Applying Semantic Web Technologies for Enriching Master Classes

Paloma de Juan
Depto. Ing. Sistemas Telemáticos
Univ.Politécnica de Madrid
paloko@gsi.dit.upm.es

José C. González
Depto. Ing. Sistemas Telemáticos
Univ. Politécnica de Madrid
jcg@gsi.dit.upm.es

Carlos A. Iglesias
Germinus XXI (Grupo Gesfor)
cif@germinus.com

Abstract

This article presents how semantic web technologies have been applied for enriching existing contents within the SEMUSICI project. The SEMUSICI project has the goal of researching on how semantic web technologies can be applied to digital libraries, and how this can improve searchability and accessibility. The project takes the results from the eContent project HARMOS, which defined a musical taxonomy for cataloguing master classes, and proposes a methodology for evolving this taxonomy into an ontology, and migrating the contents accordingly.

1. Introduction

Cataloguing standards, such as MODS [1], MARC [13] or Dublin Core [14] define metadata following a flat property value orientation, which provides textual search capabilities. In some contexts, such as the musical digital libraries, this approach is too narrow, since some of the metadata are entities themselves, such as Compositions of Composers. In the Harmos project (section 2) an object oriented taxonomy was defined, where some of the values, such as compositions, movements or composers were modeled as entity objects, and an advanced search system based on these properties was developed and is available at [17]. This article presents an evolution of this approach, where semantic technology is used for modeling the relationships of the domain model. The main advantage of this approach is its powerful retrieval and inferential capabilities.

The rest of the article is organized as follows. Section 2 and 3 give an overview of the projects Harmos and Semusici, respectively, which constitute the context of this research. Section 4 describes a generic methodology for transforming taxonomies into ontologies. This is the main contribution of SEMUSICI project for content enrichment. Finally, section 7 draws out the main conclusions of the article and the future work.

2. The HARMOS project

The European eContent HARMOS project [16] had the aim of providing access through Internet to videos of master classes from big maestros. HARMOS has produced a collection of audiovisual contents that belong to the musical heritage, where education was the principal focus and the project's main objective.

Harmos defined a pedagogical taxonomy [15] which aims to cover the whole spectrum of musical practice and teaching, focusing on pedagogical aspects. The potential semantic descriptors of this taxonomy where structured around three main concepts, the music, the musician and the musical expression, con more than 400 descriptors as detailed at [15] and more than 700 audiovisual hours of recorded master classes have been catalogued according to this taxonomy.

3. The SEMUSICI Project

The SEMUSICI project [18] aims to evolve the results of the Harmos project by introducing semantic web technologies. The Harmos system provides several retrieval facilities, which allow finding a master based on the previous selections of the user, such as a composer, a composition, a movement, a teacher that has explained this composition, etc. as described in [15]. The introduction of new retrieval possibilities required to extend the database model and huge investment in development the new consults, which should be tuned and optimised given the big volume of the database. The usage of semantic web technology, which allows an easy extension of properties and relationships with new predicates, is expected to make this feasible. In addition, semantic web technology can contribute to improve the quality of metadata, since semantic web technologies can help in checking the consistency of the cataloguing.

The inclusion of semantic web technology points out several challenges. Firstly, it is needed to define an ontology that contains the concepts of the Harmos taxonomy. Secondly, since musical analysts should not be aware of the usage of semantic web technology for cataloguing, it is needed to develop easy interfaces in order to catalogue semantically. Thirdly, it is needed migrating all the Harmos catalogued multimedia collection to the new semantic schema. Finally, it is needed to evaluate the current status of semantic web technology in terms of throughput and performance, given the size of the multimedia collection. This article will cover the first objective.

4. Methodology for transforming taxonomies into ontologies

The central aim of Semusici is to provide a semantic structure that fits the former concepts taxonomy. The purpose of this new approach is to gather information about the relationships between disjoint leaves and build a new representation of both the concepts and these relationships. This leads to a richer representation of the knowledge that is really associated to the digital audiovisual items. This implies a deep understanding of the subject domain. There is no definite methodology for this task, but a general process together with best practices has been proposed.

Once the problem has been well defined and all the requirements have been identified, a suitable structure has to be chosen. As we are looking to take advantage of the technologies of the semantic web, the most appropriate structure is an ontology. The reason why we have chosen this structure is that it provides a formal way to represent roles and their corresponding relations in a specific domain. By placing a concept in such a structure, we are stating that it has certain properties and satisfies some restrictions about his meaning. In other words, each leaf of an ontology represents the definition of a certain resource.

