A New Approach for Knowledge Management and Optimization using an Open Source Repository

GIULIO CONCAS, FILIPPO EROS PANI, MARIA ILARIA LUNESU
DIEE, Department of Electric and Electronic Engineering, Agile Group
University of Cagliari
Piazza d'Armi, 09123 Cagliari
ITALY
{concas, filippo.pani, ilaria.lunesu}@diee.unica.it http://agile.diee.unica.it

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 the context of the "Analytic Sound Archive of Sardinia" project, that aims to create an institutional archive with a linguistically annotated electronic corpus, this work proposes a new approach for management of knowledge using the tool DSpace (an open source software package developed in 2000 in the context of a joint project of the Massachusetts Institute of Technology with Hewlett-Packard): the purpose is to offer an original way to associate linguistic annotations (information associated to specific text portions) to the corpus by treating them as metadata, so as to insert and manage them in the archive of choice after formalizing them in XML. The formalization level of this approach allows for effective text retrievals through a metadata schema and easy, quick corpus interrogations, by formalizing linguistic annotation as a structured metadata schema. There is, thus, the need to have an efficient tool that could classify and store the vast amount of knowledge contained in an electronic corpus of spoken texts of Sardinian language, linguistically annotated at various levels, and that could allow a high usability in terms of ease of reference as well as ease of query and communication.

Key-Words: Knowledge Management, Open Archive, Institutional Repository, Multimedia content, Metadata.

1 Introduction

The use of User Generated Content and publishing tools facilitate the circulation of resources on the Internet. This amount of information needs approaches to gather information in an organized, reliable manner, describe it, store it, and retrieve it. All this content needs a minimum level of interoperability for tools and description parameters. In the field of scientific communication, that is what happened with the creation of knowledge management systems like DSpace [1], Eprints [2], etc., which are institutional archives modelled on the Open Access Initiative that allow structuring information and adding standardized metadata to it [3-5]. 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 [6-9].

In this paper we analyze the problem to define the electronic corpus in an IR, formed by a collection of audio recordings from poetry contests and singing performances in Sardinian language, stored and annotated on different linguistic levels.

An electronic corpus is generally a homogeneous collection of written or oral texts in digital format, processed with coherent criteria in order to build an empirical basis for language analysis. Its advantage is that it can be annotated by adding linguistic information in a specific portion of text.

The repository will be compliant with Open Access Initiative (OAI), the corpus will be included in an open IR, being therefore available for Sardinian language scholars and everyone who wishes to use it.

Linguists and musicologists, creators of the corpus, needed to study and research the documents in it, and they asked for the possibility to save their work in an readily available digital archive to store,

index and manage it for both access and communication inside the scientific community.

This needs was to save the information in an readily available digital archive to store, index and manage it for both access and communication inside the scientific community. The purpose of this study is to offer an original way to associate linguistic annotations (information associated to specific text portions) to the corpus by treating them as metadata, using an Open Source repository. In particular, an application profile has been created for the Dublin Core (DC) metadata schema, which is suitable to the nature of the audio recordings in the Analytic Sound Archive of Sardinia (ASAS). In the second section of this paper we recall some aspects about the Knowledge Management. In the third we mention some typical use of Institutional Repositories , in the fourth, we present our proposed approach for knowledge formalization and management, in the fifth the case study about the use of DSpace and the sixth section includes the conclusion and reasoning about the future evolution of the project.

2 Knowledge Management

Organization and availability of contents in KMSs basically depend on two factors: one is whether KMSs have effective tools for information indexing and retrieval; the other is how. The solution to this issue was found in the experience of the library and archive industry, which have been dealing with the issues related to organization and collection of information since way before the digital revolution. This experience suggested using metainformation, i.e. data used to describe and classify information, as a possible solution. The tools used to enter and manage contents on the Internet must allow for entering and retrieving organized and relevant metainformation, as metadata.

2.1 Metadata

Metadata have thus a fundamental role in organizing and managing digital resources, especially when there is a great quantity of available information that must be indexed and catalogued to facilitate search and retrieval [10-12]. The selection of which metadata to use in describing a resource depends on a thorough observation of the characteristics, properties, common features, and differences in the informational environment the source belongs to.

