A Digital Infrastructure for Trustworthiness

The Sapienza Digital Library Experience

Angela Di Iorio[1], Marco Schaerf[1], Maria Guercio[1], Silvia Ortolani[1], and Matteo Bertazzo[2]

[1] Sapienza Università di Roma, Rome, Italy
{angela.diiorio,marco.schaerf,maria.guercio,
silvia.ortolani}@uniroma1.it
[2] CINECA, Bologna, Italy
m.bertazzo@Cineca.it

Abstract. The building process of Sapienza Digital Library's (SDL) digital resources was designed for collecting the information required by the Open Archival Information System (OAIS) Preservation Description Information(PDI): Provenance, Reference, Fixity, Context, and Access Rights Information. The Submission Information Packages'(SIP) preservation metadata was encoded in the semantics of the PREMIS standard which is the implementation metadata set, mapped from the OAIS conceptual model. The conformant implementation of the PREMIS standard was one of the principles which permeates the SIP building process. All relevant legal aspects and formal agreements, referred to the organizations involved in the different OAIS functions of the SDL digital repository, were analyzed and structured for their inclusion into the forthcoming AIP management, and for unleashing of the preservation strategies, and for supporting the authenticity of resources.

Keywords: DL Architectures and infrastructures, Long term preservation, Metadata creation, management, and curation, OAIS, METS, PREMIS.

1 Introduction

The Sapienza Digital Library[1] (SDL) is a research project undertaken by Sapienza Università di Roma (Sapienza), the largest Europe's campus, and the Italian supercomputer center Cineca[2], the 9th in the Top500[3], which is a no profit consortium, made up of 54 Universities, 2 Research Institutions and Ministry of Education, University and Research.

The SDL project aims to build an infrastructure supporting preservation, management and dissemination of the past, present and future digital resources, containing the overall intellectual production of the Sapienza University[3].

[1] Sapienza Digital Library http://sapienzadigitallibrary.uniroma1.it (expected on May 2013)
[2] Cineca consortium http://www.cineca.it
[3] Top500 supercomputer sites http://www.top500.org/

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 59–69, 2014.

The digital infrastructure set up by the SDL project was build aiming at the conformance to the OAIS(ISO 14721:2003) functional model[1] and developing compliant services supporting the Long Term Digital Preservation (LTDP).

Both two projects participants have built two repositories that can be defined in OAIS terms like "Cooperating: Archives with potential common producers, common submission standards, and common dissemination standards, but no common finding aids". The interchanging repositories share a common metadata infrastructure based on the most spread metadata standards for digital libraries, and the Information Packages are replicated in both repositories.

The provision of a SIP, equipped with the PDI required by OAIS, was considered an essential requirement in the design of both digital repositories and in the design of the metadata framework on which is based the IP exchange.

2 The SDL Preservation Technical Strategy

Sapienza's organizations, or other organizations in legal agreement with a Sapienza's organizations, will provide digital content for the Sapienza Digital library services supporting the digital curation activities. The SDL project agreement between Sapienza and CINECA has established the commitment of both in making up the services for digital preservation. Regarding to the preservation services, the replication of storage, geographically dispersed, is one of the technical strategy for the trustworthiness[4] of the overall system.