The main difference between an ontology and a taxonomy is the kind of structure in which each of them is based. A taxonomy can be represented as a tree where each leaf is a class. No connections are allowed between disjoint branches. Relations between classes can only be established between a concept and its direct children. So an instance of a certain class can be defined as ”a kind of” its parent class. An ontology is a graph in which richer definitions can be expressed through a more extensive set of relations. This means that any class can be defined in terms of any other resource that is connected to it, not necessarily being its parent or child. Therefore ontologies can store more semantic information than taxonomies, allowing us to infer undeclared knowledge by studying the relations and restrictions of a certain class.

5.1. First step: choosing the appropriate tools

There is a wide variety of tools available to create, edit, browse and store ontologies. There are also many inference engines or reasoners, which are very important to obtain knowledge from the ontology. Several tools have been examined in order to choose the most suitable framework for our purposes. Some of these were Protégé [2], RacerPro [3], Sesame [4], SWOOP [5], WebODE [6], etc. A survey was carried out in order to find distinctive features. Therefore eleven parameters were chosen and thirteen tools were evaluated according to these key features. Some of these parameters were the supported languages, consistency check support, availability, maintenance, etc. As a result, Protégé and Sesame were chosen.

All these tools support a number of languages. Choosing the right language to implement an ontology is probably the most important step in the process. This depends on how thorough the ontology is intended to be. For Semusici, our initial choice was RDFS as it is the main language in Sesame. It proved to be complete enough to allow the building of a basic version of the ontology. Later we decided to include some restrictions to enforce the definition of the elements that we have already defined. These restrictions were also intended to help us perform consistency checks when adding new contents. For that purpose, new OWL statements were added.

5.2. Semusici knowledge base

There are two distinct parts in the knowledge base that is to be represented by the ontology. One is intended to capture all the information that is not directly related to the collection and can be useful to locate a recording. The aim of this is to answer any query that is not directly related to the contents of the recording itself. For instance, “give me all the recordings related to composers born in the 18th century”.

The other part of the knowledge base is the concepts taxonomy. The features of this structure have already been discussed. This taxonomy contains over 200 pedagogical concepts that are used as tags to describe the recordings. In the process of cataloguing the content, these recordings are to be labelled according to semantic descriptors that are part of this taxonomy.

The semantic descriptors were defined according to a tree diagram of concepts. This was based on three large branches that served as a starting point: the musician, music and musical expression. Each one of the divisions that structured the tree diagram of concepts was joined to one of these large branches. The smaller branches were then organized according to a series of categories, reaching, in the end, a didactic concept.

5.3. Building the ontology

This first ontology had to be built from scratch, as most of the concepts it should represent were new. Following a methodology is strongly recommended for this task. The goal of using a methodology is trying not to miss information in the process of transferring knowledge between the different actors that take part in the process. It also provides a set of steps to follow in order to avoid inconsistency, which would lead to undesirable rework. The quality of the ontology will be strongly affected by the choice of an appropriate methodology [7].

There is no single generic ontology-design methodology [8] that covers all the kinds of applications. This means that there is no standard way to build an ontology [9] neither a standard mechanism to evaluate a methodology. However, all published methodologies have proven to be useful, as they all have been applied to some process at least once. The key to finding the best guidelines for a certain application is to analyze the purpose those methodologies were used for and find similarities between that purpose and our application. This could be viewed as a way of reusing knowledge. Reuse is a very common practice in ontology engineering.

There are some steps that are common to almost every methodology. The first step is to identify the purpose and scope of the ontology. Both of them have already been mentioned in this document. Next, one must find out which questions is the ontology supposed to answer. These are called competency questions [10]. We gathered a list of over 50 questions and identified keywords that later would become part of the terminology of the ontology.

Next step was to decide which ones of these keywords should be represented as classes, attributes and instances. The most important thing to consider at this point is how specific we want our ontology to be. Thus we chose those concepts which we found they need a precise definition and separated them from those which constituted the most specific level of the ontology. We also considered reusing some published ontology but finally decided to define our own vocabulary.

5.4. From the concepts taxonomy to an ontology

The first step to turn this taxonomy into an ontology was to create a root class called *Concept*. Every instance of this class is assigned a concept name. This name is the same as the corresponding tag used to classify the digital recordings. Although the original taxonomy was divided into three main branches, we decided to create a first level of more specific classes. We intended to group concepts that had basic semantic features in common in order to make it easy to define relations between different classes.

The original classification grouped most of the concepts according to the instrument they were referred to. For instance, every technique that is related to a string instrument is placed in the subcategory Strings technique, child of Strings. We decided to create a main category, called Technique, to group all the specific technique related concepts, given that we can not consider that a technique “is-a” String. Thus we could establish that every instance of a subclass of Technique should be related to some type of instrument.

