Establishing a Digital Library in Wide-Ranging University's Context

The Sapienza Digital Library Experience

Angela Di Iorio[1], Marco Schaerf[1], and Matteo Bertazzo[2]

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

Abstract. The Sapienza Digital Library (SDL) is a research project undertaken by Sapienza Università di Roma, the largest Europe's campus, and the Italian supercomputer center Cineca.

The SDL project aims to build an infrastructure supporting preservation, management and dissemination of the past, present and future digital resources, that contain the overall intellectual production of the Sapienza University. The solution adopted tries to find a tradeoff between the standardization of the digital processes and products (that allows a cost-effective centralized and shared management and curation), and the preservation of the peculiarities of scientific materials, belonging to disparate knowledge disciplines (that need to be digitally available for future initiatives, more specifically tailored to the designated communities).

Keywords: Digital library, Long term digital preservation, Digital curation, OAIS, METS, MODS, PREMIS, Controlled vocabularies.

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, which is a non profit consortium made up of 47 Italian universities.

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.

Setting the future scenario of the SDL application, it has been evaluated and prefigured the large amount of research and knowledge materials, coming from such large and ancient University, as well as the variety of interests coming from such large and multidisciplinary community of stakeholders, and, last but not least, the potential uses of that material, for general and specialized communities of users.

[1] `http://sapienzadigitallibrary.uniroma1.it` (expected on January 2013).

M. Agosti et al. (Eds.): IRCDL 2012, CCIS 354, pp. 172–183, 2013.

The project was indeed conceived to manage the integration of a large volume of multiformat materials, and to enable their access through different devices, in order to fulfill the needs and the expectations of diverse communities, local, global, and future.

The actual state of experiences in digital libraries, in digital resources management, in digitization and in evolution of dissemination tools have suggested to examine new cost-effective solutions in the weaving factory of submission, archiving, and dissemination of digital resources. The solution adopted tries to find a tradeoff between the standardization of the digital processes and products, that allows a cost-effective centralized management, and the preservation of the peculiarities of scientific materials that need to be digitally available for future and specific initiatives.

2 Mission

The primary objective of the project is to provide Sapienza University with a modern digital library, comprehensive and open, which contains all digital materials produced by, held by, with ownership of, or granted to Sapienza.

The materials will be organized, catalogued, enriched and made accessible to the whole academic community and over.

3 Project's Objectives

The initial objectives that was detected are those essentially appliable to any kind of university or educational institution, and extending the vision, also to any institution that needs to manage digital material.

The objectives were firstly defined from the users point of view:

- Offering to the Sapienza's designated communities the opportunity to exploit digital materials owned and/or produced by Sapienza;
- Managing a broad variety of digital materials, born-digital and digitized;
- Archiving and preserving collections of images, audio/videos, 3D materials, scientific articles and datasets, special and valuable collections (private archives of scientists, work archives, etc.), museums/archives/libraries materials, scientific learning and teaching materials;
- Organizing, grouping, and indexing materials, supporting their browsing and searching on different dimensional views, and their reuse in different contexts;
- Optimizing, improving, and enhancing the value of digital materials throughout the web semantic technologies and social tools;
- Building a framework in the forefront, for the submission, dissemination, and preservation of Sapienza's digital assets, interconnected with the most important Italian, European, and International digital resources aggregators;
- Allowing the interoperable conversation with other kinds of information management systems (libraries/archives/musems/universities, open access repositories…).

More technical objectives connected to the more stringent organizational and technical requirements in harmonization with the global digital libraries, and the digital curation scenario, were defined as the following:

- Managing digital materials coming from former digitization projects, making retrospective conversions of existing materials;
- Gathering as much as possible information, lowering the threshold of information lost;
- Achieving a satisfying level of information not only for user needs, but also for enabling advanced services for preservation and dissemination;
- Enriching the information making it reusable and connectable with other application contexts,
- Adhering to the Open Archival Information System (OAIS)[1] functional model and developing compliant services supporting the Long Term Digital Preservation (LTDP);
- Adopting the most spread digital libraries and digital preservation metadata standards, in order to mantain and to guarantee the interoperability of the SDL system with other systems, supporting the worldwide dissemination of digital resources;
- Adopting platforms and tools based on open source solutions.

