Knowledge Formalization and Management in KMS

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, Metadata.

Abstract: Organization and availability of contents in Knowledge Management System (KMS) basically depend on two factors: one is that KMS have effective tools for information indexing and retrieval; the other is how the tools are actually understood and used by users. This work proposes a new approach for formalization and management of knowledge, in this case a group of audio recordings in a corpus and linguistic information added to that corpus with annotations. 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. The proposed approach was experimented upon and validated during a project that aimed to create the Analytical Sound Archive of Sardinia. The archive has an electronic corpus of spoken texts, linguistically annotated at various levels.

1 INTRODUCTION

Archive and publishing tools that facilitate the circulation of resources on the Internet must be able to gather information in an organized, reliable manner, describe it, store it, and retrieve it, all with 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, Eprints, etc., which are institutional archives modelled on the Open Access Initiative that allow structuring information and adding standardized metadata to it (Tansley, 2003) (Linch 2003) (Swan and Carr, 2008).

In this context, the "Analytic Sound Archive of Sardinia" project aims to create an institutional archive with a linguistically annotated electronic corpus. 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 electronic corpus in the studied Institutional Repositories (IR) will be formed by a collection of audio recordings from poetry contests and singing performances in Sardinian language, stored and annotated on different linguistic levels. The purpose of the project is the preservation, appreciation and knowledge of Sardinian oral traditions, especially improvised poetry.

In accordance 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.

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, so as to insert and manage them in the archive of choice after formalizing them in XML, the universally used markup language for representing metainformation.

In particular, an application profile was created for the Dublin Core metadata schema, which is suitable to the nature of the audio recordings in the Analytic Sound Archive of Sardinia.

In the second section of this paper we recall some aspects about the Knowledge Management. In the third and fourth, we present our proposed approach for knowledge formalization and management, and the case study. The fifth section includes the conclusion and reasoning about the future evolution of the project.

Pani F., Lunesu M., Concas G., Stara C. and Tilocca M..
Knowledge Formalization and Management in KMS.
DOI: 10.5220/0004126001320138
In *Proceedings of the International Conference on Knowledge Management and Information Sharing* (KMIS-2012), pages 132-138
ISBN: 978-989-8565-31-0

2 KNOWLEDGE MANAGEMENT

Organization and availability of contents in KMSs basically depend on two factors: one is whether KMS systems have effective tools for information indexing and retrieval; the other is how those tools are actually understood and used by users.

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, as shown by Hillman and Westbrooks (2004), Strintzis, Bloehdom, Handschuh et al. (2004), Chopey (2005), Dunsire (2008), Solodovnik (2011).

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, as shown by Heery and Patel (2000), and by Lagoze and Van de Sompel (2003).

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.

2.2 The Dublin Core Standard

A support to content management is offered by the Dublin Core metadata schema, which easily pairs up with other metadata schemas in the OAI architecture, improving granularity and refinement of their structures (Hutt and Riley, 2005).

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 (Hillman 2005) (Jackson, Han, Groetsch and Mustafoff, 2008). 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, as proven by Lunesu, Pani and Concas (2011).

2.3 Linguistic Annotations and Corpus

The so-called corpus linguistics studies great quantities of linguistic productions, either spoken or written, by observing their characteristics: lexicon, syntax, collocations, phonic chain, morphologic structures, etc. Computational linguistics, in order to aid this study, developed the first automated or semi-automated text analysis information tools, avoiding manual analysis and data research.

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, as shown by Llisterri (1996) and in the EAGLES Project.

3 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.

3.1 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", "curator", etc.

3.2 Formalization of Linguistic Annotations

The need to interrogate the corpus once entered in the KMS makes it necessary to formalize annotations in a way that permits the extraction of linguistic information without using other software agents, whose syntax may be obscure and complicated.

Since KMS are based on metadata for the organization and collection of resources, the most efficient way to use their information is formalizing them 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 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.

3.3 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.

3.4 Update of Knowledge Management System

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.

4 CASE STUDY

The "Analytic Sound Archive of Sardinia" 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.

The approach described in the previous section was applied to knowledge formalization and management, gathered in an annotated electronic corpus, in a IR based on the OAI model.

4.1 Annotations through PRAAT

The electronic corpus was annotated by linguists and musicologists through the PRAAT software, 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.

4.2 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 metainformation associated to audio recordings: title, author, object, description, performer, language, format, etc.

4.3 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.

All descriptive metainformation were analysed and formalized in a "basic" structured metadata schema.

4.4 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. All annotations are thus formalized in a structured metadata schema.

4.5 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. It could be necessary instead to map it, partially or totally, on another schema.

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. 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 3.3, 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.

4.6 Application Profile for the Analytical Sound Archive of Sardinia

In creating a specific application profile for the Analytical Sound Archive of Sardinia, 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 (customized qualifiers are in italics).

Table 1: Application profile for the Sound Archive of Sardinia.

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*

Table 1: Application profile for the Sound Archive of Sardinia(cont.).

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*

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

5 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 Analytical Sound Archive of Sardinia, 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 entered the annotated corpus in a knowledge management system, compatible with OAI standards and protocols for metadata sharing and knowledge circulation.

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.

REFERENCES

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

DSpace, http://www.dspace.org

Dunsire, G. (2008). *Collecting metadata from institutional repositories*. OCLC Systems & Services, Vol. 24, No. 1, pp. 51-58.

EPrints, http://www.eprints.org

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 LibrariesACM Press.

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.

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

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).

Lynch, C. (2003). *Institutional repositories: essential infrastructure for scholarship in the digital age*. Association of Research Libraries: a bimonthly report, no. 226.

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

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

Strintzis, J., Bloehdom, S., Handschuh, S., Staab, S., Simou, N., Tzouvatras, V., Petridis, K., Kompatsiaris, I. and Avrithis, Y. (2004). *Knowledge representation for semantic multimedia content analysis and reasoning*. In Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media technology .

Swan, A. and Carr, L. (2008). *Institutions, their repositories and the Web*. Serials Review, 34/1 (2008), p. 31, http://eprints.ecs.soton.ac.uk/14965.

Tansley R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G. and Smith, M. (2003). *The DSpace Institutional Digital Repository System: Current Functionality*. JCDL '03 Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries.

The OAI Executive (2008). *The Open Archives Initiative Protocol for Metadata Harvesting*. Document Version 2008-12-07T20:42:00Z. Retrieved from: http://www. openarchives.org/OAI/openarchivesprotocol.html