Identification Semantics for an Organization, establishing a Digital Library System

A. Di Iorio, M. Schaerf

DIAG - Department of Computer, Control, and Management Engineering Antonio Ruberti - Sapienza University of Rome, Italy

Abstract. The Sapienza Digital Library collects digital resources from the different University's Organizations, representing the multidisciplinary Sapienza University's community. The underlay of the metadata infrastructure was built on digital library standard metadata semantics and was used for exchanging package, between the archival systems that manages different services for the established digital library. The semantics adopted for the metadata infrastructure can be exploited, not only for the actual digital library services, but also for connecting the resources to the Linked Open Data Cloud through authoritative identifiers.

Keywords: Digital Libraries, Metadata Semantics, Organization metadata

1 Introduction

The paper describes a specific aspect of the development of the Digital Library System of the Sapienza university (Sapienza Digital Library `http://sdl.uniroma1.it`). The approach adopted collects information, regarding the Organizations involved in the management of the digital resources' life-cycle.
In order to manage the complexity of the Sapienza University's organizational framework, a workflow for building digital resources, based on the Organizational semantics, was designed at the first stage of the project's development.
The creation and the maintenance of an identification system, based on semantics used at national level, and mapped onto other identification systems, internationally used, was necessary, in order to make feasible the retrieval of relevant information in the Linked Open Data Cloud[1] through an authoritative identifier.
The system had been resulted essential, in the entire life-cycle of the project's development, in order to refer unambiguously to the digital resources among the project's participants. In addition it was supportive for testing and improving of the overall system's information infrastructure, for refining the metadata structures, and for curating the data.
The semantics of the SDL metadata infrastructure were used for building self-documenting packages containing metadata and objects, and for exchanging packages between different digital repository systems. The digital repositories,

[1] Linked Open Data, `http://linkeddata.org`

sharing the SDL semantics, uses the exchanging package for replicating digital resources and for distributing digital library services.

2 Background

The Open Archival Information System (OAIS) [4] defines the OAIS itself as "An Archive, consisting of an organization, [...]of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community." In addition, "The Archive is responsible for creating and preserving Provenance Information from the point of Ingest; however, earlier Provenance Information should be provided by the Producer. Provenance Information adds to the evidence to support Authenticity."
The DL.org [1] booklet remarks that "The Organization Domain stems from the Organization core concept and it is conceived to represent the main settings for characterizing the DL service..."
In our project the Organization Domain is identified by the establishing organization which indeed is, the Sapienza University. The digital resources' management of the Sapienza Digital Library(SDL)[5] was founded on the cited reference models, in particular considering the relationship between the provenance information and the Organization responsible for the management of digital resources. The production process of the OAIS Information Package (IP), used in the different functional scenario (Submission, Archiving, and Dissemination), was designed following the strategy of capturing relevant information about its custody, and exploiting the identification information associated to the Sapienza's Organizations. The self-documenting digital resource produced, can be used by other application systems sharing the standard metadata semantics, used by the SDL metadata infrastructure.

3 The long term scope of the system architecture

The system architecture was conceived with the scope of the Long Term Digital Preservation(LTDP) of materials for the multidisciplinary community of Sapienza.
The replication of the produced OAIS IPs in different repositories geographically separated, and the heterogeneity of the supporting technologies and methodologies[9][2], were considered influencing requirements, in the design of the overall architectural system.
As consequence, the initial scope of building a digital library was extended, and turned toward the conception of an infrastructure for digital library, and digital preservation services.
Following this conception, the metadata infrastructure had to be agnostic about the technological platform, in order to re-use information and objects in different digital systems, as well as in different semantic contexts. Nevertheless much

of the semantics, used for the values of the metadata elements, are often under the competence of the managing Organization. The semantics used if not well-documented and structured can be an obstacle, for the automatic management of data and documents, and consequently can have have a strong impact on the long term management of the digital resources.
Under this belief, the work-flow for building digital resources was conceived for absorbing information, conveying the custody chain of the management activities performed by different Organizations.
In other words the overall management of a digital resource, during its creation process, is permeated by the Organization's context information, connecting the digital resource to its "real" Organizations involved in the management of its production's .
An abstract representation of the main components of the overall architecture of the system is showed in the Figure 1. The main components can be divided in three categories: the pre-ingestion systems preparing the digital resources, the Digital Library Management System (DLMS)[1], performing the OAIS functional services[4], and the dissemination system. The system's components performing specific function in the architecture are briefly described in the following list.