4 Application Context

Sapienza university was founded in 1303 by Pope Boniface VIII and nowadays has 145,000 students, over 4,500 professors, almost 5,000 administrative and technical employees. The Sapienza organizational units for learning and scientific investigations, cover almost all disciplines of knowledge, and are divided into 11 colleges and 68 departments. The Sapienza memory organizations are represented by 59 libraries, 20 musems and 2 main archives, current and historical.

Collecting and managing the intellectual materials that was, is, and will be produced by that large organizational scenario, needs of a common, and a cost-effective solution, which leveraging on standardized digital resources, will allow their management and exploitation in the long term.

By the digital management point of view, the application context is extremely fragmented, because of the multiplicity of information sources that had produced digital resources in a local view and with personalized methodologies. Actually, scientists are simply more focused on their studies, researches and interests, than on the digital management of their information resources. For this reason, finding a common way to organize, manage, and exploit the content of digital materials, is essential to provide useful tools that support and ease the intellectual work, and lower the weight of the daunting task of managing digital resources.

5 Digital Materials Used

At the beginning of the project was necessary to prepare an initial census of the existing digital materials that were representative of specific type of objects like for example videos, images, digitized books, text documents, big images like historical maps. All the materials were gathered and stored into a dark repository of the Sapienza computing center. The materials were used as samples for the workflow of digital resources building that has led to the creation of the Submission Information Package (SIP) as required by the OAIS model[1].

Different types of SIPs were modelled on resources' types (i.e. image, video, map) and system's services were coherently modelled for ingesting, managing, and accessing the content. For example, even though maps and a photographs are both images, the fruition service provided by the system is different in regard to the image's dimension.

In general, the materials that, the SDL will be able to manage, are:

- Books, ancient (before 1831) and modern, prints, maps and other digitized materials;
- Scientific digital products (Ph.D. thesis, materials with rights, datasets…);
- Images, Audio/Video materials digitized and born-digital;
- Learning objects
- User Generated Contents
- Special materials: i.e. archaeological documentations, personal archives…
- 3D objects

6 SDL Reference Models and the Preservation Strategy

The reference models that lay down the digital library design and conception are the Open Archival Information System (OAIS)[1] and the DELOS digital library Reference Model (DELOS)[2].

Specifically, the DELOS Three-tier framework composed by Digital Library(DL), Digital Library System(DLS), and the Digital Library Management System(DLMS) were envisaged in the Sapienza context.

In conformance with DELOS model, the Sapienza DL is a set of "real" persons and organizational units that "*collects, manages and preserves for the long term rich* ***digital content,*** *and offers to its* ***user*** *communities specialised* ***functionality*** *on that content, of measurable* ***quality*** *and according to codified* ***policies".***

The Sapienza DLS is a distributed architecture tailored on and used by the different communities and provides specific software tools.

The Sapienza DLMS is the software infrastructure which was conceived following the philosophy of the "extensibility", in order to implement tailored services in harmonization of the integration of new content types (i.e. specific visualization tool for big images) as well as the introduction of new requirements for the system (i.e. application of new information classification system).

A supposed DLMS is usually founded on the OAIS conceptual model, and usually its archiving repository provides the basic functions like ingestion, archival storage, data management, administration, preservation planning, access. Actually, very few DLMS are equipped with a complete and overall preservation planning, likewise coherently, the relevant administration. Indeed, the LTDP needs more integration between the technological support and the digital preservation organizational need, which is especially expressed throughout the organizational commitment, as evidence of the sufficient level of awareness about the LTDP.

7 Activities Project's Overview

The project has started in January 2011 and it was divided into two work phases.

The objective of the first phase, was to release a DLMS prototype implementing all the macrofunctionalities defined by the OAIS: ingesting, archiving and access. To release the DLMS prototype, it was necessary to design the pre-ingestion activities for the SIP building, and contemporaneously, it was defined and progressively improved the metadata framework, useful to support the information management. Consequently, the outcome of the first phase was prototyping, the SDL metadata framework (7.3), the Sapienza pre-ingestion workflow (7.4), the Sapienza SIP building (7.5), the SDL DLMS (7.6).