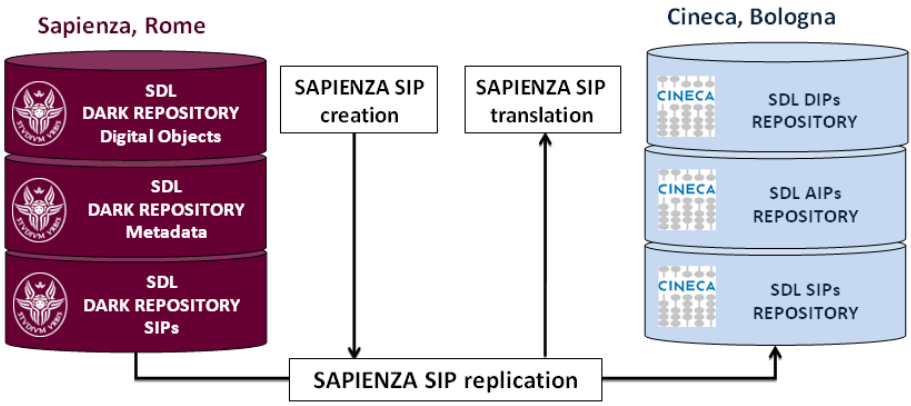

Fig. 1. SDL cooperating repositories, geographically dispersed

The first repository, where Sapienza SIPs are created, is located in Rome at the Sapienza University. Every Sapienza SIP created is then replicated into the CINECA storage, which is located in Bologna.

The Sapienza SIP replicated is ingested and translated into corresponding AIP and DIP, that are managed by the SDL management system based on Fedora Commons[4]. Both systems (Sapienza and Cineca) share semantics about the common standardized description of the original SIP, produced by the Sapienza University.

The Sapienza SIP contains metadata tailored on metadata documents' models, that CINECA technological system translates in provision of services for archiving and dissemination.

The technical level of interaction of the SDL and CINECA archives can be defined, in OAIS terms, as *cooperating* archives considering that the performed activities are based on a standard agreement and they have common SIP and DIP format and related communities of interest. The Sapienza SIPs produced by the University and stored in the Sapieza local dark archive, is replicated in the CINECA archive and ingested and translated in the SDL Archival Information Package and the corresponding DIP, updated with the Events information and provided on request.

As the OAIS specifies "The only requirement for [the Cooperating Archives] architecture is that the cooperating groups support at least one common SIP and DIP format for inter-Archive requests", the SDL framework was designed on metadata specifications that are commonly used for SIP and DIP in the Sapienza-Cineca interchange scenario. In order to support the standard agreement cooperation, "a set of mutual Submission Agreements, Event Based Orders, and user interface standards to allow DIPs from one Archive to be ingested as SIPs by another"[1] was designed, and at this moment is under implementation.

3 Designing the SDL metadata framework for LTDP

The metadata framework conceived for SDL has respected the following requirements oriented to support the LTDP:

- conformant with OAIS, in order to support the OAIS model of information, to fulfill the responsibilities for operating an OAIS Archive, and to underscore the trustworthiness of holding repositories[1];
- prearranged to hold different standard descriptions on which implementing future integration services, supporting the use of wide-ranging knowledge's materials for different designated communities;
- prearranged to the exchange with other digital library systems or other information management systems, maintaining the information about provenance;
- prearranged to the LTDP and equipped with the essential metadata, enabling the long term management.

The arrangement of consistent information supporting the LTDP has followed the structure of the Preservation Description Information (PDI), which is composed by information about PROVENANCE, REFERENCE, CONTEXT, FIXITY, and ACCESS RIGHTS.