A metadata schema is a set of structured metadata, developed for specific purposes in order to establish a standard of metadata structure and terminology, and to associate different types of metadata. Every metadata schema includes a definite number of elements, called metadata elements, each with its own meaning and purpose, i.e. describing the information resource [6-7]. However, since standardization is the purpose, it is always advisable to use largely used metadata schemas rather than creating new ones. Application profiles are made of metadata sets derived from different schemas, and are aimed to create tools for particular applications while keeping interoperability with the original base schema. This procedure and the application of common rules can make different systems interoperable, like those in libraries, museums and archives, making them able to share a part of common metadata [13-14].

2.2 The Dublin Core Standard

A support to content management is offered by the DC metadata schema, which easily pairs up with other metadata schemas in the OAI architecture, improving granularity and refinement of their structures [15-17]. The rapid spreading of DC as metadata schema was doubtlessly favoured by its remarkable simplicity, thanks to which it could adapt to many kinds of resources and usage environments. It is important, for a semantic model used in resource discovery not to be dependent on the format of the resource it needs to describe.

In the latest years, DC was increasingly used in many fields to describe, organize, manage, resources in possession of institutions and international organizations, and also to support and provide added value services, assuring a base format for aggregation and exchange of metadata collections, such as in the Open Archive Initiative, or as indispensable search tools in portals. The use of a standardized general classification system allows for metadata in such collections to be combined and for knowledge inside each collection to be shared [8-9].

2.3 Linguistic Annotations and Corpus

The linguistics corpus studies great quantities of linguistic productions, either spoken or written, by observing their characteristics: lexicon, syntax, collocations, phonic chain, morphologic structures, etc. A corpus is any complete and orderly collection

of written texts, by one or more authors, on a certain topic, or, linguistically speaking, the sample of a language as examined in the description of the same language. In order to exploit the wealth of information stored in a corpus as linguistic data, the corpus must be enriched with additional information: linguistic annotations, i.e. the adding of linguistic or metalinguistic information to different portions of a text [18].

3 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 [19]. 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.

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.

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.

4 Proposed approach

Our proposed approach for knowledge formalization and management, gathered in an annotated electronic corpus in an IR based on the OAI model, will be described below.

4.1 Formalization of Metadata Schemas

In order to manage and organize the information that makes up the corpus, KMSs associate organized and relevant information to a text when it is entered. Metadata schemas mirror the complex nature of data and are often strongly structured and hierarchical, including many kinds of metadata, with many different functions.

Building an effective system of structured metadata means creating a conceptual model to formalize and model the essential semantic characteristics of a knowledge domain.

After designing the conceptual model of the knowledge domain, a top-down approach can be used for structuring the metadata schema. If the knowledge domain is made of an electronic corpus

and its objects are its texts, essential metadata (author, title, language, publishing date, etc.) must be deducted and formalized from their semantic characteristics. Some of those metadata may be further specified according to a hierarchical structure: for example, the metadata “author” maybe further refined as “main author”, “illustrator” and, "curator".

4.2 Formalization of Linguistic Annotations

The information in the corpus are organized as informal annotations and the most efficient way to use their information is formalizing this annotation through metadata schemas. In this way, not only annotations can be associated to their texts, but they can also be used as search parameters for finding texts.

Linguistic annotations created with special software, like PRAAT [20] for audio files, are generally stored in a semi-structured manner. In fact, each annotation is distinctly represented inside the file, according to a defined, repetitive structure where the annotation texts is paired with the instant or the time interval it refers to. Moreover, the belonging of each annotation to a certain linguistic level is clearly stated in the file. The formalization of annotations in a metadata schema can be achieved using a bottom-up or inductive reasoning. Starting with the analysis of the structure of each annotations in the file and applying inductive logic, a "category" is abstracted from every linguistic level. This formalization allows for easily coding and representing of annotations though markup languages like XML, because their structure can be described with tags or markers, for metadata and their qualifiers, inside which a linguistic label is found. All annotations in the same linguistic level, e.g. phonetics, can be formalized in the XML as different occurrences of the same metadata called "phoneme", whose value can be made up of two terms: linguistic label and eventually time interval.

4.3 Choosing a Metadata Schema

The use of both a deductive and an inductive approach allows metainformation and linguistic annotations to be formalized in a single structured metadata schema. Entering metadata in a knowledge management system requires the selection of an operational criterion based on the particular needs the system has to work with.