The first phase of the project has been closed in December 2011.

The second phase started in January 2012 and the objectives are: 1) developing DLSs for making communities to interoperate with SDL, 2) enriching the prototyped elements, released during the first phase, by adding metadata enabling the digital preservation strategies implementation, 3) optimizing the overall DLMS functionalities.

The following describes the activities performed for the first phase of the project.

7.1 System's Requirements Analysis and Matching of Digital Materials Characteristics

The census of available digital resources has resulted in a first categorization of materials types, and in a list of characteristics, that need to be managed by the system's services in supporting the main OAIS functionalities like ingest, archiving and access. The characteristics of materials were modelled taking into account the user needs, and consequently the differentiated access types, the variety of searching/browsing, the preservation needs and the draft of the rights management with the *corpus* of permissions, statements and other generic constraints.

From that initial analysis was designed a workflow of the materials processing, in order to prepare them for the submission to the SDL digital repository. The processing objective is the creation of a SIP, which in real implementations is a compound object made of content objects and metadata objects.

7.2 Selection of the Digital Repository Application System

The choice of the digital repository management system is a strategic and constraining decision for the digital curation of the materials. Consequently, an initial evaluation of the most spread *open source* software projects was done, and the Fedora Commons (FC)[3] has been chosen because, in spite of its complexity is more oriented to the web services integration, the semantic web technologies, and the LTDP. Furthermore, FC gives chances to use *content models,* that can be customized in regard to the originating models of the SDL materials.

As usual in implementing FC, an analysis of the *pros* and *cons* about the atomistic and compound paradigm was done, in relation to the projects requirements. The choice was led on the atomistic model, mainly considering the long term perspective of the project, which foresees to use and reuse the digital materials in diverse contexts. The greater flexibility, in reusing digital objects, was considered a good reward respect to the major complexity in managing the atomistic model, due to the system maintenance of information about relationships among objects.

7.3 Definition and Design of the SDL Metadata Framework

The metadata framework conceived for SDL had taken into account the wide-ranging general requirements that set three specific characteristics: completeness (gathering as much as possible information), flexibility (adapting to different contexts) and extensibility (integrating with new information).

Consequently, the framework has to be able to hold any kind of resources' description. The holding of information does not mean that the managing system has necessarily to manage it, but holding information and maintaining it available, would allow its reuse in future focused projects.

The metadata framework has to support the following requirements:

- conformant with OAIS;
- prearranged to hold different standard descriptions on which implementing integration services, supporting the use of wide-ranging knowledge's materials;
- prearranged to the exchange with other digital library systems or other information management systems;
- prearranged to the LTDP and equipped with the minimal and essential metadata, enabling the long term management.

The metadata is generally categorized in descriptive, administrative, structural rights management, preservation[2], and technical and use[3], even though same metadata can

[2] *Understanding Metadata, National Information Standards Organization,* 2004, www.niso.org/standards/resources/UnderstandingMetadata.pdf

[3] Tony Gill, Anne J. Gilliland, Maureen Whalen, and Mary S. Woodley Edited by Murtha Baca Introduction to metadata Online Edition, Version 3.0 http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html

be assigned to different categories, in regard to the use perspectives. The actual scenario, in DL implementations, highlights the broad adoption of metadata standards, coming from metadata specialists international workgroups, supported by the Library of Congress. The most adopted combination is METS[4]/MODS[5]/PREMIS[6], where PREservation Metadata Implementation Strategies (PREMIS), is used for preservation metadata, Metadata Object Description Schema (MODS) for descriptive metadata and Metadata Encoding & Transmission Standard (METS) for wrapping metadata all together. Usually the metadata standards are available to the communities by means of XML schemas[4] that enable the information systems to interchange, set of information encoded in XML files.

The SDL has adopted primarily LOC standards because of the wide-adoption in DL projects and, the more the standards are spread and the more the adopting systems are interoperable. In addition, because they are open standards it follows that the longevity of their knowledge-base is likely longer.