[4] Fedora Commons Repository System `http://fedora-commons.org/`

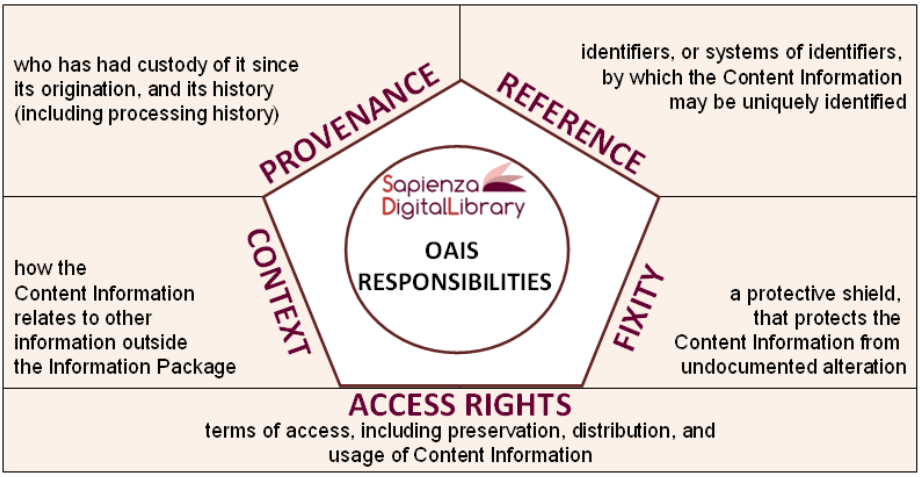

Fig. 2. SDL responsibilities discharging based on a consistent PDI

3.1 The CONTEXT and PROVENANCE Information

The CONTEXT information contains pointers to its environment by means of structured information referred to the originating organization (Sapienza's organizations libraries, museums, investigations departments). The CONTEXT information documents "why Content Information (CI) was created and how it relates to other CI objects existing elsewhere"[1].

The PROVENANCE information, which describes the source of CI, and in particular, "who has had custody of it since its origination, and its history (including processing history)" is provided by the Sapienza's organizations (both domain specific and technical) that produce, own, manage or have the custody of the CI.

Because Sapienza's University is a public institution, usually the business rules, for holding intellectual material, follow national or legal rules, like for example the Italian Author's Rights (civil law[5]) for the Intellectual Property information, or the Italian National Librarian System cataloguing rules for describing CI. It means that Sapienza's organizations, as public bodies, do already provide information following rules publicly and legally established.

Furthermore the provision of consistent CONTEXT and PROVENANCE information makes feasible to sustain both evidence to support the Authenticity of the resources, and the Trustworthyness of repository.

The system's characteristic of providing authenticity evidence, based on the assurance about the reliability of the CI, strongly consists of the ability of acquiring and maintaining unambiguous information about the CONTEXT and PROVENANCE of the managed digital resources.

[5] http://it.wikipedia.org/wiki/Civil_law

3.2 The ACCESS RIGHTS Information

The ACCESS RIGHTS information and documentation corpora (access restrictions, legal framework, licensing terms, and access control) were gathered, selected, modeled, identified and referenced to their own CONTEXT and PROVENANCE information.

The ACCESS RIGHTS Information in SDL contains the access and distribution conditions stated in the Submission Agreement, related to the third party usage and the SDL management, distribution, dissemination and preservation. The Submission Agreement involves both organizations with project responsibility: Sapienza and Cineca.

It also includes the specifications about the application of rights enforcement measures:

- Identification of the properly authorized Designated Community (Access Control, e.g. the access to some SDL objects is allowed to the Sapienza's community, the submission of resources is allowed to specific Sapienza's communities…)
- Permission grants for preservation and for distribution and dissemination (Copyright information)
- Pointers to FIXITY, CONTEXT and PROVENANCE Information
- Information about digital inhibitors like signatures, passwords and other access control mechanisms applied at submission and preservation time
- Legal and licensing framework(s)

The different layers of terms of agreements and actions allowed in the different contexts of submission, preservation, management, access and distribution were properly identified and structured in the documentation system. All specific agreements signed by Sapienza's organizations, responsible for digital curation in SDL and the third party granting the digital content, are unambiguously identified by the system, stored and referenced by related digital resources and collections. The specific agreements are referred to the general agreement, involving Sapienza as owner institution of the Digital Library and Cineca as partner, providing specific technological services, which states the general terms of the standard agreement cooperation.

3.3 The REFERENCE Information

The REFERENCE information was based on an identification system conceived in consideration of the cooperative focus of the project. Every single object, resource, and collection must have the essential information for detecting the originating entity, and the custodian entity, which has the responsibility of the digital curation.

