Protocol and Procedure Information
Data preserver

GATHER/CREATE IDENTITY INFORMATION

- The *Gather Identity Information* refers to the identification of information elements mainly related to
  - **provenance** (with reference to archival history/chain of custody, origin or source, changes since the creation …)
  - **context** (scope and content, name of action or process, archival relationships, designation, extent, medium, taxonomic systems, reference systems, registration systems, …)
  - **conditions of access and use**
  - **allied materials**
  - …

GATHER/CREATE INTEGRITY INFORMATION - 1

- The *Gather Integrity Information* refers to data integrity checks or validation/verification keys, and to procedures aimed at preventing, discovering, and correcting loss or corruption of records

GATHER/CREATE INTEGRITY INFORMATION - 2

Description: Information is extracted from Fixity Manager to generate a set of data through which evaluate the integrity of the resource and hence its authenticity. The Data Preserver evaluate the quality of information and add any missing data needed to fulfil authenticity requirements

Scenario:

1) PDS-Ingest receives an AIP with a storage request and sends a storage alert to Authenticity Management

2) Authenticity Management gathers information from Fixity Manager and sends an authenticity report to PDS-Ingest, possibly requiring more qualified data

3) When information is complete or anyway it cannot be refined anymore, Authenticity Management associates it to the resource and, in case, it sends an authenticity alert to PDS and to Administration entity

GATHER/CREATE PROTOCOL & PROCEDURE INFO - 1

- The *Gather/Create Protocol and Procedure Information* refers to definition and management of so called *Authenticity Protocols* that model **a sequence of steps** aimed at describing and evaluating the authenticity profile

- The *Gather/Create protocol and procedure information* refers to protective strategies and/or solutions adopted in order to maintain authenticity or that affect authenticity anyway.
- The recursive design of information objects emphasizes the **recursive nature of the problem of authenticity**, so **we need to manage authenticity of content/context information (metadata)** too, and define the policies for its control, for example by recording the responsibilities on creation and/or modification of content/context information

GATHER/CREATE PROTOCOL & PROCEDURE INFO - 2

Description: Information is extracted from RepInfo, PDI, Migration, and Placement Managers to generate a set of data aimed at express the value of preservation chain or rather the overall quality of system in terms of authenticity management. The Data Preserver may add any useful information, including responsibilities on creation and/or modification of content/context information, and evaluation or description of authenticity of content/context information.

Scenario:

1) PDS-Ingest receives a storage request with an AIP and sends a **storage alert** to Authenticity Management

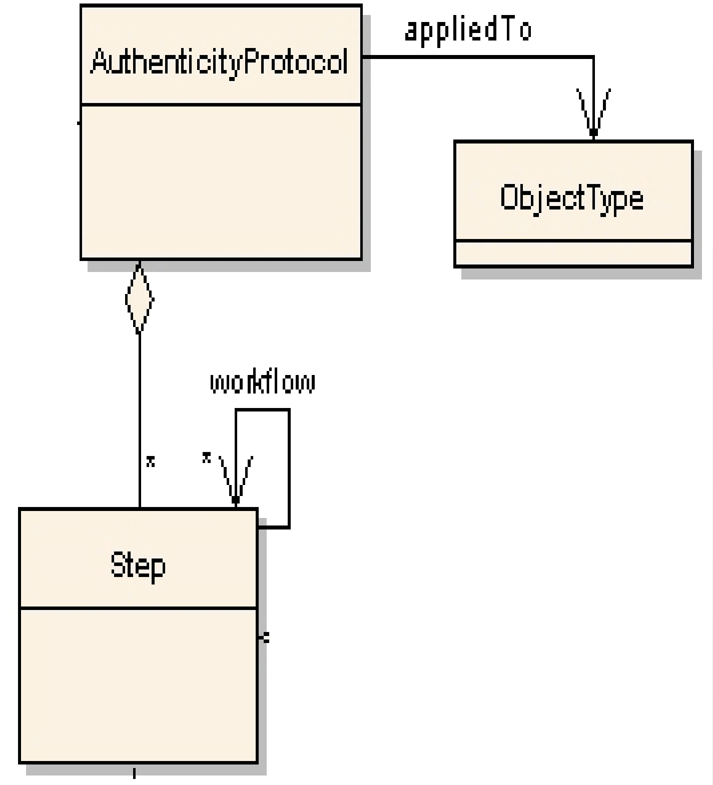
2) Authenticity Management gathers information from RepInfo Manager, PDI Manager, Migration Manager and Placement Manager, and sends **an authenticity report** to PDS-Ingest, **eventually requiring more qualified data**

