Optimization of Knowledge Availability in an Institutional Repository

Filippo Eros Pani, Maria Ilaria Lunesu, Giulio Concas, Carlo Stara and Maria Pia Tilocca
Department of Electrics and Electronics Engineering, University of Cagliari, Piazza d'Armi, Cagliari, Italy

Keywords: Knowledge Management, Open Archive, Institutional Repository, Multimedia Content, Semantic Metadata.

Abstract: The Institutional Repositories (IRs) based on Open Archives represent one of the main free access tools for the results of scientific research, and their diffusion is continuously growing. In fact, more and more universities and research centers rely on IRs in order to optimize the management and diffusion of scientific work. During an industrial project aimed to the creation of the Analytic Sound Archive of Sardinia, was born the idea to create an Institutional Archive to solve the problems of organization and availability of information. The Archive would contain a linguistically annotated corpus to preserve, enhance and support the oral traditions in the Sardinian language. The distinctive feature of this work is the proposal of a new approach for formalization and management of knowledge using the tool DSpace (a Knowledge Management System typically used for the organization and management of text documents) to store, manage and query the electronic corpus made of audio clips. In this specific case we worked with a group of audio recordings in a corpus and linguistic information added to that corpus with annotations. The customization of the structures and interfaces of the chosen KMS, to ensure the availability and sharing of knowledge, are also closely tied to this research.

1 INTRODUCTION

Internet has become the most popular place for integration, exchange and sharing of information. The significant increase in its user base is flanked by an even more significant increase in the amount and types of available content, mainly due to the process of digitization of existing information and the creation of new information generated on the Net. The problem of content availability and organization depends on two strictly related issues: availability of appropriate capabilities of indexing and retrieval inside knowledge management tools, and the ability of users to understand and use these features.

Experience suggests that the solution of this problem can be found in the use of organized schemas and relevant standardized metadata, or data tiers, that allow us to describe, classify, and organize basic information, allowing retrieval and use (Heery and Patel, 2000) (Lagoze and Van de Sompel, 2003) (Lunesu, Pani and Concas, 2011).

The trend to the standardization and interoperability of the tools and system for archiving highly structured information has brought, in recent years, to establish institutional interoperable repositories created on Open Archive Initiative (OAI) architecture, based on standardized and shared metadata schemas (Chopey, 2005) (Solodovnik, 2011).

IRs have become the privileged place for the dissemination and use of scientific knowledge, for archiving and preservation of digital resources, as well as for the exchange of metadata, in order to promote and support the organization and management of digital resources created by institutions and their members. The Dublin Core (DC) metadata schema represents a reliable support to manage such contents, as it easily matches with other metadata schemas, increasing the granularity and refinement of their structures.

Building solid, well-structured and shared metadata schemas as DC significantly improves the classification and the availability of stored resources while ensuring greater knowledge availability for end users (Hutt and Riley, 2005). This is the contexts where the “Analytic Sound Archive of Sardinia” project belongs. The project aims to create an IR that contains a linguistically annotated corpus to preserve, enhance and support Sardinian language oral traditions, especially improvised poetry.

The electronic corpus will be made of a collection of audio clips, including poetry contests called cantada, canto a chitarra, mutetu longu, tenore, etc. The Sardinian samples are stored and

Eros Pani F., Ilaria Lunesu M., Concas G., Stara C. and Pia Tilocca M..
Optimization of Knowledge Availability in an Institutional Repository.
DOI: 10.5220/0004173203020306
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 302-306
ISBN: 978-989-8565-30-3

recorded at different language levels, from acoustics-phonetics up to linguistic and paralinguistic levels. In compliance with OAI, this corpus will be inserted into an open IR and therefore made available to researchers of the Sardinian language, as well as all individuals with an interest for the matter. The proposed work therefore aims to show a concrete solution for the realization of such an archive based on the use of the open-source software DSpace.

This document is organized as follows: Section 2 shows the development of institutional repositories and the Dublin Core standard; in Section 3 we present the approach we used to build the Analytic Sound Archive of Sardinia; Section 4 shows the description of the steps undertaken to customize the tools we used. The last section hosts our final observations about the work done and the possible future.