- The Massive conversion system performs the retrospective conversion of existing digital materials, and related content's description, standardized or not standardized: it was developed for the need of Sapienza, extending a PHP/Mysql application, Bringing Digital Environment (BriDgE)[2].
- The Cataloguing system properly developed for describing collections of heterogeneous materials to be digitized.
- The DLMS as defined by the Delos Reference Model[3]: was developed extending services of Fedora Commons[4].
- The web portal of SDL, which manages the public interface of the system.

The Cataloging system and the web portal had been developed using Drupal[5] that uses services managed by the DLMS. The Italian University consortium Cineca[6], as technological partner of Sapienza for SDL, has developed the DLMS and the Cataloguing and the web portal systems.
Actually the repository, archiving the digital resources managed by the DLMS, is located in Bologna (the location of the Cineca's headquarter).
The exchange of IPs between pre-ingestion systems (Massive conversion and Cataloguing) and the DLMS, is performed between Sapienza repositories in Rome, and the DLMS's repository located in Bologna. This preservation strategy respects the influencing requirements of the LTDP: the digital resources replication in different repositories geographically separated, and the heterogeneity of the supporting technologies and methodologies[9][2].

[2] Bringing Digital Environment (BriDgE), `http://bri-dge.sourceforge.net/`
[3] DELOS Reference Model for Digital Libraries, `www.delos.info/ReferenceModel`
[4] Fedora Commons, `http://fedora-commons.org/`
[5] Drupal, `http://www.drupal.org/`
[6] Cineca website, `http://www.cineca.it`

The OAIS IPs produced by the pre-ingestion systems are the exchanging packages used by the systems supporting the different services. Consequently the IPs produced by the pre-ingestion systems has to be self-documenting, on the base of metadata and identification semantics shared by the SDL systems, geographically separated.