We followed these same criteria to create the main categories and build the first level of our ontology. We also defined some properties, such as “relatedTo”, “partOf” and “elementOf”. The first one was defined as a symmetric property and was meant to connect concepts that could be interesting to the same users. For instance, if a user searches for a lesson about hammers, he will probably be interested in videos about keyboards too.

Both “partOf” and “elementOf” are transitive properties. This means that if a first concept is part/element of a second one and this one is part/element of a third one, we can state that the first concept is also part/element of the last one. The difference between them is that if concept A is part of concept B, every instance of B has A (i.e. the frog is part of the bow, because every bow has a part called frog). However, if concept A is element of concept B, that means that only some instances of B have A (i.e. the reed is element of the embouchure, because there are wind instruments that have no reed in their embouchure). Considering this difference, we can state that if concept A is part of concept B and this concept is related to concept C, A is related to C. This is not true if A is element of B though.

Some of these semantic relations were established between concepts that stood under disjoint classes, in order to help the system make future recommendations. Finally, we also used restrictions to enforce the definition of the classes to make it easy to preserve the consistency of the ontology when expanded. Most of the important decisions were taken as a consequence of a thorough analysis of the distribution of the concepts. This analysis led to follow a bottom-up strategy, in order to find the most natural way of classifying the elements of the original taxonomy.

6. Conclusions and future work

As a result of the conceptualization of the subject domain, a list of classes and properties was elaborated. The formalization was carried out using Protégé. This tool provides all the means to code the ontology and visualize some of its elements. The resulting ontology was tested with Sesame. A set of custom rule was arranged in order to support some OWL reasoning.

Almost 1,500 statements were generated as a result of the codification. This is only the ontology, as the knowledge base has not been integrated yet. This includes more than 150 classes and almost 50 properties.

We are currently working on interlinking our ontology with some other data sources in order to improve searching. We would like to incorporate information from the CIA Factbook [11] to perform geographical reasoning. We would also like to add biographical information about the composers. We are testing some datasets from DBpedia [12].

Our second line of work is that of developing a consistency check system. Our purpose is to provide our ontology with a means to preserve consistency and coherence in case that there are several annotators working on the same dataset.

7. References

[1] MODS (Metadata Object Description Schema). Available at

[2] *The Protégé Ontology Editor and Knowledge Acquisition System*, http://protege.stanford.edu.

[3] *Racer Systems GmbH & Co. KG*, http://www.racer-systems.com.

[4] *Sesame: RDF Schema Querying and Storage*, http://www.openrdf.org.

[5] *SWOOP - Hypermedia-based OWL Ontology Browser and Editor*, http://www.mindswap.org/2004/SWOOP.

[6] *WebODE Ontology Engineering Platform*, http://webode.dia.fi.upm.es/WebODEWeb/index.html.

[7] S. Hakkarainen, D. Strasunskas, L. Hella, and S. Tuxen, "Choosing appropriate method guidelines for web-ontology building," in *Proceedings of the 24th Conference on Conceptual Modelling (ER 2005), ser. LNCS 3716*. Springer-Verlag, November 2005, pp. 270–287.

[8] N. F. Noy, and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology", *Tech. Rep. SMI-2001-0880*, Stanford University School of Medicine, 2001.

[9] M. Uschold and M. Grüninger, "Ontologies: principles, methods, and applications", *Knowledge Engineering Review*, vol. 11, no. 2, pp. 93–155, 1996.

[10] M. Grüninger and M. S. Fox, "Methodology for the design and evaluation of ontologies," in *IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing*, April 13, 1995.

[11] *CIA Factbook*, http://www4.wiwiss.fu-berlin.de/factbook

[12] *DBpedia. Querying Wikipedia like a Database*. http://wiki.dbpedia.org

[13] MARC Standards by the Library of Congress, available at http://www.loc.gov/marc/.

[14] Dublin Core Metadata Initiative, available at http://dublincore.org/

[15] C. Iglesias, M. Sánchez et al. A Multilingual Web based Educational System for Professional Musicians, M-ICTE 06, Current Developments in Assisted Education, 2006.

[16] Harmos project official web site available at http://www.harmosproject.com

[17] Magister Musicae web site, available at http://www.magistermusicae.com

[18] The SEMUSICI project official web site, available at http://semusici.germinus.com

Table of Contents

AXMEDIS 2007: proceedings of the 3rd International Conference on Automated Production of Cross Media Content for Multi-channel Distribution
edited by Jaime Delgado, Kia Ng, Paolo Nesi, Pierfrancesco Bellini
ISBN 13: 978-88-8453-678-5