3) When information is complete or anyway it can’t be refined anymore, Authenticity Management associates it to the resource and, in case, it sends an authenticity alert to PDS and to Administration entity

6. ELEMENTS OF AN AUTHENTICITY CONCEPTUAL MODEL FOR PRESERVATION

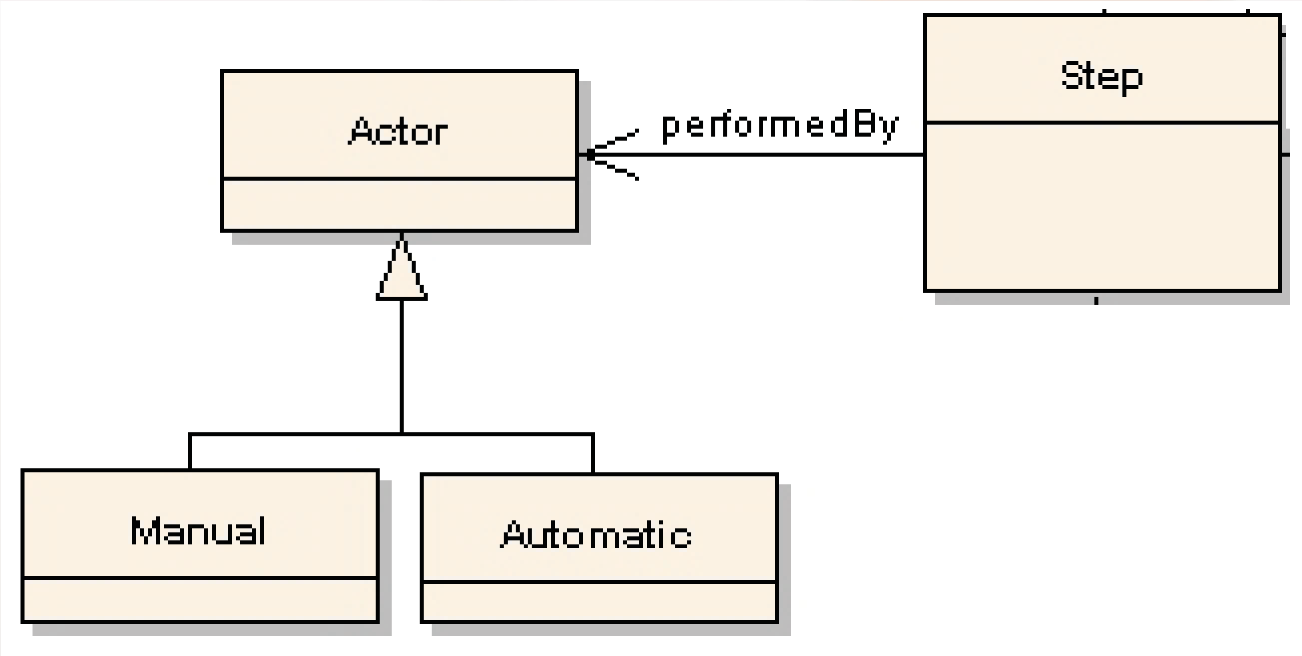
AUTHENTICITY PROTOCOL (AP)

- The protection of authenticity and its assessment is a **process**. In order to manage this process, we need to define the procedures to be followed to assess the authenticity of specific type of objects
- We call one of these procedures an **Authenticity Protocol** (abbreviated as **AP**). An AP is a set of interrelated steps (**Authenticity Step - AS**).