2 INSTITUTIONAL REPOSITORIES

Since the Nineties of the last century, a new phenomenon has affected the process of scientific communication and knowledge sharing: the appearance and spread of digital repositories of scientific contributions in order to make the movement of information more "agile". In 1991, Paul Ginsparg, at Los Alamos National Laboratory (USA), paves the way to the arXiv, a repository of works on Physics and Mathematics.

In June 1994 Stevan Harnad sent a “subversive” proposal to the mailing list of the Virginia Polytechnic Institute: they ought to share their ideas through the contributions of self-archiving on the internet, in order to communicate their results more effectively.

A new kind of open archive begins to emerge: the IR, supported and managed by an institution, such as an university, which incorporates the contributions of its researchers.

In October 1999 a group of researchers and librarians in Santa Fe (USA) marked the turning point: the rise of the OAI, essential to the management of technical aspects such as protocols and data exchange standards, localization and subsequent retrieval of scientific contribution, and software such as operating tools and for indexing. The OAI consolidates the experience and the previous techniques and, above all, embodies a sort of “philosophical” awareness.

At the beginning of the new century, when the archives have already opened and operational, the expression “Open Access” is used for the first time in a public document: Budapest Open Access Initiative manifesto (2002). It suggested for the first time to adopt both strategies, called “complementary”, to encourage the spread of the open access system: the "self-archiving", i.e. archiving in institutional and disciplinary "open electronic archives", of articles by researcher and "open access journals," the new generation of scientific open access journals.

3 OPTIMIZATION OF KNOWLEDGE MANAGEMENT SYSTEM

During an industrial project aimed to the creation of the Analytic Sound Archive of Sardinia, the idea to create an Institutional Archive to solve the problems of organization and availability of information came forth. There was, thus, the need to have an efficient tool that could classify and store the vast amount of knowledge contained in an electronic corpus of Sardinian language, and that could, at the same time, allow a high usability in terms of ease of reference as well as ease of query and communication.

3.1 Formalization of Knowledge

In this context, knowledge is represented not only by the texts of the corpus, but especially by the meta language and linguistic annotations that enhance them.

For each audio clip, a set of metainformation describing the content is needed in order to enable the search and retrieval of data by local author, title, date of recording, to more particular features like linguistic variety or singing type. Each audio clip is also enhanced by a set of linguistic annotations.

The insertion of the audio clips in the chosen KMS required the formalization of all the associated metainformation in the form of a structured set of metadata.

Linguists and musicologists working on the Sardinian Linguistic Sound Archive chose a list of possible annotation levels (syllable, tone, morpheme, syntagm, accents, etc.), useful for both linguistic and musical analysis of audio recordings.

3.2 Choice and Customization of the KMS

DSpace, an open source software package developed in 2000 in the context of a joint project of the Massachussetts Institute of Technology with Hewlett-Packard, provides all the necessary tools for creation and management of an IR based on the Open Access model. Such an IR can collect, store, index, preserve and make accessible the information output created by universities and research institutes in a digital format.

DSpace is designed as a central storage facility able to collect all kinds of content from the community relating to the institution through a user interface as simple and intuitive as possible. It can collect various types of digital resources including text, images, video, audio, articles and preprints, technical reports, working papers, datasets, and learning objects directly from the creators.

DSpace was chosen to realize the Analytic Sound Archive of Sardinia as it fulfills all the requirements asked by linguists and musicologists. It is in fact completely customizable, supports natively Qualified DC metadata schema and is compatible with OAI with the support of OAI-PMH. The proposed approach allows to insert the corpus and the associated knowledge inside of DSpace, ensuring the maintenance of its structure and the ability to interrogate and update it easily by adding or modifying its contents. Each text of the corpus is inserted into a DSpace item so that it can be uniquely associated with all of the metadata needed for the linguistic analysis. The audio file contains the registrations and the original files with the annotations are loaded inside of the item as a bitstream, while the metadata are stored in the system database.

The first step consisted in the insertion of the customization of new qualifiers for the Dublin Core descriptive metadata representation and a new scheme called "asas" for the representation of the annotations. When inserting the corpus into DSpace it was decided to create a specific item for each of audio clip. It was therefore necessary to set the release wizard offered by DSpace by changing the specific XML file responsible for entry forms (input-forms.xml). The descriptive metadata, identified by researchers, such as title, author, type of song, instrument, etc., and all metadata corresponding to linguistic annotations (phono, morpheme, word, etc..), was associated to each item, together with the original file containing the audio recording and the original file of annotations.