Most archives use Qualified Dublin Core as main schema for indexing and displaying metadata and Simple Dublin Core to show them through the OAI-PMH standard. There are four main criteria for choosing a metadata schema, with different approaches in metadata organization: 1) mapping of native metadata on existing DC elements; 2) mapping of native metadata on DC elements and creation of new customized qualifiers for DC elements; 3) creation of a customized metadata schema, identical to the native metadata set; 4) creation of DC metadata records as abstraction of native metadata records and entering of the latter as attachments to the resource. Out of the criteria mentioned above, the first one is the least satisfactory for preservation and reuse of descriptive metadata of resources, while the third one is the most preserving of the integrity and granularity of original metadata but needs great efforts for the creation of a customized metadata schema, together with high maintenance costs for the archive. The second and fourth criteria combine preservation and granularity needs with archive management costs better than the other two. Choosing between them depends solely upon the particular requirements of the archive.

Once the decision on which criterion to use is settled, the archive must be configured so that it is compatible with the approach of choice for metadata management. In particular, if the second criterion is adopted, the DC schema must be updated with new, customized qualifiers; if the third criterion is chosen, the entire metadata schema created ad hoc must be entered into the system. In this way, customized metadata and qualifiers can be used to describe texts of the corpus inside the archive.

Generally, metadata schemas can be configured through the user interface of the archive. However, schemas rich in elements and qualifiers are better configured with the import tools provided by management systems, after having encoded them with the XML markup language. XML is used by archives to manage the import-export of metadata.

Compilation of metadata records associated to texts in the corpus may be usually done with either a user interface or with batch import tools. Instead, when big quantities of metadata need to be associated to one resource, like with linguistic annotations, there are specific batch import tools that require the specification of all metadata as attribute-value pairs, coded in an XML file.

5 Case study

The ASAS project (http://asas.flosslab.it) aims to create an IR with an annotated spoken language electronic corpus that could become a platform for the preservation, study, communication and appreciation of oral traditions of the Sardinian language, especially improvised poetry.

5.1 Annotations through PRAAT

The electronic corpus was annotated by linguists and musicologists through the PRAAT software [19], which, besides performing spoken language analysis, allows for multilevel segmentation and linguistic annotations of audio files. The software has a graphic interface with waveforms and voice spectrum that make annotators' work easier and make visible those acoustic phenomena that can be found by an accurate spectrum analysis, followed by annotation levels. 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.

5.2 Metainformation Associated to Audio Recordings

Musicologists and Linguists, other than with annotations, wanted to complete every audio recording by describing it with a number of information, chosen among the most relevant features of the recordings. The information could be used to manage recordings in the archive, because by describing them they allow for selection and organization, facilitating efficient retrieval and usage. Metainformation range from something closely related to cataloguing, like author, title, object, recording date, etc., up to more technical information like the different singing types, speech types, accompaniment or instruments. Linguists and musicologists selected 38 metainformations associated to audio recordings: title, author, object, description, format, etc.

5.3 Formalization of Semantic Characteristics: Top-Down Approach

After designing the conceptual model of the knowledge domain, a top-down or deductive approach can be used for formalizing the semantic characteristics of texts. Through a continuous dialogue with the scholars, audio recordings were analysed for their essential and basic properties, needed to organize and retrieve texts in the corpus. Twelve general metadata were found: title, author, publisher, object, contributor, date, place, occasion, document accessibility, language, description and format. Those metadata outlined the necessary information to describe spoken texts in the corpus, conveying in particular singing or speech type, the occasion in which the audio was recorded, and the linguistic variety it belongs to.

The top-down approach proceeds to further specialize the metadata. More specific, or qualified, metadata are represented by adding a qualifier to the name of the more general metadata and using the common syntax metadata.qualifier.

Lastly, "relational" metadata are defined as well, in order to define a certain relation among two or more different objects belonging to the corpus. An inclusion relation must be specified in order to describe the belonging of one or more objects to the same recording set, for example different songs in a singing contest.

5.4 Formalization of Linguistic Annotations: Bottom-Up Approach

The formalization of annotations in a metadata schema can be achieved using a bottom-up or inductive reasoning, as explained in the previous section. The structure of annotations is analysed with the PRAAT software. Annotations are organized with a precise structure: each annotation is made of a time interval and a text label or by an instant and a marker with its text. All annotations in the same linguistic category are collected in the same tier (or annotation level), which can be considered as the category they belong to, giving its name to the corresponding metadata. In this way, a repeatable metadata is found in each annotation level of the TextGrid (the text file where PRAAT stores all Tier with their own segmentations and annotations) and each annotation can be represented as multiple occurrences of that metadata.

