Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 38 (2014) 128 – 132

Procedia Computer Science

10th Italian Research Conference on Digital Libraries, IRCDL 2014

An Integrated Management System for Multimedia Digital Library

N. Barbuti[a]*, S. Ferilli[b], D. Redavid[c], T. Caldarola[d]

[a]*Dept. of Classical and Late Antiquity Studies - University of Bari, Piazza Umberto I, 1 70121 Bari, Italy*
[b]*Dept. of Computer Science - University of Bari, Via E. Orabona, 4, 70125 Bari, Italy*
[c]*Artificial Brain S.r.l., Via Piave 63, 70125 Bari, Italy*
[d]*D.A.BI.MUS. S.r.l., Via Quintino Sella, 268, 70123 Bari, Italy*

Abstract

Contemporary libraries have changed quickly their social role and function due to the proliferation and diversification of multimedia digital documents, becoming complex networks able to support communication and collaboration among the various distributed users communities. Technologies have not grown in step with the needs generated by this new approach, except in specific areas and implications. Hence the need to design an integrated digital library architecture that covers by advanced techniques the whole spectrum of functionality, without which the same social and cultural function of a modern digital library is at risk. This paper briefly describes an architecture that aims to bridge this gap, bringing together the experience, expertise and software systems developed by university and companies researchers. A prototype of the system is under development.

Peer-review under responsibility of the Scientific Committee of IRCDL 2014

Keywords: Digital Library; Digital Library Management System; Digital Recognition; Layout Analisys

1. Introduction

Here Librarianship has evolved and taken new study and research fields in order to conceive and design new systems able to renew the management of information and, at the same time, to promote both cooperation between users and integration of heterogeneous information resources, ushering in the era of *Digital Libraries* (DL)[1]. Over the years, DL considerably has evolved from simple digital interface of public libraries' physical collections in

* Corresponding author. Tel.: +39-080-571-4712 ; fax: +39-080-571-4712 .
E-mail address: nicola.barbuti@uniba.it

Peer-review under responsibility of the Scientific Committee of IRCDL 2014
doi:10.1016/j.procs.2014.10.021

complex networks, able to support communication and collaboration among different users communities worldwide. By contemporary DL citizens can access, discuss, evaluate and develop different types of information content[2]. Facing to the progress in research aimed at creation of DLMS able to support DL containing homogeneous digital collections and metadata, there are still hard difficulties in designing and implementing DL effectively integrating, managing and making accessible multimedia digital objects and related metadata. This paper proposes an "integrated" DL architecture, designed as a prototype for the project D_ISRAELI, an approved project under the "Living Labs Smart Puglia 2020" call by Apulia Region, presented by University of Bari Aldo Moro and CeRDEM – Center for Research and Documentation in Judaism in the Mediterranean. The prototype, under development, aims at creating a DL on Jews history and culture in Apulia designed especially for librarianship and archival areas. The paper is structured as follow: section 2 presents a comparison about the major open source Digital Library Management Systems (DLMS) finalized to choose the best one to our aim. Section 3 details the components that should be integrated or enhanced in the choosen DLMS architecture in order to realize the set of innovative features identified by the component itself. Section 4 gives an overview of the major characteristics of the proposed DLMS, in particular the management of different types of metadata and text extraction and indexing from scanned digital objects. Section 5 gives the conclusions.

2. Open-Source Technological Base

The definition of the proposed architecture has required an analysis of available open source DLMS[3] in order to choose the one more suitable to our aims. Among the most known and used we have considered and compared *dSpace*, *EPrints* and *Greenstone*[4]. *Fedora Commons* has not been taken into account in this comparison because fundamentally it is a repository oriented to the digital data preservations rather than to the fruition. Furthermore, MARC protocol and the bibliographic data exchange protocol Z39.50 are not adequately supported by *Fedora Commons*, both non-negligible characteristics for the purposes of the DL to be realized. From the comparison between the three identified DLMS, the common features are: 1) OAI-PMH is supported, 2) storage and management of any type of content, 3) Multilingual Interface-oriented capabilities to the end user, and 4) production of statistical reports based on the count of records. Instead, they differ on the following characteristics:

- As unique identifier for the DL objects, dSpace uses CNRI Handle System (www.cnri.reston.va.us), Greenstone uses OAI Identifier (www.openarchives.org/OAI/2.0), while Eprints does not rely on any standard convention.
- In addition to the Dublin Core and METS metadata, common to all three DLMS, dSpace supports MARC/MODS too, while Greenstone NZGLS and AGLS.
- The search functionalities supported by dSpace (Field Specific, Boolean Logic and Sorting options) are a superset of those of EPrints and Greenstone.
- For the browsing functionality, EPrints and Greenstone allow the use of any field, while only Author, Title, Subject and Collection can be used with dSpace.
- User authentication in dSpace is possible via LDAP or Shibboleth, in Eprints only via LDAP, and User Groups in Greenstone.
- The databases that can be used are Oracle and PostgreSQL by dSpace, also MySQL and Cloud by EPrints, while Greenstone has an its own implementation.
- OAI-ORE, SWORD, SWAP are supported by dSpace (which adds SRW/U as extension of Z.39.50) and EPrints (which adds RDF), while Z39.50 by Greenstone.

