Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 38 (2014) 4 – 11

Procedia
Computer Science

10th Italian Research Conference on Digital Libraries, IRCDL 2014

The Sapienza Digital Library from the holistic vision to the actual implementation

Tiziana Catarci, Angela Di Iorio, Marco Schaerf

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

Abstract

The Sapienza Digital Library (SDL) was released in December 2013 as result of a research project undertaken by Sapienza University of Rome and the Cineca consortium, since 2011. The digital library has been collecting materials coming from different kind of organizations including departments, libraries, and archives, belonging or donated to Sapienza University. The main result of the project was the development of an information framework supporting multidisciplinary organizations in managing digital materials, maintaining scientific, organizational and operational responsibilities. The technical solution adopted has found a trade-off between the standardization of the digital processes and products, and the preservation of the scientific materials' peculiarities. The automatic standard translation, and the enrichment of the digital resource's metadata have reached the main goal of providing digital resources with the essential information necessary to their management in different technological contexts. The reuse of the digital information and contents, in different application contexts, has converted the holistic vision of a digital library in the implementation of an information infrastructure, setting the foundation for the long-term access and usability of its digital assets.

Keywords: Digital Libraries, Long Term Digital Preservation

1. Introduction and motivations

In the Digital Library (DL) world, the increasing need of managing huge amount of digital resources, produced and consumed in different, and sometimes unpredictable contexts, requires information system that can adapt to a changing environment. In this paper we present our experience in the development of Sapienza Digital Library (SDL). SDL has started as research project undertaken by Sapienza University of Rome (Sapienza), and the Italian super-computer center Cineca, in January 2011. The mission of the project was to provide Sapienza University with a modern digital library collecting multidisciplinary material, coming from the Sapienza's communities, that represent almost all knowledge disciplines. The first requirement was to build a central point for digital libraries services, useful to provide access to the material respondent to different communities' needs, and useful to enable access through different contexts. The digital library management system is the mean for making digital material available to other systems, and providing services for specific contexts and objectives, in an interoperable manner.

[*] Angela Di Iorio. Tel.: +39-0649913866. Email: angela.diiorio@uniroma1.it

doi:10.1016/j.procs.2014.10.002

The Sapienza Digital Library, following this requirement, adopted the strategy of making digital material accessible in the long term by means of the adoption of community wide-spread digital library models and standards[1]. The workflow used for producing the building blocks of the information infrastructure allowed for the agile production and maintenance of digital resources, while structuring them for their archiving and reuse. The project's vision was to provide Sapienza's community with a digital library supporting the use of digital material managed, owned and produced by the Sapienza University. Initially expressed by the Sapienza's Libraries Committee, the main requirement was dealing with heterogeneous data and material, both scientific and educational, coming from University's libraries, museums, archives, and research departments. The Sapienza Digital Library, as information system, would have been a forefront infrastructure, flexible, interoperable, and connected to the most relevant international projects and initiatives, enabling the reuse of digital material and supporting the exchange of knowledge in different contexts and for different communities. Consequently, the general requirements are:

- Better dissemination of the research results, by means of web communication tools (social, professional and content specific), with a particular concern to the persistence of the resources, and the assessment of its impact.
- Enrichment of the resources by including representation information describing multidimensional relationships, revealing connections among contents, and using multiple classification systems and authority lists.
- Increasing the return of investment, by means of advanced technological services and uses open to the Sapienza community, not only for the fruition of material but also for the discover of new contexts, and unpredictable uses.
- Empowerment of the University information infrastructure by means of the interdisciplinary knowledge of the contributing research groups (scholars, researchers, librarians, archivists, museologists, IT specialists...) cooperating to allow the enhanced management of different contents and languages, inside of the same technological infrastructure.