The identification system manages a mechanism for creating identifiers families that are strictly connected to the "real Sapienza's organization", which is responsible for the custody chain (digital curator). At every level of the digital resource, it is possible to get unambiguously information about the origin and, consequently, the history of the resource. Every single SDL object can be reused and repurposed in different contexts, and is provided of all bounding information about its PROVENANCE, and

its originating CONTEXT, which points to the PDI of the source, expressed at collection level. The opportunity of exploiting resources in a referable manner, also allows the flawless interchange with other repositories.

3.4 The FIXITY Information

Automatic production of FIXITYs information is provided at the early SIP creation stage. The FIXITY information is one of the technical requirement about the overall management of the digital resources accessioned by both SDL archives. This means that, likewise the REFERENCE information, at every layer of the SIP building the FIXITY information is automatically produced, following a bottom up method: from the single content objects, going up the metadata objects and finishing with IPs.

3.5 Discharging Responsabilities of the SDL Organizations

The design of the system at this moment was focused on the essential services of ingestion, archiving and dissemination, waiting the forthcoming implementation of a robust preservation management system. Nevertheless the SIP creation workflow was conceived for gathering all information necessary for supporting LTDP and covering information necessary to the "mandatory responsibilities that an organization must discharge in order to operate an OAIS Archive"[1].

The negotiation between SDL archive and the Sapienza's organizations is based on an agreement, which will formally cover all resources submitted to the digital repository. The agreement (at this moment in draft form) establishes the acquisition of properly selected CI, produced by Sapienza organizations, and requires the provision of the bare minimum of information necessary for a consistent PDI, specifically oriented to the Designated Communities.

A similar agreement model was defined for terms of services between Sapienza and external organizations, not belonging to Sapienza University. Those organizations, that are willing to donate resources and to use SDL services, need to sign a legal agreement with one of the Sapienza's organizations, which is declared as "digital curator" of the donated resources.

The aim of the agreement model is obtaining "sufficient control for preservation"[1], gathering copyright implications, intellectual property and other legal restrictions on use, and acquiring the right level of authority to modify Representation Information, in the future contexts of migration.

The Organizations responsible for the content are deputed to define its own Designated Community of consumers, with the support of the SDL Scientific committee, taking into account the harmonization of the domain specific information, with the existing SDL descriptive information, necessary "to enable the Designated Community to discover and identify material of interest" [7].

At this moment, the documentation about policies and procedures is not yet completed and is under revision, but will be easily integrated by the system once the responsible Organizations will be agreed on it. The SDL archive agreement will be also

integrated by the constraint which claims the conformance to the policies and technical rules established by the SDL management.

4 The Preservation Metadata Implementation

Considering the LTDP strategy adopted, the overall SDL SIP building workflow must ensure the basic provision of the preservation metadata, considered mandatory by the PREMIS standard[2], which is the preservation metadata framework mapped from the conceptual structure of the OAIS model. The SDL metadata framework was designed to guarantee the conformance with the PREMIS standard, both on semantic unit and data dictionary level. As stated by the conformance guidelines on the PREMIS implementation[5], the SDL framework design has followed requirements and constraints, defined in the PREMIS Data Dictionary[6], and the SIP building workflow has collected all metadata necessary to support the PREMIS conformance requirement(4.1).

The PREMIS Data Dictionary defines preservation metadata as "the information a repository uses to support the digital preservation process". The SDL PREMIS implementation was based on the underlying belief that, the trustworthiness of a repository system relies on the ability of tracking back information about the custodianship of the objects. The custodianship's history allows to trace responsibility's chains back to the agents responsible for the events that occurred in the digital history of the resources.

4.1 PREMIS Conformance Requirement

The PREMIS conformance declared by a repository system, means that the implementation of PREMIS information adheres to the principles of use, stated by the PREMIS conformance: use of semantic units, and use of the data dictionary.

The metadata elements held by Sapienza black repository management system, shares names and definition, and respects the use requirements of the PREMIS semantic units. The obligation, repeatability and application rules are respected (semantic units principle of use).

Moreover, all mandatory semantic units, related to the PREMIS entities, are supported and used by the repository system (data dictionary principle of use).