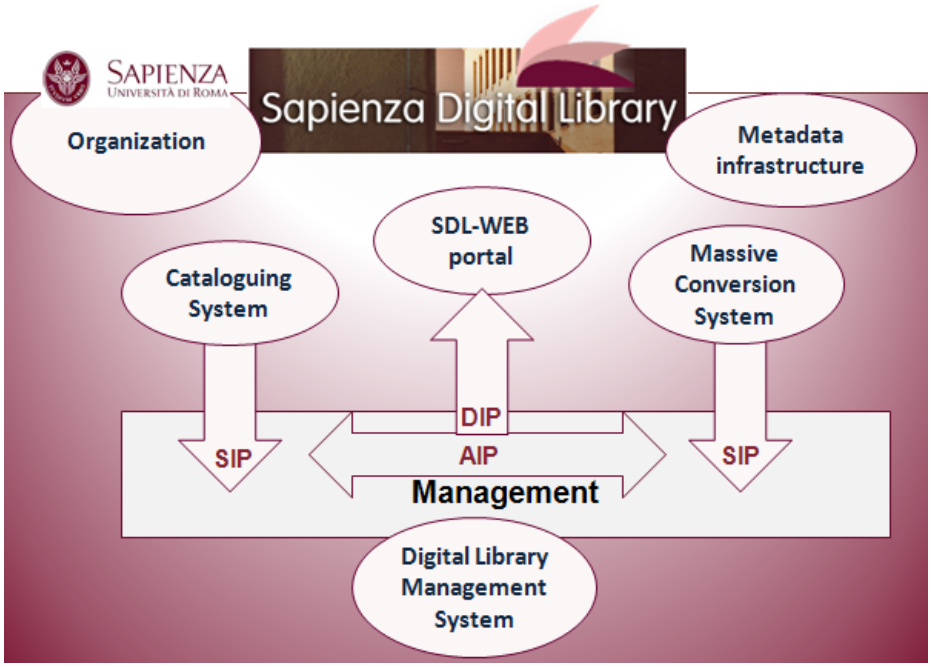

Fig. 1. Abstract overview of the SDL architecture components

4 Approaching the organizational complexity

The Sapienza University's is a complex Organization composed by 63 investigation departments, 56 libraries, 21 museums, 8 administration departments and some research center. We have conceptually considered the Sapienza's Organizations as Organizational units belonging to the Sapienza University.
In order to deal with the Organizational complexity of the Sapienza University, it was deemed essential to devise a metadata infrastructure, not only based on semantics world-wide known, but also with identification semantics aiming to identify unambiguously the Sapienza's Organizational units, involved in the work-flow production of the digital resources.

Furthermore, the long-term focus implies that the metadata infrastructure is able to record information referring to the real evolution of the Organizational units, that are involved in the management of the digital life-cyle of resources. The conception of an holistic approach referring to the Organizations' custody chain, recorded and expressed by the metadata infrastructure was based on the two reference model cited in the section 2 [4][1]. In addition the "Certification (TRAC): Criteria and Checklist"[3] that now is an ISO standard[7], focused on the repository's trustworthiness certification, proves that the first aspect in the checklist, influencing the trustworthiness of the digital repository, is the Organizational infrastructure. Consequently, the information about Organization, establishing an information system, has not to be neglected, but has to be curated and considered as relevant OAIS Preservation Description Information.
Considering the reference models, the long term aspect, and the complex organizational application context of Sapienza, the following requirements for designing the metadata infrastructure and the supporting identification semantics, were deemed essential:

- the unambiguous identification of the Sapienza's Organizations producing digital resources;
- the maintenance of the naming information history, connecting the evolution of the real Organizations with the digital management of the resources;
- the establishment of an identification hierarchy based on the concept of the Organizational Collection.

5 The Digital Library system and the metadata infrastructure

The digital resources managed by the SDL system constitute the digital representation of the Intellectual Entities[10], that are managed under different types of conditions (creation, holding, management etc.), by the Sapienza's Organizational units. The definition of Intellectual Entity, is borrowed from the PREMIS Data Model[10], which defines the intellectual entities as: "a set of content that is considered a single intellectual unit for purposes of management and description: for example, a particular book, map, photograph, or database. [..] An Intellectual Entity may have one or more digital representations." In the SDL system an Intellectual Entity is technically represented by a Digital Resource (DR), that can be considered as the digital embodiment of an intellectual item, and is equivalent to the OAIS IP [4].
By the implementation point of view a DR is physically composed by the set of objects files, that together represent the OAIS Content Information, and the set of metadata represent the OAIS Preservation Description Information.

5.1 The digital library standards adopted

The metadata infrastructure was conceived for supporting different kind of DRs. The DRs can be represented in different formats (still and moving images, texts,

sounds, cartographics, etc) and can be representing different kind of intellectual contents (multidisciplinary knowledge). In order to manage the materials' diversity and to deliver centralized digital library services, based on the metadata, we had considered metadata standards with a sufficient degree of granularity, as well as a sufficient level of semantic interoperability. The analysis of the standards adopted in the digital libraries' scenario had driven to the choice of a very well known standards combination:

- Metadata Objects Description Standard(MODS) which describes the intellectual contents and follows libraries semantics, derived by the MARC 21 semantics[7], the pillar standard of all libraries information systems;
- PREservation Metadata Implementation Strategies(PREMIS) for managing preservation metadata;
- Metadata Encoding and Transmission Standard(METS)[8] for wrapping together metadata belonging to the DR.