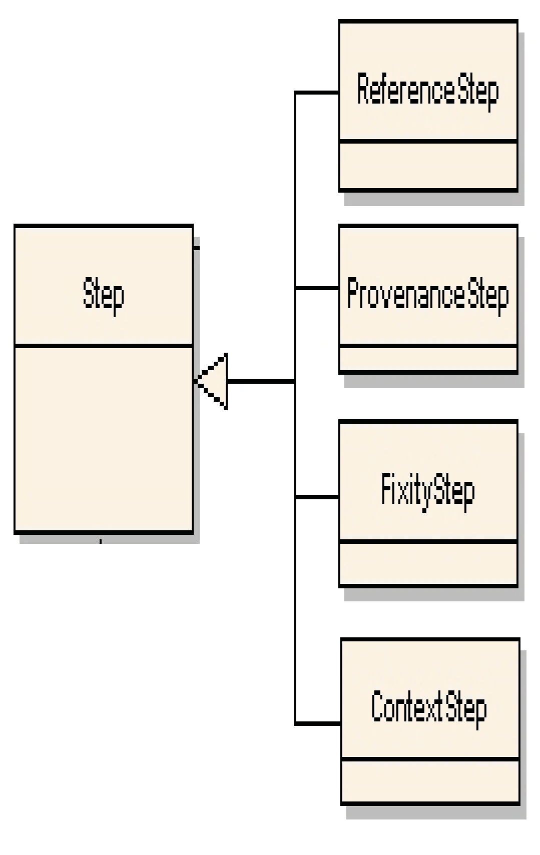
MODELING AUTHENTICITY STEP

- An AS is performed by an **Actor**, who can act either in an automatic (hardware, software) or in a manual (person, organization) way:



AUTHENTICITY STEP (AS)

- There can be several types of ASs. Following OAIS, we distinguish ASs based on the kind of Preservation Description Information required to carry out the AS. Consequently, we have 4 types of steps:
  - Reference Step
  - Provenance Step
  - Fixity Step
  - Context Step



AUTHENTICITY STEP (AS): INFORMATION AND RECOMMENDATIONS

- Since an AS involves a decision, it is expected that it contains at least information about:
  - the criteria that must be satisfied in taking the decision
  - good practices or methodologies that must be followed
  - the actors who are entitled to take the decision.
- Moreover, an AS is defined (possibly) following **Recommendations** and is disseminated as established by a dissemination indication:

AUTHENTICITY PROTOCOL EXECUTION (APE) - 1

- APs are executed on objects belonging to a specific **typology**, in the context of **Authenticity Execution Sessions**.
- The execution of an AP is modelled as an **Authenticity Protocol Execution** (APE for short).
- An APE is related to an AP via the **IsExecutionOf** association and consists of a number of execution steps (Authenticity Step Executions, ASEs for short)

AUTHENTICITY PROTOCOL EXECUTION (APE) - 2

- Every ASE, in turn, is related to the AS via an association analogous to the **IsExecutionOf** association, and contains the information about the execution, including:
  - the **actor** who did the execution
  - the **information** which was **used**
  - the **time**, **place**, and **context** of execution
  - possibly the **outcome** of the execution. Not every step necessarily implies a decision, some steps simply imply collecting information related to a specific aspect of the object, e.g. title, extent, dates, and we are only interested into declaring the step has been done, without any form of evaluation. From a modelling point of view, we could classify steps as *decisional* (and the outcome is the decision) and *non-decisional* ones (having a different kind of outcome as an attribute, e.g. “step done" or "step not completed for such and such reason")

AUTHENTICITY PROTOCOL EXECUTION (APE)- 3

- The *Authenticity Protocol Report* will include also the information on which the execution/evaluation of *Steps* was based

AUTHENTICITY MODEL: AN EXAMPLE

Prot	OBJ	Descr	Steps
Prot1	PDF	…	S1, S2, S3
Prot2	MP3	…	
Prot3	JPEG	…	

Steps		Actor	Def	Rep
S1	R1	Boss of CASPAR	•Do this	Date, Place, Actor
S2	P1	Anyone	•Do other	…
S3	F1	Internal Operator	•Do that	…

Prot	OBJ	Descr	Steps
Prot1-070726	Myfile.pdf	…	exeS1, exeS2, exeS3

exeSteps		Actor	Def	Rep
exeS1	R1	David	• Do this	07-07-26, Rome, David
exeS2	P1	Monica	• Do other	…
exeS3	F1	Carlo	• Do that	…