The multidisciplinary vision, and the multifaceted aims required the cooperation of different expertise, likewise specialist skills, belonging to the University. The interdisciplinary research center Digilab was responsible for the project management and coordinated the selection of the scientific material for the digitization. The computing center (InfoSapienza) and the University Library System (SBS) had the responsibility for the technical coordination and design of the metadata infrastructure, and the libraries digitization projects. The University's organizations involved in the project's development have been cooperating to set the organizational general objectives, reflecting the *desiderata* about the use of the digital material in the medium and long term. The objectives defined are essentially applicable to any kind of university or educational institution, and, extending the vision, to any institution that needs to manage digital material, under similar perspective. The objective defined and shared by Sapienza's Organizations are:

- collecting, archiving and preserving broad varieties of digital material (born-digital and digitized), supporting intellectual and valuable contents, making it available to different kinds of exploitation objectives;
- supporting typical browsing and searching services on the metadata, and whereas applicable the full-text search, facilitating the access to the relevant digital contents;
- enabling the opportunity of reusing digital material and/or its parts in different contexts;
- optimizing and enhancing digital library services through the semantic web technologies and social tools;
- supporting services for submission, dissemination, and preservation of the Sapienza's digital assets, allowing the highest level of interoperability with external systems.

The Sapienza Libraries Committee worked on an initial evaluation of the available software solutions, focusing on the metadata standards supported, and concluded that the solutions were more focused on supporting specific types of material(special collections, open access, digitized material), and could not cover the needs defined by the Sapienza vision. Thus decided to undertake the path of the development of an architecture based on the integration of different information management systems, contributing to the information infrastructure of the Sapienza Digital Library, and sharing the same metadata infrastructure for the exchange of the resources. Furthermore, the long term perspective of the digital library project, required a strategy for the Long Term Digital Preservation (LTDP), which has been a requirement for the overall technological architecture. The holistic vision of a digital library setting the foundation for an information infrastructure, aiming to manage scientific and cultural material, and the long term concern for digital assets, has driven the choice of the technological partnership with the Italian University consortium Cineca. The Cineca consortium supports the Italian research institutions and the scientific communities with a technological infrastructure for supercomputing, usually among the first 100 supercomputer sites. The Cineca consortium was chosen as the responsible for the development and the implementation of the digital library management system and the digital library services. The commitment of both partners was to build an infrastructure for digital library, and preservation services, in the long term, exploitable by other Organizations. The multi-organizational scale of the Sapienza University could have been an useful test-bed for experimenting the technical implantation of a system for the third party digital library and preservation services.

2. The reference models for digital libraries

The design of the project's information infrastructure followed the most adopted approaches in the international scenario of the DLs projects. The functional model Open Archival Information System (OAIS)[2] which defines the OAIS as "An Archive, consisting of an organization, which may be part of a larger organization, of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community", was the first reference model taken into account, considering that the definition is tailored to the SDL project's vision. The Information Package(IP), defined by the OAIS, was the focus of the experimentation for building a consistent set of information, covering the information needs, required by the different OAIS functional scenarios: Ingestion (Submission IP), Archiving (Archival IP) and Access (Dissemination IP). The digital library and the preservation services would have been based on the information conveyed by the IP, which is enriched by the different components, supporting the management of the information infrastructure.

The DELOS Network of Excellence reference model[3], and its evolution, the "DL.org Digital Library Interoperability, Best Practices and Modelling Foundations"[4], was the second reference model taken into account for the development of the entire SDL architecture. The three-tier framework[3 4] is represented by the Digital Library(DL), the Digital Library System(DLS), and the Digital Library Management System(DLMS), where "the concept of Digital Library is intended to capture an abstract system that consists of both physical and virtual components, the Digital Library System and the Digital Library Management System capture concrete software systems. For every Digital Library, there is a unique Digital Library System in operation (possibly consisting of many interconnected smaller Digital Library Systems),(...) The DL is thus the abstract entity that 'lives' thanks to the software system constituting the DLS and the DLMS is the software system that is conceived to support the life cycle of one or more DLSs." The different DLSs and the DLMS that were developed or experimented in the SDL research project are graphically represented in Figure 1. The incremental design, which has driven their development, is represented by the circular flow, passing through the main concepts underlying all digital library systems[3]. The different DLSs and the DLMS composing the Sapienza Digital Library are related to the super domain "Organization"[4], which in this case is linked to the establishing Institution, the Sapienza University and, consequently to its components, the Sapienza's Organizational units. The metadata infrastructure captures not only the set of information belonging to digital objects, but also the Organization's super domain context which is the profiling information structure for each IP, and it is used to connect and exchange information among the different DLSs. The metadata infrastructure supports the management of the overall Digital Library information system. Taking into account the conceptual reference models, the research Project has started its development in January 2011. The project was developed in three main phases. The first consisted in the design and the development of all main architectural components, abstractly represented in Figure 1:

- the repository and service system for archiving, managing and preserving resources (Repository, Service Delivery Platform and DLMS);
- the web portal as system interface, for user and digital service (SDL web portal);
- the ancillary system of massive retrospective conversion of digital material, for providing standard conforming IP and for supporting and testing the overall workflow of IPs' production;
- the metadata infrastructure necessary for exchanging resources between systems, DLSs and DLMS.

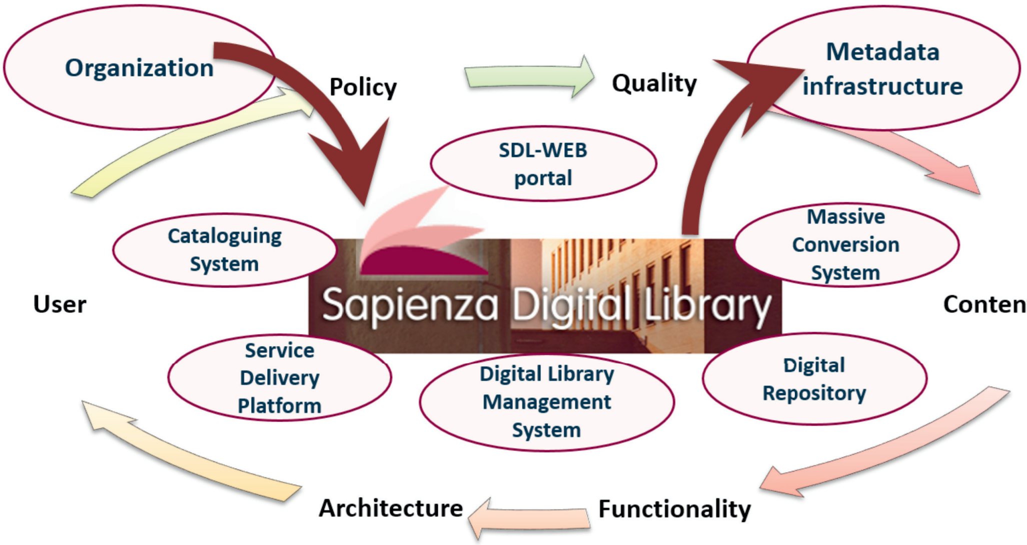

Fig. 1. The reference model basis for the SDL's architecture

The first phase was concluded by the presentation of the first prototype of the system, to the Sapienza community at the end of 2011. The second phase of the project has consisted in:

- the enrichment of the metadata infrastructure, based on the ingestion of more structurally complex resources;
- the development of a new system for the cataloguing and the production of resources coming from new digitization projects;
- the improvement of the overall architecture system's components (DLSs, DLMS);
- the web portal customization with the Sapienza Institutional graphical template;
- the massive improvement of the data quality and its visualization, following advice and requests of the responsible scientific working groups, and coherently to the new graphical interface data disposition.

The second prototype of the system was released at the end of 2012, and was submitted to the Sapienza community users for using and testing. The third phase of the system has consisted in the optimization of the overall architecture, based on the Sapienza community users feedback and with the increment of new collections coming the from the release of the Cataloguing system. The SDL web portal was publicly published as beta version at the end of 2013.