Descriptive Metadata Set

The SDL metadata framework was designed to support as much as possible information, even coming from different kinds of knowledge provider. The MODS metadata is the “core description” on which are configured the DLMS’s services for searching and browsing of the SDL collections. A stable MODS profile will be released during the second phase of the project, and it will be the reference descriptive framework for describing new digital materials. All the varied information sets, collected from the different Sapienza organizational units are mapped and encoded in MODS, and enriched by controlled values taken from the MODS controlled vocabularies as well as, the SDL controlled vocabularies.

The translation of different information sources into the MODS has respected and followed the Digital Library Federation/Aquifer Implementation Guidelines for Shareable MODS records[7]. The elements required by the DLF/Aquifer requirement level, has been adopted as one of the SDL policies for the basic requirement level in resource’s description. Furthermore, for special collections it has been taking into account the Master Data Element List of Library of Congress Metadata for Digital Content[8].

The MODS has been used for describing materials, not only at the single item level, but also at the collection level. Every item or resource (here meant as a discrete unit, conceptually equivalent to the OAIS Information Package (IP), in this article specifically qualified as SIP in 7.1), existing in SDL, must belong to an identifiable collection, that indeed is described by MODS elements.

The MODS was considered more suitable for the SDL metadata framework, because is richer than the easiest implementable Dublin Core, it being understood that the DL system can dumb down from MODS, to Dublin Core[5], and similar in simplicity, as well as to map toward open data standards[6].

[4] XML Schema, `http://www.w3.org/XML/Schema`

[5] The Dublin Core Metadata Initiative, `http://dublincore.org/`

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

Mostly, all advantages and features, listed on MODS official website, were considered important in implementing it, but one of them deserves to be cited: MODS can be used also for Search/Retrieval via URL(SRU)[7] as specified format, enabling federated searching and similar automated queries via URL, which means to offer further advanced services to the users.

Metadata Container and Structural Metadata

For packaging all metadata together into the defined SIP, the SDL metadata framework deviced, has exploited the flexibility of METS, making the system available 1) to collect other kinds of metadata set, over those specifically adopted, and 2) to dumbing down toward standards less complex like Dublin Core or European Semantic Elements[8].

Whenever is necessary to collect resources' descriptions more detailed than MODS, these descriptions are stored "as is" and summarized, according to the SDL MODS profile, into the SDL MODS core description.

Thanks to the METS flexibility, during the development of the per-ingestion activities and the improvement of the metadata framework design, other metadata standards have been embedded, like for example technical standards specific for different kinds of materials (MIX[9], VideoMD[10]).

A stable METS profile will be released during the second phase of the project.

Preservation Metadata

The overall SDL SIP building workflow was pervaded by the LTDP philosophy, ensuring the basic provision of the preservation metadata, considered mandatory by the PREMIS, that is the preservation metadata framework mapped from the conceptual structure of the OAIS model. The SDL metadata framework was designed to guarantee the minimum conformance with the PREMIS standard both on semantic unit and data dictionary level, following requirements and constraints, and by collecting all the metadata defined as mandatory by the PREMIS Data Dictionary[9]. Although PREMIS is not formally and completely adopted yet, all the mandatory information were encapsulated into the metadata framework, and all the other useful information were stored by the SDL black repository system and will be encapsulated during the project's second phase.

This means that the SDL resources are already equipped for the management of the LTDP strategies, even though the DL prototype is not managing them yet. The preservation planning and administration will be implemented in the second phase of the project.

[7] Search/Retrieval via URL, `http://www.loc.gov/standards/sru/`

[8] European Semantic Elements, `http://www.europeanalocal.eu/eng/Document-Library/Reports/ESE-Semantic-Elements-ver-3.1`

[9] NISO Technical Metadata for Digital Still Images Standard, `http://www.loc.gov/standards/mix/`

[10] AudioMD and VideoMD technical metadata for Audio and Video, `http://www.loc.gov/standards/amdvmd/index.html`

7.4 Pre-ingestion Workflow

Considering the extension of the University, in numerical as well as in geographical terms, the workflow was designed following a distributed view of the work, even though in order to design the workflow, at the beginning a centralized system of the materials' treatment is necessary. The workflow will be applied and distributed in the different institutional units belonging to Sapienza at the first stable system release.