5.5 Choosing a Metadata Schema for KMS Entering

Depending on the interoperability needs that must be met, importing the metadata schema that was just created into the knowledge management system

may not be appropriate or convenient. Most archives use Qualified DC as main schema for indexing and displaying metadata and Simple DC to show them through the OAI-PMH standard. Therefore, the adoption of Dublin Core must be thoroughly evaluated when an archive is needed to be compliant with the interoperability principles required by OAI. Our of the four criteria listed in section 4.3, the most suitable technique for the case study is an hybrid model between the second (mapping of native metadata on DC elements and creation of new customized qualifiers for DC elements) and the third one (creation of a customized metadata schema, identical to the native metadata set). The third criterion is more convenient for linguistic annotations, so that a dedicated metadata schema can be created to preserve their granularity; while the second criterion is best suited for all other metadata, because it combines the advantages of granularity as provided by qualifiers to interoperability provided by DC metadata.

5.6 Application Profile for the Analytical Sound Archive of Sardinia

In creating a specific application profile for the ASAS, a "conservative" approach was used towards the original Qualified DC elements and qualifiers in order to use as many of them as possible for the formalization of descriptive and relational metadata. A special schema, identified by the prefix "asas", was created instead for annotations. Its metadata were entered into the DC application profile as outlined below.

Metainformation or ASAS Annotation	DC Application Profile Metadata
Title	dc.title
Author	dc.creator
Publisher	dc.publisher
Object	dc.type
Description	dc.type.category
Contributor	dc.contributor
Annotator	dc.contributor.annotatore
Location	dc.coverage.spatial
Date	dc.date.created
Occasion	dc.subject
Source	dc.relation.isbasedon
Document Accessibility	dc.rights
Performer	dc.contributor.sperakerPerformer
Performer's Age	dc.description.speakerPerformer
Performer's Place of Origin	dc.description.speakerPerformer
Language	dc.language
Source Completeness	dc.description.integrità
Source No.	dc.relation.ispartofseries
Source Section No.	dc.relation.ispartofseries
Document Type	dc.format.audioVideo
Format	dc.format.medium
Acquisition Method	dc.format.modoAcquisizione
Reading Type	dc.type.lettura
Interview Type	dc.type.intervista
Monody Type	dc.type.monodia
Unison / Heterophony	dc.type.unisonoEterofonia
Accompaniment Type	dc.type.monodiaAccompagnamento
Polyphony Type	dc.type.polifonia
Instrumental	dc.type.strumentale
Instrument	dc.type.strumento
Singing Type	dc.type.tipoCanto
Other	dc.description
Syllable	asas.annotazione.sillaba
Tone	asas.annotazione.toni
Morpheme	asas.annotazione.morfema
Phone	asas.annotazione.fono
Word	asas.annotazione.parola
Part of Speech	asas.annotazione.pos
Syntagm	asas.annotazione.sintagma
Sentence	asas.annotazione.frase
Information Structure	asas.annotazione.strutturaInformativa
TurnPerf	asas.annotazione.turnPerf
Musical Syllable	asas.annotazione.sillabaMusicale
Metric Segment	asas.annotazione.segmentoMetrico
Musical Segment	asas.annotazione.segmentoMusicale
Tonal Centre	asas.annotazione.centroTonale
Notation	asas.annotazione.notazione
Ornamentation	asas.annotazione.ornamentazione
Accents	asas.annotazione.accenti
Melismatic Syllable	asas.annotazione.sillabaMelismatica
ADD1	asas.annotazione.annotazioneLibera

Table 1: Application profile for the ASAS

The next step is to enter metadata in the knowledge management system: once metainformation have been organized and structured, the KMS is configured so that it can be adapted to the selected metadata schema.

5.7 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 [1]. 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 [21].

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.

5.7.1 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.

5.7.2 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.