Consequently, the PREMIS internal conformance is respected and, at this stage of the implementation, the external conformance is supported just in the form of export. Further implementation will be allowing the import form.

The actual resources' metadata, archived in the SDL archives, are encoded and collected into the METS[6] container. The encoding of PREMIS semantic units is under development but the conformance level internal and external was already reached. All the information, deemed essential for supporting the trustworthiness of the system and the authencity of resources, are owned and managed by the Sapienza black repository

[6] www.loc.gov/standards/mets

management system, and the undergoing implementation will be encapsulating PREMIS semantic units into the existing metadata framework.

4.2 PREMIS Enrichment Workflow

The PREMIS enrichment workflow consists of those activities necessary to the information adjustement for extending the actual PREMIS internal conformance of the SDL system, toward the PREMIS external conformance. The workflow essentially shapes the SDL existing metadata in the form of export for cross-repository interactions.

The workflow essentially consists of the following activities, that gathers information, and enriches the base of data with the information needed:

- Detecting Intellectual Entities and assignment of the SDL identifier, created by the pertaining collection's identifier and a unique identifiers for the corresponding resource:
  - Identifiers coming from the originating records (bibliographic catalog or original database, spreadsheet…)
  - SDL record identifier assigned by the SDL resources acquisition function.
- Getting the information about Objects related to the Intellectual Entities. At this moment the information automatically gathered and provided by the system are more than that required by the PREMIS conformance:
  - a unique identifier for the object (type and value),
  - fixity information message digest, algorithm and the application used,
  - size,
  - format,
  - original name of the object,
  - information about its creation,
  - where and on what medium is stored,
  - relationships with other objects and other entities (via identifiers).
- Getting the information about the Events occurred in the lifecycle of the Objects until the SIP production:
  - a unique identifier for the event (type and value),
  - type of event (creation, replication, message digest calculation, validation),
  - date and time,
  - detailed description of the event,
  - a coded outcome of the event,
  - detailed description of the outcome,
  - agents (via identifiers), involved in the event and their roles,
  - objects (via identifiers), involved in the event and their roles.
- Getting the information about Agents, engaged in activities impacting on the Objects' digital history
  - a unique identifier for the agent (type and value),
  - agent's name,

- designation of the type of agent (person, organization, software),
- extended description of the agents connected to the Sapienza's organization context,
- events (via identifiers) that the agents has determined,
- rights statements (via identifiers), to which the agent is related.

- Getting the information about Rights statements that impact on the Objects management:
  - a unique identifier for the rights statement (type and value),
  - basis of right (copyright, license, statute, or other),
  - more detailed information about the rights statements,
  - actions allowed by the rights statement,
  - restrictions on the action(s),
  - term of grant, or time period in which the statement applies,
  - objects (via identifiers), to which the statement applies,
  - agents (via identifiers), involved in the rights statement and their roles.

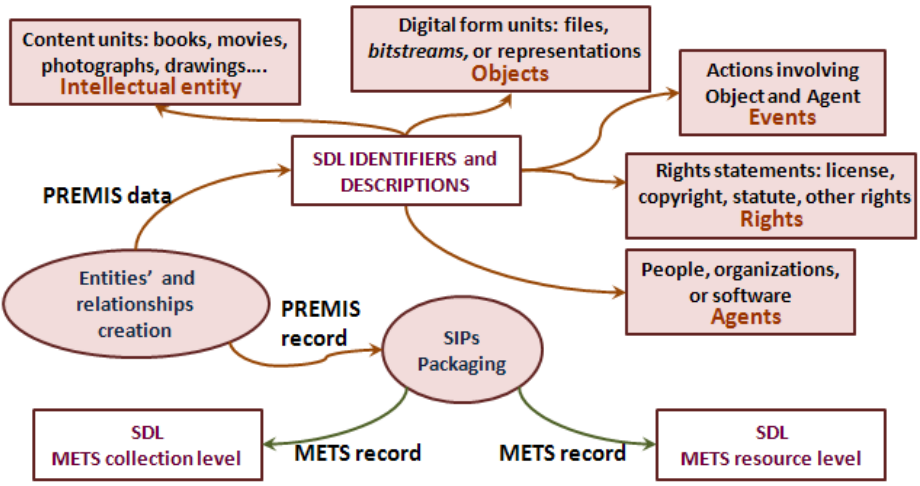

Fig. 3. Data flow diagram of the PREMIS entities implementation in the SDL metadata framework