The materials gathered into the SDL dark repository were of different varieties, in terms of contents and metadata structure. The first step was to maintain the provenance source of materials identifying firstly the real Sapienza Organizational Unit, that asked to submit materials to the system, and secondly, identifying the collections and items contained.

At the beginning, the workflow was designed and tested on a sample collection of almost 2000 images, and progressive tests of SIP building processes(7.5) had fixed and integrated both metadata framework(7.3), pre-ingestion workflow activities and the SIP building processes.

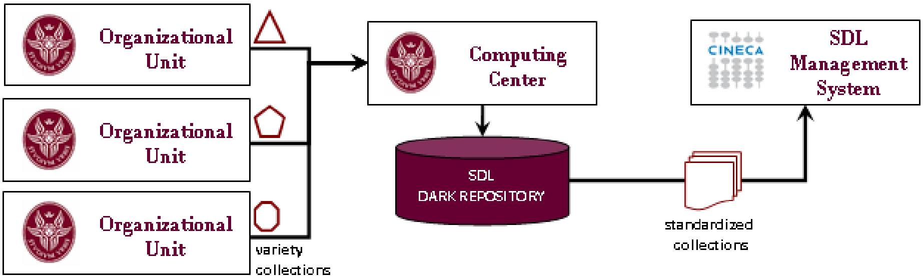

Fig. 1. Pre-ingestion workflow overview, from variety to standardized collections

The pre-ingestion workflow consists of all those activities necessary to prepare digital objects and metadata for the automatic building of the SIP, in conformance with the SDL metadata framework defined. In practice it organizes, structures and enriches the flow of digital materials from Sapienza organization units toward the Sapienza SIP building DLS.

The design and development of the workflow has essentially consisted of the following activities:

- Detection of available digital materials;
- Analysis, normalization and enrichment of both metadata and digital objects;
- Modelling of resources and resources' collections information for the submission objective;
- Modelling of provenance information being collecting in pre-ingestion activities;
- Designing and development of a local database as "metadata nursery" for the production of the final SIP's version.

7.5 Selection of Representative Samples and SIP Building

From the analysis of the existing materials were identified the representative samples of diverse types of materials(collections, videos, images, maps, books), to which applying specific content model. The building process of the SIP was applied to the samples selected and was performed throughout activities like the normalization of files' naming, the organization of collections, resources and digital objects, the creation and encoding of metadata files. The building process was tested and integrated many times during the experiment of the *ingestion*, until the maturity of the metadata framework.

At the conclusion of the first phase of project the SIP building was tested in 6 retrospective conversions of existing collections, compounding images and video, and described by spreadsheets and database information.

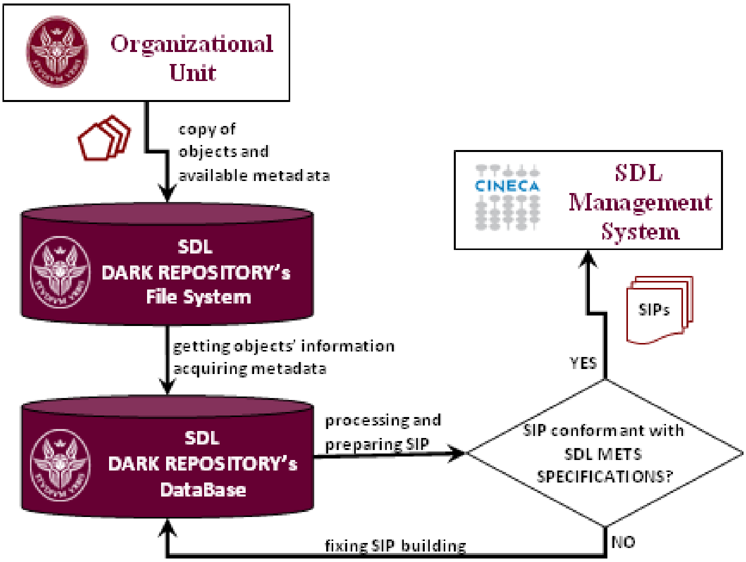
Organizational Unit
copy of objects and available metadata
SDL DARK REPOSITORY's File System
getting objects' information acquiring metadata
SDL DARK REPOSITORY's DataBase
processing and preparing SIP
SIP conformant with SDL METS SPECIFICATIONS?
YES
NO
fixing SIP building
SIPs
CINECA
SDL Management System

Fig. 2. SIP building development

For each sample were defined constraints on the information structure, which characterizes the originating model and unleashes the relevant FC content model.