3. Developing SDL

The first phase of the project development has used the IPs produced by the conversion of the existing digital material, here summarized in the following list:

- a sample collection of 1954 images, coming from a scholar's archive concerning teaching material of theatrical culture and history, and a special collection of 191 digital images scanned (TIFF and JPEG) from transparency, coming from the Architecture library, and containing architectural photographs donated by a well-known Sapienza's Faculty member;
- a video collection of 4647 items describing all mediateca holding and containing 927 digital video objects;
- two museums' collections coming from the two Sapienza museums (Chemistry, and the Origin) and comprehending the description of 947 items, and 669 digital image objects;
- Around 300.000 digital objects produced by the first Sapienza's libraries digitization project Prodigi (`http://prodigi.uniroma1.it`) run in 2007, comprehending around 500 items (ancient books, and maps), coming

from 9 libraries, mostly scanned in high resolution TIFF format, and reproduced also in the derivative format JPEG for the internet fruition; 362 PhD thesis coming from the Physics' library provided in pdf format; and 3 ancient maps coming from the Architecture library.

About the material used at the beginning of project, it is important to underline three influencing aspects on the building of digital resources conforming with OAIS IP structure :

- the descriptive metadata provided with the material were provided in both standard description (bibliographic description coded in ISO2709) and non-standard description. The non-standardized descriptions were usually structured into local databases or spreadsheets often not normalized;
- no other metadata then the descriptive metadata about the intellectual content were provided with digital objects;
- no controlled vocabularies were shared by the heterogeneous material;
- the digital objects were multi-formats, differently structured, and differently related to the descriptive metadata, and in many cases not consistently.

Consequently, it was necessary to identify, normalize, classify, organize, enrich, and package the incoming metadata and objects into a consistent set of digital objects and metadata related to them. The package had to be structured and coded in the standards adopted, and profiled by designing specific content models, shared by both partners for conveying information about resources and for applying the relevant services. The standards adopted for SDL are:

- Metadata Objects Description Standard (MODS) for describing the intellectual contents;
- PREservation Metadata Implementation Strategies (PREMIS) (`http://www.loc.gov/standards/premis/`) for managing preservation metadata;
- Metadata Encoding and Transmission Standard (METS) (`http://www.loc.gov/standards/mets/`) for packaging metadata belonging to the digital resource.

The choice was technically made, considering the flexibility of METS, and the higher level of information granularity of MODS with respect to other wide spread descriptive standards, like Dublin Core (DC) deemed more interoperable. In addition the MODS native controlled vocabularies are actually exposed as Linked Data Service Authority and Vocabularies (LC Linked Data Service, `http://id.loc.gov/`) by the Library of Congress, allowing to connect the SDL resources to the Linked Data Cloud (Linked Open Data, `http://linkeddata.org`). The metadata infrastructure, built on the standard structure and shared by other systems, could have been the basement for the exchange of the SDL OAIS IPs among internal DLSs and toward other external systems. Furthermore, it will ease interoperable translations toward different data and metadata structures, considering the wide spread adoption of the standards by the communities, the numerous implementations, tools and mappings available to the communities, and the endorsement of the maintenance by the Library of Congress with the support of International editorial committees. The IPs are referred in the SDL domain as Digital Resources DRs, that are the "building blocks" of the digital library. 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 inside of the system, as a set of different kind of metadata related to object files. Each box is labeled with the name of the related standard XML schema name. The descriptive metadata, pointed by the blue arrows, is coded into two descriptive standards. MODS which reflects the granularity of MARC21, and DC, commonly adopted in other contexts, not strictly related to the libraries world, and consequently considered more interoperable. The third descriptive box represents the original descriptive record, coming from the data provider, and stored as an external file. This choice was made in order to maintain the original information provided with the digital content, and to preserve also those information that could have not been mapped completely, toward the standard metadata semantics. The administrative metadata, pointed by the orange arrows, is coded in the specific entities, that constitute the PREMIS data model[5]. The inventory metadata listing the files' names and locations, and the structural metadata, pointed by the red arrow, are coded in two specific 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. Finally, the objects part of DR, pointed by the purple arrow is represented by the set of files. The XML coding of the DRs is the transfer format, by which