In other words, if we express the metadata set information in natural language, it should result as: the Intellectual Entities, manifested[4] by different kinds of digital Objects, are produced by SDL Organizations made of people using tools. People, Organizations and tools are considered Agents responsible for specific actions. Actions are considered as Events in digital curation workflow, performed under specific conditions, formally defined and linked to the relevant Rights statements.

5 The Development of the Repository's Trustworthiness Value

Does the preservation metadata implementation support the trustworthiness of a system?

The maintenance of information conveying the digital history of the digital objects by means of preservation metadata related to OAIS[1], does support the future implementation of a trustworthy repository. The SDL SIP provision of a comprehensive set of preservation metadata, based on an international consensus-based standard like PREMIS, assure the availability of essential data for creating evidence of the trustworthiness of the archival systems. The AIP management, which will depends on the AIP data derived from the SDL SIP, will have a consistent set of information, available for implementing preservation services.

Any process of assessment, audit or certification strongly relies on the availability of consistent structured metadata. "Constant monitoring, planning, and maintenance of the repository, as well as conscious actions and strategy implementation will be required of repositories to carry out their mission of digital preservation" [7]. The cited management functions are strongly based on digital objects' metadata, that could negatively impact on the digital objects management, in case of absence, inconsistency, incompleteness.

Moreover, if the preservation metadata are encoded in PREMIS standard that is consensus-based of an international community of experts, the evidence of the repository's trustworthiness can be conveyed, by means of the consistent base of global standardized semantics, and on the hopeful alignment of the PREMIS conformance to the OAIS conformance.

In conclusion, the more the metadata framework will be updated, maintained and connected to the parallel "real" business evolution of the responsible Organization, showing evidence of the custody chain, the more it will be possible to have the result of the trustworthiness expected by the preservation digital repository.

References

1. Consultative Committee for Space Data Systems Reference Model for an Open Archival Information System (OAIS), Recommended Practice CCSDS 650.0-M-2 Magenta Book (June 2012), http://public.ccsds.org/publications/archive/652x0m1.pdf
2. PREservation Metadata Implementation Strategies (PREMIS), http://www.loc.gov/standards/premis/
3. Di Iorio, A., Schaerf, M., Bertazzo, M.: Establishing a Digital Library in Wide-Ranging University's Context. In: Agosti, M., Esposito, F., Ferilli, S., Ferro, N. (eds.) IRCDL 2012. CCIS, vol. 354, pp. 172–183. Springer, Heidelberg (2013), http://link.springer.com/content/pdf/10.1007%2F978-3-642-35834-0_18

4. McDonald, R.H., Walters, T.O.: Restoring Trust Relationships within the Framework of Collaborative Digital Preservation Federations. Journal of Digital Information (2010), http://journals.tdl.org/jodi/index.php/jodi/article/view/757/645
5. PREMIS Editorial Committee, Conformant Implementation of the PREMIS Data Dictionary (October 2010), http://www.loc.gov/standards/premis/premis-conformance-oct2010.pdf
6. PREMIS Editorial Committee: PREMIS Data Dictionary for Preservation Metadata version 2.2 (July 2012), http://www.loc.gov/standards/premis/v2/premis-2-2.pdf
7. Audit and Certification of Trustworthy Digital Repositories. Magenta Book. Issue 1 (September 2011), http://public.ccsds.org/publications/archive/652x0m1.pdf