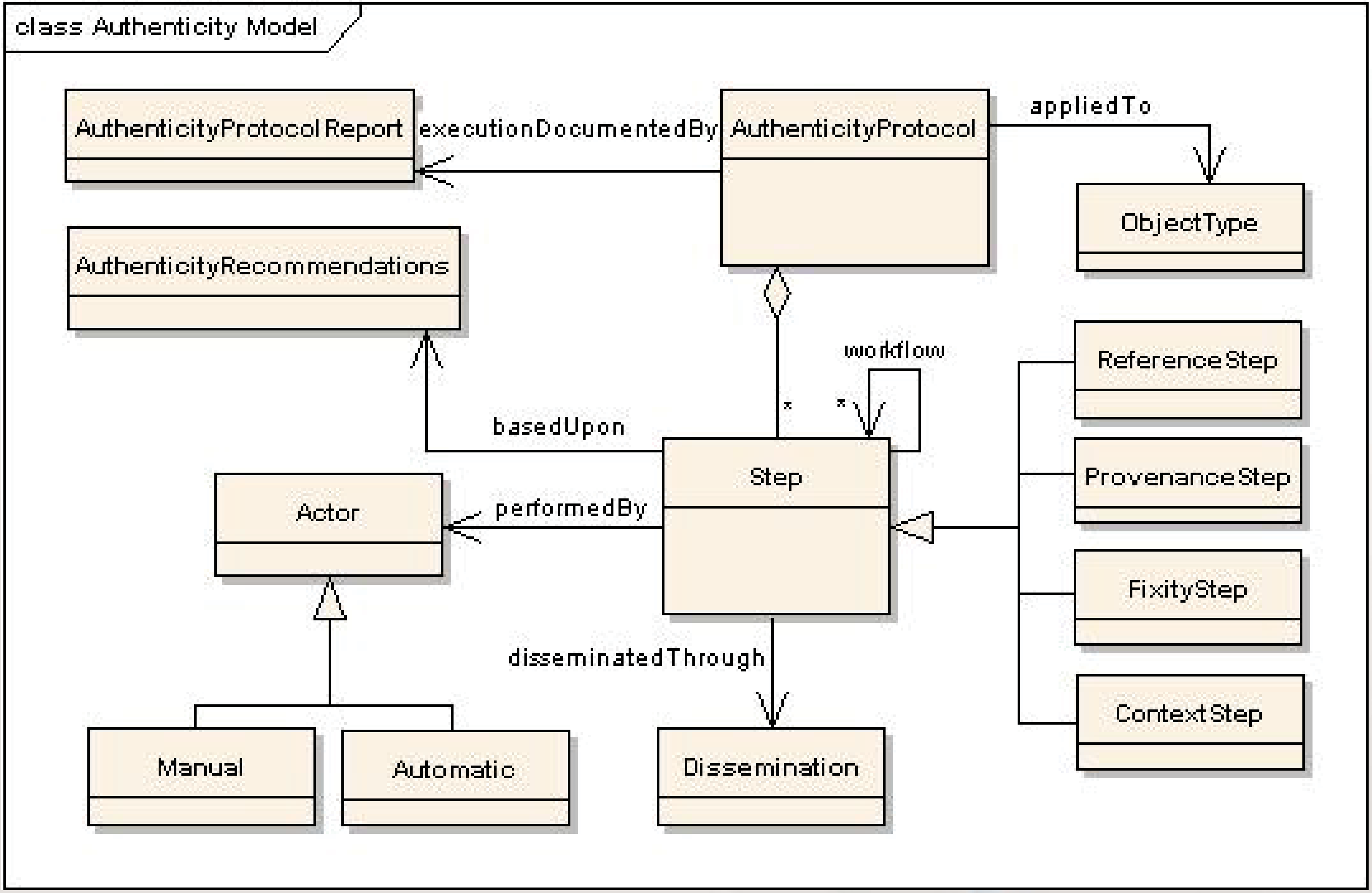
AUTHENTICITY STEP EXECUTION (ASE): ITS EVOLUTION

- Because we are dealing with preservation, we also want our model to be able **to cope with the evolution of both APs and their executions over time**