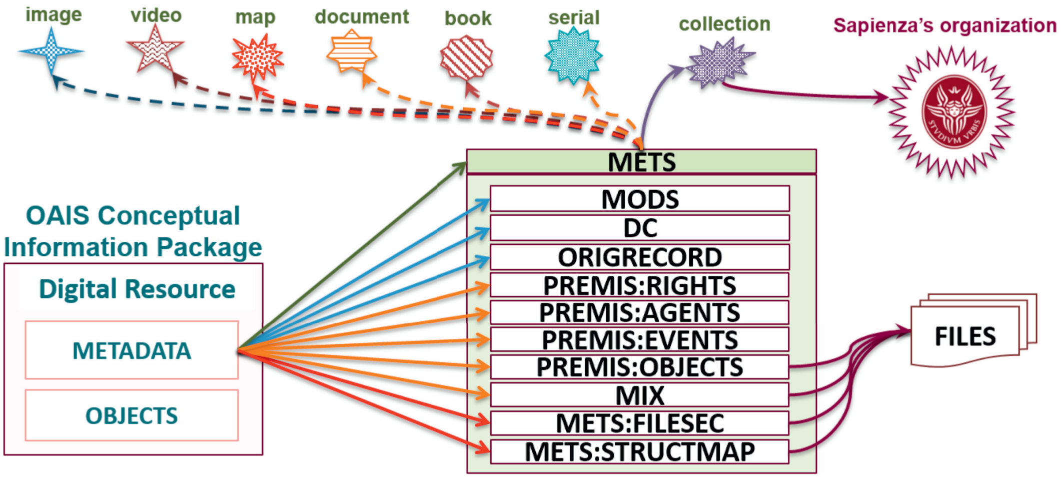

Fig. 2. SDL Digital Resource structure and content models

the OAIS IP must be produced in order to be transferred, ingested and stored, into the digital repository system and managed by the DLMS, which provides the coherent translation of the metadata profile into the proper management services: submission, archiving and dissemination[2].

The chosen repository software was Fedora (Fedora Commons, `http://fedora-commons.org/`) because of its compatibility with the standards adopted, its commitment to provide "the basis for ensuring long-term durability of the information" and the semantic web technology compatibility of Fedora Commons that is based on RDF (Resource Description Framework, `http://www.w3.org/RDF/`).

The Sapienza and Cineca technical working group agreed on the adoption of the repository software, and the SDL content metadata models were matched with the content technological models developed by Cineca, and integrated with the services deployment system for e-learning already managed by Cineca[6]. The agreement on the mandatory information, for the exchange of DRs, was necessary in order to transfer zipped DRs from the dark archive of Sapienza into the geographically separated area (from Rome to Bologna) for submitting them to the ingestion service. Figure 3 shows how the tasks of the DRs' production(Massive conversion and Cataloguing), are coded as SDL METS SIP. The METS SIPs are ingested by the DLMS, which archives the packages and makes it available to the dissemination services (web portal and others). Considering that the standards adopted are usually released to the communities, accompanied with the XML schema (`http://www.w3.org/XML/Schema`), which is the formal definition of the XML semantics, during the project was developed a software properly customized to the SDL requirements for the DRs production conforming with the standards' XML schema. The development of this software was a practical choice for the DRs production conforming to the metadata infrastructure and the content metadata models designed for the Sapienza digital material. The incremental update of the overall customization of the metadata infrastructure, adopting the evolution of the projects and the new choices, coming from both technical and scientific working groups, has relied on the flexibility of this solution.

The software was used to test the consistency and completeness of the metadata infrastructure, to properly define the content metadata models, to check the correspondence of the SIP package to the DIP package, and to improve the quality of data and homogeneity of the data rendering. In addition, the data management distributed and replicated on different software solutions and different architecture covers the LTDP requirements for risk management and maintaining the architectural heterogeneity[7]. The system also includes a lightweight data-entry front-end for digital-inventory that allows to precisely structure the descriptive information about heterogeneous material of interest for the Humanities community.