Mostly these standards, made up in combination, have covered the need of providing sufficient granularity of information for the intellectual content (MODS), sufficient granularity of information for the digital preservation management (PREMIS), and sufficient granularity and flexibility for supporting the need of managing an Organization infrastructure, using DRs variously structured (METS). Indeed, the encapsulating mechanism provided with METS has allowed not only to include other standard semantics, more relevant to specific aims (like for example Dublin Core (DC)[9] (more interoperable), or NISO Technical Metadata for Digital Still Images Standard MIX[10]), but also supporting the exchange of packages between the architectural components of the SDL infrastructure (see Sect.3).

5.2 Metadata infrastructure and the building blocks

The metadata infrastructure is coded in the adopted standard semantics and is organized on the DRs, that are the essential bricks, building the digital library. Both the massive conversion system, and the cataloguing system produce DRs, encoded in XML[11], and conforming to the metadata standards adopted by the project (see the following Section).
The DLMS ingests DRs produced by both two pre-ingestion systems, in order to start the management of their digital life-cycle[4].
The Figure 2 is a simplified representation of the SDL's DR. On the left is visible how the conceptual OAIS IP is generally divided into two parts: the metadata, and the content objects. On the right is represented how is physically composed

[7] MARC 21 Format for Bibliographic Data, `www.loc.gov/marc/bibliographic/`
[8] Metadata Encoding Transmission Standard, `www.loc.gov/standards/mets/`
[9] The Dublin Core Metadata Initiative, `dublincore.org/`
[10] NISO Technical Metadata for Digital Still Images Standard, `www.loc.gov/standards/mix/`
[11] Extensible Markup Language (XML), `http://www.w3.org/XML/`

inside of the system, as a set of different metadata semantics and a set of object files. Each box is labeled with the name of the related standard XML schema[12] name(see Sect. 5.1).

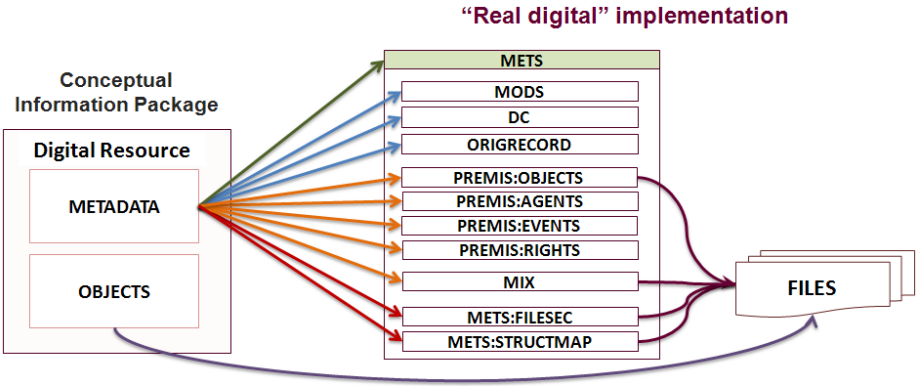

Fig. 2. Digital resource's structure

The descriptive metadata, pointed by the blue arrows, is coded into two descriptive standards. The MODS which reflects the granularity of MARC 21. The DC, which is commonly adopted in other contexts, not strictly related to the libraries world, is consequently considered more interoperable.
The inventory metadata, listing the files' names and locations, and the structural metadata, pointed by the red arrow, are coded in the two relevant METS sections. Both sections of metadata are connected together by METS, which is essentially used for conveying the whole structure of the DR in the XML format. All metadata blocks are unambiguously identified and referred to the Organizational context[8], related to the DRs production.
The system was publicly opened the 20th of December 2013, as Beta version 1.0[13], and is under testing by the communities.
The DLMS is actually providing access, and discovery services to the communities and has ingested more than 11.000 DRs distributed in 22 collections, belonging to 10 different Organizational units. By the end of the year more than 30 new Sapienza's Library will be incrementing the number of digital resources.