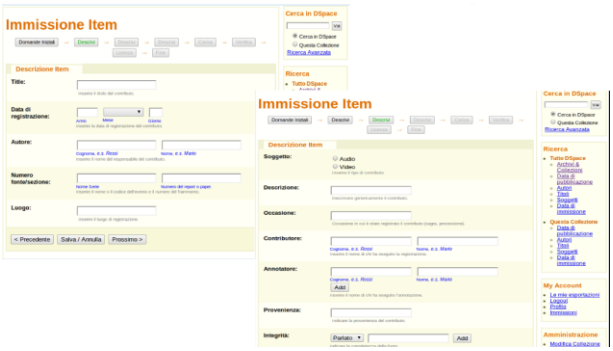

Figure 3: Interface personalization input metadata

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, 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.

5.7.3 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.

6 CONCLUSIONS

The purpose of this work was to offer a new approach to formalization and management of knowledge represented by a set of audio recordings belonging to a corpus plus the linguistic information added to the same corpus with annotations. The approach was applied to formalize knowledge in the ASAS, a joint project by linguists and musicologists at University of Cagliari. The project aimed to present a study on improvised poetry in Sardinian language, using an electronic corpus they created and annotated. In order to make the resources openly accessible through the Internet, as per our aim, we used DSpace as a specific tool to organize and manage a big quantity of information coming from the audio recording. We have chosen this tool 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. The formalization of a structured metadata schema was reached through the creation of an application profile for the Qualified Dublin Core metadata schema, where customized qualifiers were added to the standard elements and qualifiers. Metadata in non-standard schemas could then be better represented. Linguistic annotations were formalized as well through a metadata schema. Corpus interrogation was thus made easier and quicker, since it used the knowledge management system's search tool.

This work leaves space for future research on ways to improve the service. A dedicated website or the integration of this system in an institutional portal through an exploration interface would be particularly interesting. Another feature that could be implemented may be a virtual map where recordings can be explored by geographic location.

In future works we want to use the same approach for knowledge formalization and management as the one represented in a set of scientific papers [22-36] on different topics of Software Engineering. Knowledge in papers is usually limited to keyword management; however, we think this approach could be effectively used for other types of knowledge as well.

References:

[1] DSpace, http://www.dspace.org
[2] EPrints, http://www.eprints.org
[3] Tansley R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G. and Smith, M., The DSpace Institutional Digital Repository System: Current Functionality, *JCDL '03 Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, 2003.
[4] Lynch, C., *Institutional repositories: essential infrastructure for scholarship in the digital age*, Association of Research Libraries: a bimonthly report, no. 226, 2003.
[5] Swan, A. and Carr, L., *Institutions, their repositories and the Web*. Serials Review, 2008, p. 31. ttp://eprints.ecs.soton.ac.uk/14965
[6] Heery, R. and Patel, M., *Application profiles: mixing and matching metadata schemas*, Ariadne, 2000.
[7] Lagoze, C. and Van de Sompel, H., *The making of the Open Archives Initiative protocol for metadata harvesting*, Library Hi Tech, 2003.
[8] Lunesu, M. I., Pani, F. E. and Concas, G., An approach to manage semantic informations from UGC, *International Conference on Knowledge Engineering and Ontology Development (KEOD)*, 2011.
[9] Lunesu, M. I., Pani, F. E. and Concas, G., Using a standards-based approach for a multimedia knowledge-base, *International Conference on Knowledge Management and Information Sharing (KMIS)*, 2011.
[10] Chopey, M. A., *Planning and Implementing a Metadata-Driven Digital Repository*, Haworth Press Inc., 2005.
[11] Dunsire, G., *Collecting metadata from institutional repositories*, OCLC Systems & Services, Vol. 24, No. 1, 2008, pp. 51-58.
[12] Solodovnik, I., Metadata issues in Digital Libraries: key concepts and perspectives, *Italian Journal of Library and Information Science*, Vol. 2, No. 2, 2011.
[13] Pani, F. E., Lunesu, M. I., and Concas, G., Optimization of Knowledge Availability in an Institutional Repository, *International Conference on Knowledge Engineering and Ontology Development (KEOD)*, 2012.

[14] Pani, F. E., Lunesu, M. I., and Concas, G., Knowledge Formalization and Management in KMS, *International Conference on Knowledge Management and Information Sharing (KMIS)*, 2012.

[15] Hutt, A. and Riley, J., Semantics and Syntax of Dublin Core Usage in Open Archives Initiative Data, *Joint Conference on Digital Libraries, ACM Press*, 2005.