The set of data managed for the cataloguing description reflects the aim of integrating knowledge domains. Because the collection selected concern art history, theater, archeology, cultural archives, and could potentially serve

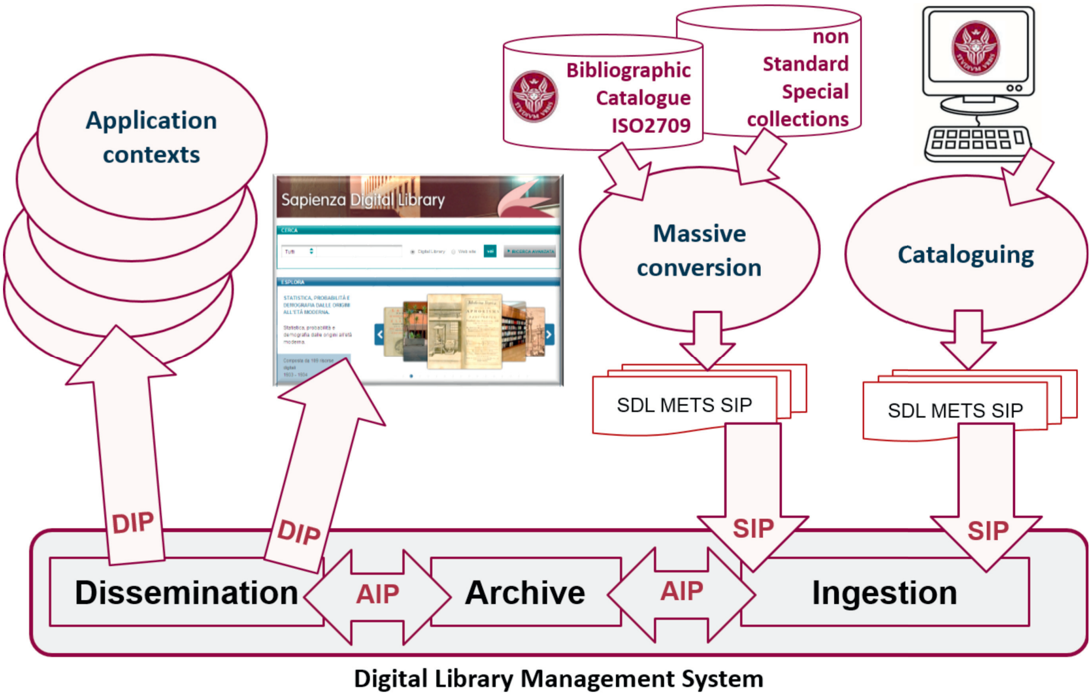

Fig. 3. Digital resource structure

all Sapienza scientific community, it has been decided to choose the reference controlled vocabularies and authority lists following the Italian and the International practices: the use of the Subject Headings of Florence Library (Nuovo Soggettario di Firenze, `http://thes.bncf.firenze.sbn.it/`), the CulturaItalia PICO Thesaurus (`http://www.culturaitalia.it/pico/thesaurus/4.3/thesaurus_4.3.0.skos.xml`), the Getty Thesaurus of Geographic Names (`http://www.getty.edu/research/tools/vocabularies/tgn/`), the database of the geographical names GeoNames (`http://www.geonames.org/`), and the use of identifiers of the Virtual International Authority File (VIAF, `http://viaf.org/`) were deemed necessary to for covering the multidisciplinary material's description need. The system was publicly opened the 20th of December 2013, as Beta version 1.0 (Sapienza Digital Library, `http://sdl.uniroma1.it`) 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. The Table 3 shows the increment of the resources' submission differentiated by the workflow and phases of the project. During the development was also performed an experimentation of import of metadata records coming from the Sapienza's research catalogue and the export of more than 100.000 metadata objects for the Europeana portal (Europeana think culture, `http://www.europeana.eu/`) portal, for the project Linked Heritage (`http://www.linkedheritage.eu/`). The tasks performed has substantiated the idea of infrastructure for exploiting the digital assets. After all the improvements to the beta release, and the corrective intervention on the harmonization of the overall metadata infrastructure between the two DLSs, all new collections, under preparation, will be ingested. The quality assurance of the digitized books coming from the mass-digitization Google Books project (`http://books.google.com/`), which has involved Sapienza as participating partner[8], will provide more than 30 new collections containing around 20 millions of digital objects.