6 The identification semantics for Digital Resources managed by an Organization

The SDL DR's production starts from the main constraint of existence about the identifier of one of the Sapienza's Organizational units, which has the man-

[12] XML schema, `www.w3.org/XML/Schema`
[13] Sapienza Digital Library, `sdl.uniroma1.it`

agement responsibility of the DRs. Consequently, the *conditio sine qua non* for the existence of every single DR must be its identification by an identifier based on the Sapienza's Organizational units' identifier. This identifier abstractedly defines the concept of "Organizational collection", that gathers all DRs belonging, owned or managed by a Sapienza's Organizational unit. Consequently, all objects belonging to a DR are identified extending the Organizational collection identifier, which is the root of the identification.
The identifier is necessary for capturing information about the Organization context, which has some responsibility in the SDL DR production: scientific or technical responsibility, objects digitization, metadata editing or management responsibility. The long term focus of the digital library requires to deal with an ever-growing amount of DRs and the re-use in the long term of a DR could result difficult or inconsistent, if it is not possible to have agents of reference about its management.
The semantics adopted for the whole process of SDL's DRs production is based on an identification system that, first of all, aims to identify the Sapienza University ownership of the digital library service. In addition it identifies the Sapienza's Organizational unit, having the initial management responsibility of the resource's digital born under the Sapienza domain (selection or creation of the digital materials). The production method designed for building DRs allows to produce self-documenting IPs, where the documentation is based on the structured semantics, referring to the Organizational context.

6.1 The Organizational collection and the identification family

The Organizational collection in the conception of the SDL is the digital embodiment of the Organization's collecting actions, that consist of the digital production, preservation and fruition. The collected digital item is represented physically by the OAIS IP which is the DR in the SDL context.
The Organizational collection is the set of the whole digital production made, managed or owned by the Sapienza's Organizational unit that has the responsibility of the DRs created for the SDL. The abstract concept of the Organizational collection refers the contextual information about the Organization and set the basement of the identification semantics of the referred DRs.
By means of the Organizational collection identifier, we captured information about the organizational context where the DR was born, and produced for the ingestion in the SDL's DLMS. We have also leveraged on the identification information for relating other information, about context and provenance[4][6] related to the DRs.
This is the reason why the related Organizational collection's identifier is considered the first mandatory information, for submitting the resources to the system.
In order to respect the LTDP requirement, allowing the DRs re-use, we have considered essential to use identification semantics, already used by a national identification system, where the main organization Sapienza and its Organizational units are hierarchically represented.

Respecting the hierarchical structure of the University, the SDL identification system has adopted an identifiers' family derived and extended from the Italian National Bibliographic System[14] where the Sapienza University is identified by the identifier "RMS".
This is the main identifier, which associated with descendant identifiers, unambiguously identify the Organizational collection, and build relationships with other entities involved in the DRs management: objects, agents, events and rights[10].
The well-defined structure of the SDL identification system has allowed to enrich resources and the pertaining objects with contextual information about Sapienza organization.
In addition, the registration of the Sapienza University to the international identification MARC organization code [15], identified by "itrousr", and semantically mapped to the same level of the italian "RMS" identifier, allows to set the DRs context also at international level. Indeed, the replication of such code as mandatory administrative metadata in each SDL's DR, makes possible its connection to the Linked Open Data Cloud[16].
The open world "itrousr" identifier, exposed by the Library of Congress Linked Data Service(LCLOD)[17] in the Cultural Heritage Organization identification system as authoritative identifier, makes each DR, belonging to the local Sapienza domain, worldwide reachable through the exposed identifier "`http://id.loc.gov/vocabulary/organizations/itrousr`", and by virtue of the mapping between the local ("RMS") and global identifier("itrousr").