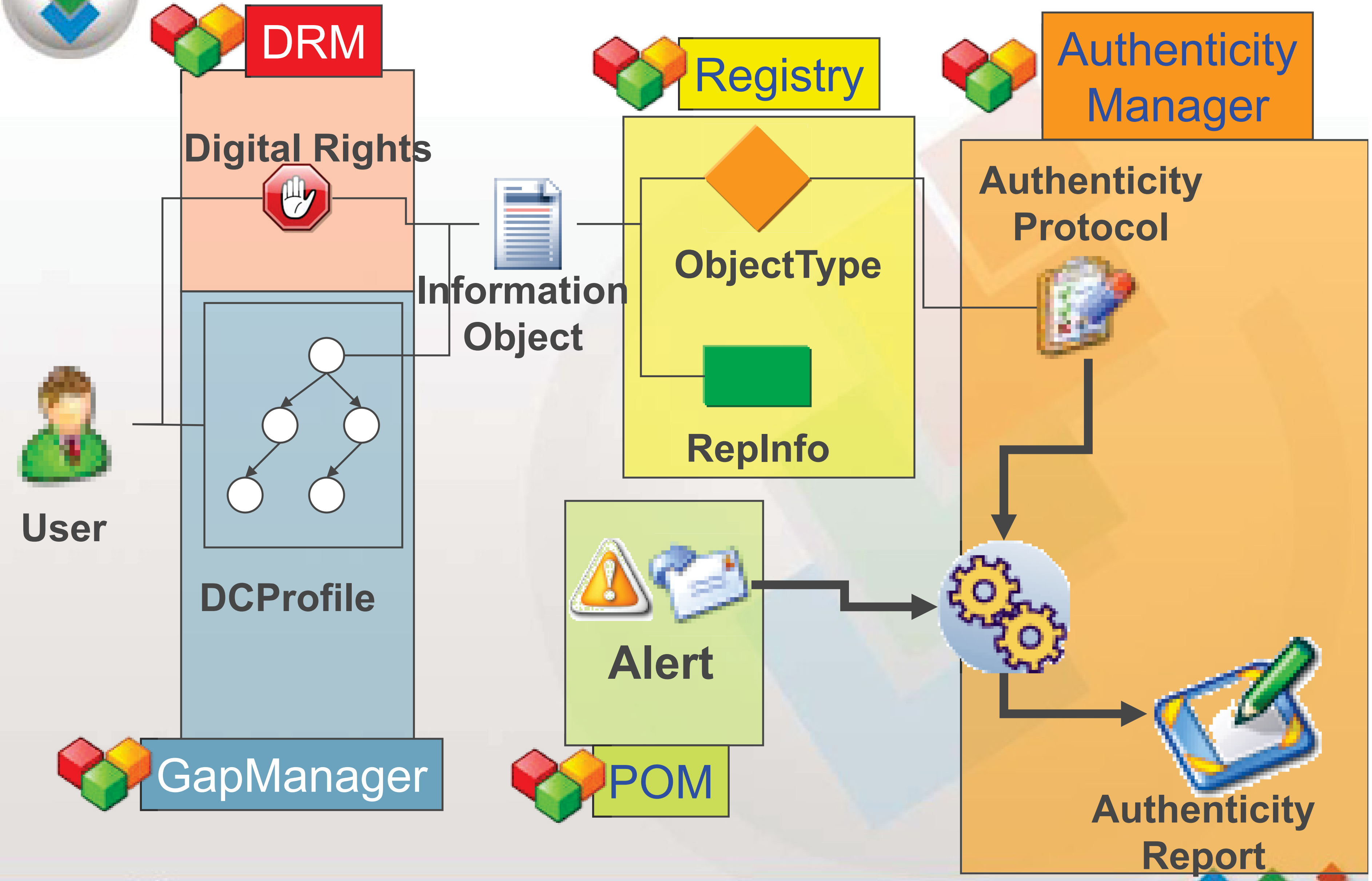
- The evaluation of an AP may concern the **addition, removal or modification of one of the steps making up the AP**. In any case, both the old and the new step are retained, for documentation purposes

- When an AS of an AP is changed, all the executions of the AP which include an ASE related to the changed step, must be revised and possibly a new execution is required for the new (modified) step. Also in this case, the old and the new ASEs must be retained

OVERALL AUTHENTICITY MODEL



Authenticity within the overall CASPAR framework



Future Work

- Refinement of Authenticity Protocols and Steps
- Validation and Test of authenticity procedures in specific testbed(s)
- Evaluation fo compliance with Trusted Digital Repositories recommendations for auditing
  - **Digital Repository Audit and Certification Working Group** **http://wiki.digitalrepositoryauditandcertification.org**
  - **Authenticity tools prototype**

Some conclusions

- We need a consistent terminology specifically if we look for cross domains research
- Cross domains research is fruitful especially within a persistent cooperation environment made available thanks to networks (national and international) of competence centres and becoming more relevant in relation to the increasing complexity of the preservation function (more we know in the field, more we are aware we do not know enough)
- The existence of an open flexible but also persistent space for the research community and the professionals and the institutions involved in each domain is a requirement for a successful approach to the digital preservation
- InterPARES and OAIS seem to offer this basic framework
- National networks, focused on domains but also open for large cooperation, are the key nodes of a fruitful and persistent network able to transform occasional events into a stable program thanks to strong scientific and technical relationships among stakeholders.

Thank you for your attention