Figure 1: Customization of DSpace metadata's Register.

After the insertion of metadata, the interface was customized by replacing the standard forms provided by DSpace using modules specfically designed to allow the creation of items and the release of DC metadata according to the specific needs of the project. The metadata on the annotations were inserted instead using direct import because the high number of occurrences for each item made it difficult to enter them manually, as shown by Hillman and Westbrooks (2004).

Finally, we proceeded to customize the search interface of DSpace in order to adapt it to new metadata and to the particular needs of the Analytic Sound Archive of Sardinia. In essence, all metadata corresponding to linguistic annotations needed to be indexed in DSpace's search engine so that we could find a certain audio clip even through the search of an associated record. Furthermore, some descriptive metadata such as location, type of performer and contribution were indexed to allow effective searching that exploited the granularity of the metadata.

3.2.1 Metadata Schemas

The metadata are stored and managed by DSpace through a special tool, the Metadata Registry, where the Qualifed Dublin Core schema is configured by default. It can nevertheless be changed, and new customized schemas can be added. The system offers two ways to configure the register: one is the graphic interface named Manakin, and the other can be used by the terminal. Each of them has a specific purpose. The first method allows an authorized user to act on the diagrams through an easy and intuitive web interface.

Once you create a schema, the metadata can be added one at a time, with any qualifiers and related notes. This feature is crucial for the updating and maintenance of the system as it can make adjustments quickly and easily without the intervention of a computer expert. Likewise, you can

choose the second solution where using a specific command, the metadata schema expressed in XML can be imported in the register according to a specific syntax.

During the creation of the Sound Archive, metadata was to be gradually defined and refined by linguists and musicologists, so it was decided to insert and manage it through the DSpace web interface.

We obtained a customized Qualified DC schema with new specific qualifiers and a new schema "asas" with the metadata to use for the insertion of linguistic annotations, as you can see in the Figure 2.

Figure 2: The new schema "asas".

Once we reached the final version of the scheme, however, it was decided to formalize the metadata in XML so that should the system be reinstalled or transferred, the register could be quickly configured via the import metadata command.

3.2.2 Insertion of Metadata in the KMS

DSpace is preset to use, during the phase of the insertion of an item, the Qualified Dublin Core metadata schema but at the same time allows to customize it, or to create new metadata schemas according to the users and their requirements. The first step in order to submit the metadata in DSpace was to organize and structure them as metadata schemas. The second step was, instead, to configure the knowledge management system so that it could be adapted to the particular chosen metadata schema.

The Manakin graphic interface offers a tool to manage the Metadata Register that allows to customize the preconfigured Dublin Core schema in a rapid and intuitive way.

In the Metadata Register can be entered all new qualifiers with their related comments to obtain the desired application profile, which coexist in the Qualified Dublin Core standard with its qualifiers, qualifiers and made for the specific project,

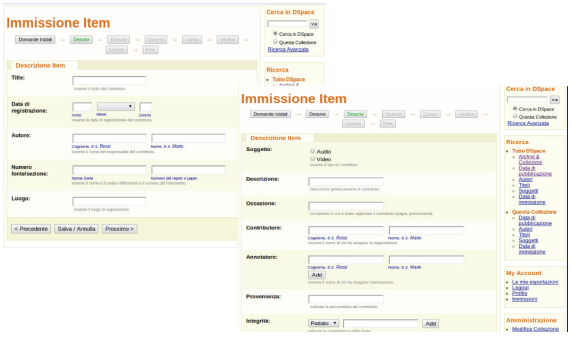

Figure 3: Interface personalization input metadata.

Analytical Sound Archive of Sardinia.

In DSpace this entity is called "item" and is "built" with a wizard that, using the pre-configured modules, allows to specify the values of the meta-information to be formalized as metadata. These modules are ready for the insertion of the metadata belonging to the Qualified Dublin Core schema, but in the case study they were fully customized to fit the new specific qualifiers for the archive.

3.2.3 Research Modules

DSpace offers a powerful search interface that is configured to use the main DC metadata (like title, author, language) or the free full-text research as parameter for the search.