6.2 The Organization as the source of the identification system

The SDL identification system is structured on four layers, extended from the main layer, represented by the "RMS" identifier of the Sapienza Digital Library, and going down to the following hierarchical layers, that are also sampled in the Fig.3:

- the root identifier corresponding to the Organizational Collection (see subsect. 6.1), in the showed case, "RMSAR" identifies the Sapienza's Library of Architecture;
- the digital collection identifier, corresponding to the SDL aggregation level for managing DRs, which in many cases is directly identified by the Organizational collection itself. In the showed case, the Library of Architecture collects the digitized books from its holdings, directly collected as DRs of the Organizational collection "RMSAR". In addition the same library collects a special collection "RMSAR_SEVERATI" collecting images, donated by an Architecture's Faculty member;

[14] Anagrafe Biblioteche Italiane `http://anagrafe.iccu.sbn.it/opencms/opencms/`
[15] MARC code list for organizations `http://www.loc.gov/marc/organizations/org-search.php`
[16] Linked Open Data, `http://linkeddata.org`
[17] Library of Congress Linked Data Service, `id.loc.gov`

– the DR (Figure 1) identifier, in the figure the "RMSAR_00000025" is a digitized book of architecture, and "RMSAR_SEVERATI_00000001" is a photograph digitized and containing Brasilian buildings relevant for the architecture interest;
– the digital objects identifier, represented by the DR's identifier and the order number of the object, as example the book's page "RMSAR_00000025_0324".

The replication of the higher layer's identifiers over the identifiers of the lower layers, allows to reuse the single objects in other contexts, without ambiguity about the pertaining DR of the objects, and from the root identification layer, back to the responsible Organization. The multiple representing format (in the example jpg and tif) are managed by the system, using the reference of the digital object's identifier.

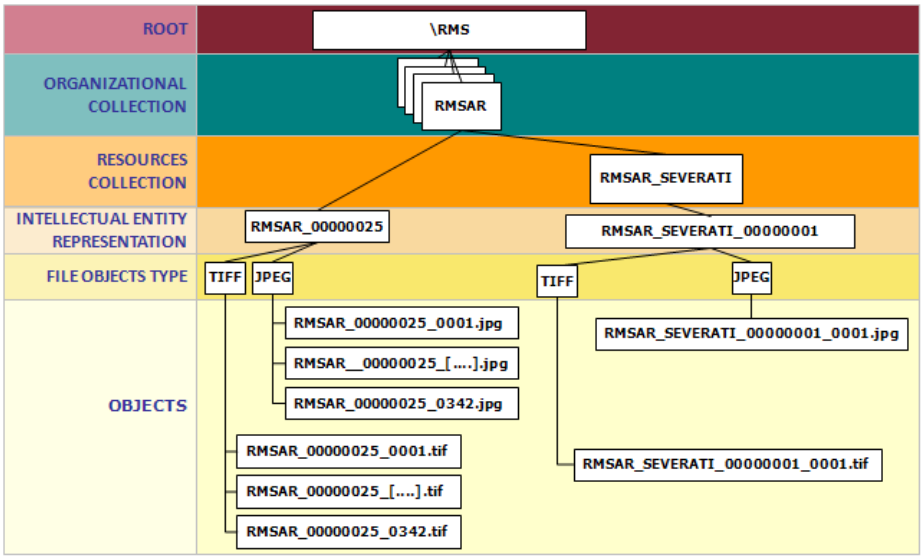

Fig. 3. SDL resource identification layers

7 Conclusions and future developments

The management of the identifiers based on semantics, derived by the Organizational collection conception, matches with two Ontologies, recommended recently by the W3C.
The Organization Ontology(Org-O), originally developed for use by data.gov.uk[18], represents the formal definition resulted from real implementations and uses. The