The SIP building process consists of transforming the sources information (databases/datasets/spreadsheets/systems folders) in metadata, enriching it with provenance, context, reference, fixity (OAIS) metadata, and with basic knowledge domain semantics, encoding it in XML files valid for the XML schemas specifications, and combining them following the METS profile specification of the SDL metadata framework.

The SIP built has the following characteristics:

- based on METS as metadata container;
- encompasses different descriptive and administrative standards;

- conformant to specific metadata standard guidelines (i.e. DLF aquifer);
- conformant to Sapienza customized FC content models;
- exists independently and redundantly apart from the DLMS;
- open to the inclusion of other metadata standards that are more descriptive than MODS core description.

7.6 SDL DLMS Prototype

The SDL DLMS gets in ingestion the SIP built in conformance with the SDL metadata framework, it archives the objects and information in FC, and makes the SIP's resources available to the SDL exploration's system which allows to navigate, browse and search resources.

The prototype is internally available in order to give the opportunity to the project's community of participating to the optimization of the UI and Users' services, as well as testing functionalities, and improving the architectural information structure of the web interface. At this moment the prototype was peopled with 6 collections containing almost 15,000 items, differentiated in four resources types, images, books, videos, and maps. Furthermore, was uploaded a collection created *ad hoc* for the prototype's homepage, that was created by reusing resources, already ingested in the system. This has verified that the system holds all necessary information useful to create new digital objects (items or collections), that are aggregations of existing resources.

8 Future Developments

The second phase of the project will provide the DL with a set of specific DLS which will essentially support 1) the submission of new digital items and collections, 2) the participation of Sapienza community for collecting new materials, 3) the optimizazion of dissemination tools by means of customized interfaces (web users, OAI-PMH, Web Services…), and the integration of services for the exhibition/export of metadata for third party resources aggregators: InternetCulturale[11], Europeana[12], World Digital Library[13], 4) the integration of metadata supporting preservation planning and administration.

The workflow of the overall materials submission will be ruled by SDL guidelines wherein will be defined the Sapienza digital policies. The guidelines will harmonize the way of creating, and producing digital resources for the DL: specific paths for different content models will be described in order to support the community in being aware about the digital resources' management.

Stable METS and MODS profiles will be released during this phase.

At the end of 2012, the system will be released in production for the main functionalities, and will be open to other partners.

[11] http://www.internetculturale.it/opencms/opencms/it/

[12] http://www.europeana.eu/portal/

[13] http://www.wdl.org/en/

References

1. Consultative Committee for Space Data Systems Reference Model for an Open Archival Information System (OAIS), Blue Book. Issue 1 (January 2002), http://public.ccsds.org/publications/archive/650x0b1.pdf
2. Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V., Schuldt, H.: The DELOS Digital Library Reference Model - Foundations for Digital Libraries, Version 0.98 (February 2008), http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf
3. Fedora 3.5 Documentation, Content Model Architecture, https://wiki.duraspace.org/display/FEDORA35/Content+Model+Architecture
4. Metadata Encoding & Transmission Standard (METS), http://www.loc.gov/standards/mets/
5. Metadata Object Description Schema (MODS), http://www.loc.gov/standards/mods/
6. PREservation Metadata Implementation Strategies (PREMIS), http://www.loc.gov/standards/premis/
7. Digital Library Federation/Aquifer Implementation Guidelines for Shareable MODS Records, https://wiki.dlib.indiana.edu/confluence/download/attachments/24288/DLFMODS_ImplementationGuidelines.pdf
8. Metadata for Digital Content Developing institution-wide policies and standards at the Library of Congress, http://www.loc.gov/standards/mdc/elements/
9. PREMIS Editorial Committee, PREMIS Data Dictionary for Preservation Metadata version 2.0 (March 2008), http://www.loc.gov/standards/premis/v2/premis-2-0.pdf