The choice has fallen on dSpace essentially for the following three factors: 1) it supports better than the others the various metadata standards and protocols for interoperability, 2) it has been developed using only one programming language, and 3) it has a more complete documentation and there are various communities that provide support.

3. Architecture of the proposed system

The proposed DLMS therefore extends the architecture of dSpace adding specific modules for the management of innovative features. The architecture is divided into three levels, according to the conceptual framework in Fig. 1. The Application Layer includes the following components for access to the system:

- Web UI is the module for the Web-based access both to the back-office area (via IDPs) and into front-end of the Digital Library through various portals. The back-office area allows the insertion and editing of digital content and its metadata, as well as managing users for access. The front-end allows visualization and rendering of contents through a Web interface that combines all advanced file formats guaranteeing multimedia, multichannel and protection. It also presents the front-end for the collaborative tagging (operated by Web 2.0 module).
- Mobile Devices allows viewing and rendering of contents through tablets and smartphones.
- Monitoring is the module that allows you to observe the behavior of the system and to produce reports in Excel format, XML, and PDF with the possibility of representation through graphs.
- The I/O module include interfaces that allow the metadata exchange using OAIS, OAI-PMH, Z39.50 and OAI-ORE. It also enable the rendering of content as Open Data.

The Business Logic level includes the following modules suitable for the realization of the system functionalities:

- Core Tools is the module containing basic entities for system configuration and logging.
- Search Engine is the module that implements the functionality to support the information finding. Content indexing is done by the Lucene open source tool that implements techniques based on terms at the state of the art.
- Web 2.0 is the module for the management of collaborative tagging, i.e. it allows to perform the necessary functions to support the active participation of the portal users to the published contents.
- Access Management is the module for user authentication and profiling. It allows the access management both in normal mode or through IDP.
- Text Extractor is the module that enables the extraction of text from documents via ICRPad (see Sect. 4).
- Content Manager is the module designed to manage the objects, collections and licenses within the DL.
- Cataloging is the module that allows to manage digital contents that will populate the DL. The MAG/ICCU standard (version 2.0.1) compatible with standard METS will be used.
- Georeferentiation is the module that allows to reference the content in accordance with their geographic coordinates. This module will allow the storage and retrieval of content on the basis of spatial queries.

The Storage level manages access to the physical resources of the system deals with the organization of the content, including metadata, information about users and permissions associated with them, the status of the approval flow during the insertion of a content. Specifically: RDBMS Wrapper is the module that allows read/write access to the particular implementation of the DB (in this project is PostgreSQL, but is easily extendable to other types, e.g. Oracle). Bitstream Storage Manager is the module that permits the storage on the file system or SRB (backup tool). The DL operations will involve the various architectural modules. In addition to functionalities of a classical DL in the next section will be described some of the innovative aspects that will be integrated.

4. Innovative aspects of the proposed solution

To enable the efficient management and retrieval of content with different representation formats, we adopt a standard representation for administrative and descriptive metadata involving the use of different languages and hardly compatible with each other. To this end a good provision about the integration of tools that enable a separate management of metadata according to the various representation standards is necessary from the beginning. In order to facilitate interoperability, an extension of the dSpace basic architecture to support exchange protocols is required.

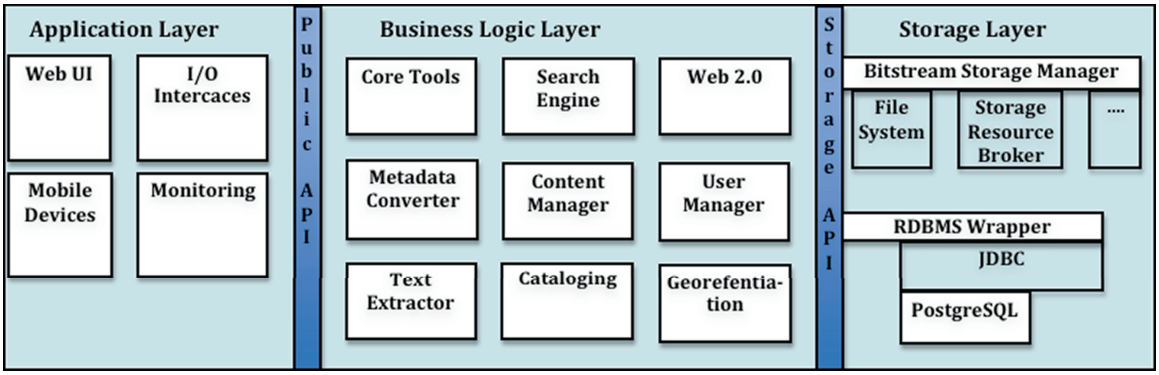

Fig. 1. Architecture of the proposed system.