[18] Opening up government `http://data.gov.uk/`

core class in the ontology is the "Organization" class which represents "a collection of people organized together into a community or other social, commercial or political structure". The main class "Organization" of the ontology Org-O, semantically speaking, matches to the Sapienza University. While the Org-O subclass "OrganizationalUnit", matches with the Sapienza's Organizational units.
The matching conceptualization between the "OrganizationalUnits" class of the Org-O and Sapienza's Organizational units associated to the SDL's Organizational Collections, and unambigously identified, will drive the reasoning systems to retrieve information about DRs belonging to the pertaining "Organization" or "OrganizationalUnit".
The identification system based on semantics locally defined, but world wide processable by means of dereferenceable URI like the LCLOC identifier (see Sect.6.1), allows to make all belonging DRs reachable by URI through the Organization ontology support.
Coherent to this scenario is the ontology aimed to model the information about "entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness"[19], known as Provenance Ontology.
The Prov-O Ontology(Prov-O) [11] is provided with a Data model, that simply defines three core types of classes: Agent, Entity, and Activity and related relationships. Focusing on the topic of this paper we underline the fact that the Agent defined as main class in the PROV-O data model, can be connected to the Org-O's Organization concept by means of the Agent subclass "Organization". The Organization subclass is defined in the Prov-O as "An organization is a social or legal institution such as a company, society, etc.". Also in this case the matching of PROV-O definition with the SDL Organizational units, and its Organizational collection digital conceptualization, allows to connects classes of information and relationships with the information collected in SDL, where the identification semantics drive to the relevant values.
The recommendation by W3C of this two ontologies demonstrates the global interest, around the traceability of digital assets back to the Agents responsible for their management, harmonically with the SDL's Organizational collection conception, where the agents belong to the context information referred to the Organization.

References

1. Candela, L., Athanasopoulos, G., Castelli, D., Al., e.: The Digital Library Reference Model. Tech. rep., DL.org: Coordination Action on Digital Library Interoperability, Best Practices and Modelling Foundations (2011), http://bscw.research-infrastructures.eu/pub/bscw.cgi/d222816/D3.2bDigitalLibraryReferenceModel.pdf
2. Caplan, P., Kehoe, W., Pawletko, J.: Towards interoperable preservation repositories (tipr). In: Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop. p. 16. ACM (2010)

[19] PROV-Overview, http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

3. CLR (Center for Research Libraries and RLG Programs): Trustworthy repositories audit and certification checklist (2007), http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf
4. Consultative Committee for Space Data: Reference Model for an Open Archival Information System (OAIS), Recommended Practice CCSDS 650.0-M-2 Magenta Book (2012), http://public.ccsds.org/publications/archive/652x0m1.pdf
5. Di Iorio, A., Schaerf, M., Bertazzo, M.: Establishing a digital library in wide-ranging university's context: The Sapienza Digital Library experience. In: Digital Libraries and Archives, 8th Italian Research Conference on Digital Libraries, IRCDL 2012, vol. 354 CCIS, pp. 172–183. Springer (2013), http://www.scopus.com/inward/record.url?eid=2-s2.0-84873865280\&partnerID=40\&md5=d8b5b1f12a673c347ec521d4a4e8b391
6. Di Iorio, A., Schaerf, M., Guercio, M., Ortolani, S., Bertazzo, M.: A digital infrastructure for trustworthiness The Sapienza Digital Library experience. In: Bridging Between Cultural Heritage Institutions, pp. 59–69. Springer (2014)
7. ISO (the International Organization for Standardization): Iso 16363:2012 - space data and information transfer systems – audit and certification of trustworthy digital repositories (2012), https://www.iso.org/obp/ui/#iso:std:iso:16363:en
8. Parsons, M.A., Godøy, Ø., LeDrew, E., De Bruin, T.F., Danis, B., Tomlinson, S., Carlson, D.: A conceptual framework for managing very diverse data for complex, interdisciplinary science. Journal of Information Science 37(6), 555–569 (2011)
9. Payette, S.: The state of technology for digital archiving. arXiv.org http://arxiv.org/pdf/1403.7748v1.pdf
10. PREMIS Editorial Committee: PREMIS Data Dictionary for Preservation Metadata version 2.2. (2012), www.loc.gov/standards/premis/v2/premis-2-2.pdf
11. W3C: Organizational Ontology (2013), http://www.w3.org/TR/vocab-org/