The researchers asked for specific search criteria, besides the standard ones, so you can search for audio clips by place of recording, the participation to a particular event, but also the annotations they contain. The last criterion, in particular, is the most important function for the study of the language corpus as it allows to perform statistical analysis on the text easily and quickly, without the need to use specific and complicated software interrogation.

The search indexes were modified so that all needed metadata were selected as criteria in the search interface and information like place of recording, performers' information, the number indicating the event place and all metadata corresponding to the levels of annotation were specifically added.

4 CONCLUSIONS

In the context of an industrial project that aimed to create the Analytic Sound Archive of Sardinia, we proposed a new approach using DSpace as a specific tool to organize and manage a big quantity of information coming from the audio recording, in

order to formalize knowledge in the Analytical Sound Archive of Sardinia. We have chosen DSpace because after some investigations we found that more Universities and research Institutes (i. e. Brunel University, Cornell University and Massachusetts Institute of Technology) use DSpace, in fact it is a very efficient tool easy to use, customizable and flexible to allow the management, the classification and the storage of a vast amount of knowledge contained in an electronic corpus of Sardinian language, and that could, at the same time, allow a high usability in terms of ease of reference as well as ease of query and communication.

REFERENCES

arXiv (Cornell University Library), http://arxiv.org/

Barton, J., Currier, S., Hey, J. M. N. (2003). Building Quality Assurance into Metadata Creation: an Analysis based on the Learning Objects and e-Prints Communities of Practice. In Sutton, S., Greenberg, J. and Tennis, J., Eds. *Proceedings 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice - Metadata Research and Applications*, Seattle, Washington (USA).

Becker, H., Chapman, A., Daviel, A., Kaye, K., Larsgaard, M., Miller, Nebert, D., Prout, A. and Wolf, M. P. (1997). *Dublin Core element: Coverage*. http://www.alexandria.ucsb.edu/publicdocuments/metadata/dc_coverage.html

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003). Retrieved from: http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/

Brunel University Research Archive (BURA), Brunel University London, UK, http://dspace.brunel.ac.uk/

Chopey, M. A. (2005). Planning and Implementing a Metadata-Driven Digital Repository. Haworth Press Inc.

Cornell University Library (Cornell University, USA), http://ecommons.library.cornell.edu/index.jsp

DSpace, http://www.dspace.org/

DSpace@MIT, MIT's institutional repository (Massachusetts Institute of Technology, USA), http://dspace.mit.edu/

Dublin Core Metadata Initiative (DCMI).

Gartner, R. (2008). Metadata for Digital Libraries: state of the art and future directions. *JISC Technology and Standards Watch Reports*. Retrieved from: http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf JISC.

Heery, R. and Patel, M. (2000). Application profiles: mixing and matching metadata schemas. Ariadne. http://www.ariadne.ac.uk/issue25/app-profiles/

Hillmann, D. I. (2005). Using Dublin Core. Dublin Core Metadata Initiative Recommendation. Retrieved from: http://dublincore.org/documents/usageguide/

Hillman, D. I. and Westbrooks, E. L. (2004). Metadata in practice. *American Library Association*.

Hutt, A. and Riley, J. (2005). Semantics and Syntax of Dublin Core Usage in Open Archives Initiative Data. *Joint Conference on Digital Libraries*.

Jackson, A. S., Han, M. J., Groetsch, K. and Mustafoff, M. (2008). Dublin Core Metadata Harvested Through OAI-PMH. *In Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*.

Lagoze, C. and Van de Sompel, H. (2003). The making of the Open Archives Initiative protocol for metadata harvesting. *Library Hi Tech*.

Lunesu, M. I., Pani, F. E. and Concas, G. (2011). An approach to manage semantic informations from UGC. *International Conference on Knowledge Engineering and Ontology Development* (KEOD).

Lunesu, M. I., Pani, F. E. and Concas, G. (2011). Using a standards-based approach for a multimedia knowledge-base. *International Conference on Knowledge Management and Information Sharing* (KMIS).

Open Archives Forum: OAI-PMH Tutorial, http://www.oaforum.org/tutorial/

Open Archives Initiative, http://www.openarchives.org/

Solodovnik, I. (2011). Metadata issues in Digital Libraries: key concepts and perspectives. *Italian Journal of Library and Information Science*, Vol. 2, No. 2.