The standards supported by the DLMS interface require the integration of tools which ensure the proper functioning and approaches that permit the mapping among the different metadata representation usable in the standard interface. One of the most innovative aspects will be related to the high interactivity that the DL will ensure to different user communities in response to their cognitive demands. This will be made possible by the innovative feature on the indexing of extracted texts directly from the digital document images including manuscripts with, e.g., historical content. This feature will allow to build a search engine able to handle directly the information contained in these documents enabling the application of semantic indexing techniques. To this end, into some modules will be implemented features that are not natively supported by dSpace.

In *Web UI* module will be integrated recognition features, graphic matching and text extraction from digital objects into PDF/A with printed and manuscripts content using the digital platform for the recognition ICRPad[5]. It will support Text-based lexical indexing through state of the art technology provided by Lucene and those based on co-occurrences and semantics provided by the system DOMINUS[6], which also includes advanced features for the selective reading of interesting portions of digital format documents. The DLMS will support digital objects in different formats, in particular FITS (Flexible Image Transport System) for the images.

In *Profiling and Authentication* module will be implemented a user profiling behavior functionality aimed at customization of services: using advanced techniques developed in the Artificial Intelligence field for tracing the interactions of each individual user with services, it will make possible to infer specific information regarding various spheres such as the special interests, the preferences of interaction, the goals, routine activities, and more. Based on this information, customized service for each user will be provided, making its experience with the DL easier and more productive[7].

In *Cataloging* module will be included a common higher-level metadata schema based on Semantic Web languages. This will allow to take advantages offered by the abstract properties of the metadata during the search operation of the DL contents.

5. Conclusions

All Digital libraries have been radically changed by the increase of new technologies, evolving from mere digital counterpart of public libraries' physical collections in complex networks, able to support communication and collaboration among users communities worldwide. This increase created the need for integrated systems able to manage DL by advanced functionality, today still unsatisfied. This article has proposed an innovative DL architecture that aims to bridge this gap by new features, such as the integration of technologies for processing documents covering all phases: acquisition, content extraction, indexing, searching and enjoyment. A prototype system is being developed with very interesting prospects which could eventually lead to realize a DL model that functions as an integrated system for preservation, management and use of complex multimedia digital objects.

Acknowledgements

This work is partially funded by the project "VINCENTE – A Virtual collective INtelligenCe ENvironment to develop sustainable Technology Entrepreneurship ecosystems" (PON 02_00563_3470993) funded by the Italian Ministry of University and Research (MIUR).

References

1. Arms WY. The1990s: The Formative Years of Digital Libraries. Library Hi Tech, 30(4), 2012, pp. 579 – 591.
2. Agosti M. Digital Libraries. *Mondo Digitale* settembre 2012;**43**:1-13.
3. Andro A, Asselin E, Maisonneuve M. Digital libraries: Comparision of 10 software. *Library Collections, Acquisitions, and Technical Services* 2012;**36**(3-4):79-83.
4. Tramboo SH, Shafi SM, Gul S. A Study on the Open Source Digital Library Software's: Special Reference to DSpace, EPrints and Greenstone. *International Journal of Computer Applications* (0975– 8887) 2012;**59**(16).
5. Barbuti N, Caldarola T. An innovative character recognition for ancient book and archival materials: A segmentation and self-learning based approach. In Agosti M, Esposito F, Ferilli S, Ferro F, editors. Communications in Computer and Information Science. Vol. 354: *Digital Libraries and Archives*, IRCDL 2012, Heidelberg: Springer, (pp. 261-270).
6. Ferilli S, Esposito F, Basile TMA, Redavid D, Villani I. DOMINUSplus - DOcument Management INtelligent Universal System (plus). In Agosti M, Esposito E, Meghini C, Orio N, editors. *Digital Libraries and Archives - Post-proceedings of the 7th Italian Research Conference* (IRCDL-2011), Communications in Computer and Information Science 249, 123-126, Springer, 2011.
7. Semeraro G, Costabile MF, Esposito F, Fanizzi N, Ferilli S. Machine Learning Techniques for Adaptive User Interfaces in a Corporate Digital Library Service. In *Machine Learning and Applications, Proceedings of the ACAI-99 Workshop W03 on Machine Learning in User Modeling*, 21-29, Chania, Crete, Greece, July 5-16, 1999.