4. Conclusions and Acknowledgments

The technological exploitation of the digital material, by means of the connection to Linked Data Cloud as well as the availability of APIs for the use and reuse of contents, metadata, and tools, will be providing more and more re-

Table 1. Rate of exchanging packages distinguished in work-flow, resources and related digital objects, and divided per years and project's phases

Phase	YYYY-MM	Workflow	Resources	Objects
1st	2011-11	Massive	7647	3764
2nd	2012-11	Massive	603	108856
2nd	2012-11	DLMS Import	2543	0
3rd	2013-05	Massive Export	105489	105489
3rd	2013-05	Massive	3829	72914
3rd	2013-11	Massive	293	169372
3rd	2013-11	Cataloguing	2240	37063

fined services for users, under the condition of long term access provision. Coherently to the European trend focused on ICT-based infrastructures and services connecting and supporting a broad range of scientific and cultural disciplines, the digital library project has started the process of translation from the local fruition of material, to the global exploitation making it more reusable and easily linked to new application scenarios. The holistic vision aiming to provide Sapienza with a digital library conceived as an information infrastructure, serving multidisciplinary communities, and overcoming all the organizational and disciplinary differences, will leverage the research and development of the Sapienza University.

We would like to thank all the people that have participated to the working groups, and that in different way and under the different perspective of their own expertise, scientific and technical, have supported and collaborated to the test and the evaluation of the overall Sapienza Digital Library system, allowing improvements of the system and refinements of the services. In particular we would like to thank Giovanni Ragone, Giovanni Solimine, Mariella Guercio and Gianfranco Crupi, the members of DIGILAB involved in the project, Cecilia Carloni, Salvatore Casaluci, Giorgio Felici, Silvia Ortolani, Isabella Tartaglia, the members of Sistema Bibliotecario Sapienza Adriana Magarotto, Ezio Tarantino, Maura Quaquarelli and the members of InfoSapienza, Luciano Longhi and Matteo Righetti. Last but not least, we thank Cineca for their valuable partnership and the technological support.

References

1. 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*; vol. 354 CCIS of *8th Italian Research Conference on Digital Libraries, IRCDL 2012*. Springer. ISBN 18650929 (ISSN); 9783642358333 (ISBN); 2013, p. 172–183. URL: `http://www.scopus.com/inward/record.url?eid=2-s2.0-84873865280&partnerID=40&md5=d8b5b1f12a673c347ec521d4a4e8b391`. doi:10.1007/978-3-642-35834-0_18.
2. Consultative Committee for Space Data, . Reference Model for an Open Archival Information System (OAIS), Recommended Practice CCSDS 650.0-M-2 Magenta Book. 2012. URL: `http://public.ccsds.org/publications/archive/652x0m1.pdf`.
3. Candela, L., Castelli, D., et al. The DELOS Digital Library Reference Model - Foundations for Digital Libraries. Version 0.98. Tech. Rep.; ISTI-CNR at Gruppo ALI; 2008. URL: `http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf`.
4. Candela, L., Athanasopoulos, G., Castelli, D., et al. The Digital Library Reference Model. Tech. Rep.; DL.org: Coordination Action on Digital Library Interoperability, Best Practices and Modelling Foundations; 2011. URL: `http://bscw.research-infrastructures.eu/pub/bscw.cgi/d222816/D3.2bDigitalLibraryReferenceModel.pdf`.
5. PREMIS Editorial Committee, . PREMIS Data Dictionary for Preservation Metadata version 2.2. 2012. URL: `www.loc.gov/standards/premis/v2/premis-2-2.pdf`.
6. Bertazzo, M., Di Iorio, A.. Preserving and delivering audiovisual content integrating fedora commons and mediamosa. *Journal of Digital Information* 2012;**13**(1).
7. Caplan, P., Kehoe, W., Pawletko, J.. Towards interoperable preservation repositories (tipr). In: *Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop*. ACM; 2010, p. 16.
8. Magarotto, A., Quaquarelli, M., Vallania, M.. Il progetto di digitalizzazione google books presso le biblioteche della sapienza, università di roma. *Digitalia, Rivista del digitale nei Beni Culturali* 2013;**2**.