[16] 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.

[17] Hillmann, D. I., *Using Dublin Core*, Dublin Core Metadata Initiative Recommendation, 2005.

[18] Llisterri, J., *Text Corpora Working Group Reading Guide*, EAGLES (Expert Advisory Group on language Engineering Standards) Document EAG-TCWG-FR-2, CNR, Istituto di Linguistica computazionale, 1996.

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

[20] PRAAT, http://www.fon.hum.uva.nl/praat/

[21] Hillman, D. I. and Westbrooks, E. L., *Metadata in practice*, American Library Association, 2004.

[22] Concas, G., Marchesi, M., Murgia, A., Pinna, S., Tonelli, R., Assessing traditional and new metrics for object-oriented systems. *In Proceedings of the 2010 ICSE, Workshop on Emerging Trends in Software Metrics (WETSoM '10)*, 2010, pp. 24-31.

[23] Melis, M., Turnu, I., Cau, A., Concas, G. Evaluating the impact of test-first programming and pair programming through software process simulation. *In Software Process Improvement and Practice* 11 (4), 2006, pp. 345-360.

[24] Concas, G., Marchesi, M., Murgia, A., Tonelli, R., An empirical study of social networks metrics in object oriented software. *Advances in Software Engineering*, Vol. 2010 (ID: 729826).

[25] Concas, G., Marchesi, M., Murgia, A., Tonelli, R., Turnu, I., On the distribution of bugs in the Eclipse system, *IEEE Transactions on Software Engineering* 37 (6), art. no. 5928349, 2011, pp. 872-877.

[26] Turnu, I., Concas, G., Marchesi, M., Pinna, S., Tonelli, R., A modified Yule process to model the evolution of some object-oriented system properties, *Information Sciences* 181 (4) ,2011, pp. 883-902.

[27] Locci, M., Concas, G., Tonelli, R., Turnu, I., Three algorithms for analyzing fractal software networks, *WSEAS Transactions on Information Science and Applications* 7 (3) , 2010, pp. 371-380.

[28] Tonelli, R., Concas, G., Locci, M., Three efficient algorithms for implementing the preferential attachment mechanism in Yule-Simon Stochastic Process, *WSEAS Transactions on Information Science and Applications* 7 (2), 2010, pp. 176-185.

[29] Concas, G., Lisci, M., Pinna, S., Porruvecchio, G., Uras, S., Open source communities as social networks: An analysis of some peculiar characteristics, *Proceedings of the Australian Software Engineering Conference (ASWEC)*, art. no. 4483227, 2008, pp. 387-391.

[30] Turnu, I., Melis, M., Cau, A., Setzu, A., Concas, G., Mannaro, K., Modeling and simulation of open source development using an agile practice, *Journal of Systems Architecture* 52 (11), 2006, pp. 610-618.

[31] Destefanis, G., Tonelli, R., Concas, G., Marchesi, M., An analysis of anti-micro-patterns effects on fault-proneness in large Java systems, *ACM Symposium on Applied Computing*, 2012, pp. 1251-1253.

[32] Murgia, A., Tonelli, R., Marchesi, M., Concas, G., Counsell, S., McFall, J., Swift, S., Refactoring and its relationship with fan-in and fan-out: An empirical study, *The European Conference on Software Maintenance and Reengineering (CSMR)*, art. no. 6178854, 2012, pp. 63-72.

[33] Concas, G., Marchesi, M., Murgia, A., Pinna, S., Tonelli, R Assessing traditional and new metrics for object-oriented systems. *International Conference on Software Engineering*, 2010, pp. 24-31.

[34] Turnu, I., Marchesi, M., Tonelli, R. Entropy of the degree distribution and object-oriented software quality, *3rd International Workshop on Emerging Trends in Software Metrics (WETSoM)*, art. no. 6226997, 2012, pp. 77-82

[35] Concas, G., Marchesi, M., Murgia, A., Tonelli, R., Turnu, I., On the distribution of bugs in the Eclipse system, *IEEE Transactions on Software Engineering* 37 (6), art. no. 5928349, 2011, pp. 872-877.

[36] Concas, G., Marchesi, M., Destefanis, G., Tonelli, R. An empirical study of software metrics for assessing the phases of an agile project, *International Journal of Software Engineering and Knowledge Engineering* 22, 2012, pp.525-54.