Conceptual framework and chain of custody for sustaining the digital trustworthiness

Maria Guercio
Università degli studi di Urbino
Vienna, 12 November 2009

Basic assumptions

- The creation and preservation of digital resources are both **ongoing functions**
- Metadata offer partial solutions and in some cases create new problems to face if their nature and function is not clearly articulated (**attributes/component** of the resource, **external elements** or **implicit information** within the procedural, technological or juridical context
- a **pragmatic effort** is required, but strongly rooted on **consistent theory** and **principles**
- the interdisciplinary approach has to take into account the **promising outputs** of the most recent research projects in the field (PLANETS, CASPAR, INTERPARES, PREMIS)

The methodological outputs from the research environment

- **OAIS as a reference model** to be implemented not only as generic architecture
- **InterPARES as conceptual framework** for interrelating principles, policies and procedures and developing a consistent frame to compare and assess quality and consistency of the digital practices
- **CASPAR** had the specific aim of building a **standardized digital environment** for cultural, scientific and performing arts domains (that is for dynamic sectors which require more complex and really evolving solutions).

The methodological outputs from the research environment: *CASAPAR*

- Specifically the CASPAR conceptual model has included relevant results achieved in the field at international level with the aim of creating institutional digital repositories based on an integrated approach to be applied for differentiated and complex archival and information systems :
  - **InterPARES**,
  - **OAIS**,
  - **TRAC** – Trusted Repository Audit Checklist,
  - **RAC** – Repository Audit and Certification (ISO guidelines),
  - **PREMIS** – Metadata for digital preservation,
  - **CIDOC** Conceptual Reference Model - ISO standard for developing ontologies and mapping metadata schemas with semantic approach and capacity

The conceptual framework and the principle of trustworthiness

- The information and record curation is increasingly based on concept of **trust**, specifically in the digital environment
- 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.

Trust and digital certification the CCSDS recommendations - 1

- "The overall aim of certification is to give confidence to all parties that a management system fulfils specified requirements. The value of certification is the degree of public confidence and trust that is established by an impartial and competent assessment by a third-party. Parties that have an interest in certification include, but are not limited to
- a) the clients of the certification bodies,
- b) the customers of the organizations whose management systems are certified,
- c) governmental authorities,
- d) non-governmental organizations, and
- e) consumers and other members of the public".

http://wiki.digitalrepositoryauditandcertification.org/bin/view/Main/ReqtsForAuditors

Trust and digital certification the CCSDS recommendations - 2

- It requires the identification of reference principles able to inspire *confidence.* This kind of principles includes:
- "- impartiality,
- - competence,
- - responsibility,
- - openness,
- - confidentiality, and
- - responsiveness to complaints".

Trust and digital certification the CCSDS recommendations - 3

In relation to trustworthiness the efficient use of metadata could:

- **foster the credibility** of the repository as trustworthy custodian on the basis of its capacity of securing integrity and authenticity of their digital contents through a standardized accumulation of descriptive and management information,
- **control the cost of descriptive function** "by using a simple encoding scheme and by ingesting metadata on transfer from public sector institutions",
- **enlarge to range of interrelations** by "exchanging finding aid metadata with metadata harvesters from all kinds of communities".

Metadata: the ambiguity of the term

- *A machine or human readable assertion about a resource*
- The term is used today "so ubiquitously and in so many different ways by different communities that it is in peril of losing any specificity"
- It is commonly recognized that metadata are relevant and meaningful if they are themselves "trustworthy and comprehensively managed for as long as they are required", if they are "sufficient, appropriate, understandable and of high quality

Metadata: the timing

- The **timing** of the metadata creation and extraction is crucial.
- InterPARES makes a clear distinction between the metadata identified in the creation phase (as part of **benchmark requirements** that need to be "explicitly expressed and inextricably linked to a record in order for its identity and integrity to be asserted") and the metadata specified by the preserver as trusted custodian (as part of **baseline requirements** "to support the production of authentic copies of digital records after they have been transferred to the preserver's custody".

Metadata: the profiles

- The profiles vary considerably from implementation to implementation and can include:
  - only essential elements,
  - the elements that a given system is able to support or that an institution/individual has sufficient expertise to create,
  - rich and rigorous metadata schemas well specified and provided with analytical framework and standardized documentation

Metadata: dynamic capture

- The dynamic approach for their capture is going to be increasingly relevant and complex, specifically in relation to the range of ways in which they can be automatically acquired
  - business processes,
  - documentary forms,
  - file properties
  - logs,
  - audit trails of any changes of the functionality of the original technical environment,
  - elements to distinguish authoritative resources from draft and derivative versions,
  - links between stored data and manifested content
  - e-mails details for receipt, dispatch and transmission.

Metadata and chain of 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

Metadata are relevant because authenticity and integrity could be evaluated:

- on the basis of the elements on the *face/form* of the resource and its **attributes/metadata**,
- from the **circumstances** carefully **documented and tested through metadata** of its maintenance and preservation: "an unbroken chain of responsible and legitimate custody is considered an insurance of integrity until proof to the contrary"
- from the **integrity of essential metadata** related to the resources handling and preservation as a further requirement for attestation of integrity and authenticity:
  - individuals/offices involved,
  - 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,
- **as inference on the basis of the trustworthiness of the document/information system** in which the documents/information exist.

What is missing

- **consistent and accepted terminology** and **definitions** used across domains and **well understood**
- **development of interrelations** and concrete and open **cooperation** among relevant projects and standardization process with the aim of building an interoperable framework
- **integration** of models, schemas and business solutions to be developed in the **application scenarios** for handling relevant tasks as:
  - authenticity and its presumption,
  - storage systems in independent environment,
  - automation of metadata extraction

What is missing: credible solutions for metadata extraction

- Text categorization (based on machine-learning and supervision)
- Document clustering (based on information retrieval)
- Document classification (based on controlled vocabulary and term extraction): is here ontoly useful in the creation process?
- Mixed approach

But what to do in case of functional classification? Is genre classification based on records/documents type useful?

Sources

- InterPARES 2. *Part Six. Investigating the Roles and Requirements, Manifestations and Management of Metadata in the Creation of Reliable and Preservation of Authentic Digital Entities. Description Cross-domain Task Force Report*, pp. 305-360, www.interpares.org

- Kai Naumann, Christian Keitel, Rolf Lang , "One for Many: A Metadata Concept for Mixed Digital Content at a State Archive", *The International Journal of Digital Curation*, 2009, 2,
- http://www.ijdc.net/index.php/ijdc/article/viewFile/120/123

- M. Day, *Preservation metadata*, http://www.slideshare.net/michaelday/preservation-metadata

- Y. Kim, S. Ross, "The Naming of cats. Automated Genre Classification", *International Journal of Digital Curation*, 2007, 1, http://www.ijdc.net

- Pikka Heutonnen, "Creating Recordkeeping Metadata", *Atlanti*, 9 (2009), pp. 67-76. For the FinnONTO project see www.seco.tkk.fi.