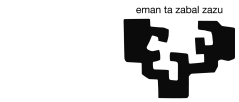
Technical Report
EHU-KZAA-TR-2012-01



UNIVERSITY OF THE BASQUE COUNTRY
Department of Computer Science and Artificial Intelligence

Ontologies for Representation of Folk Song Metadata

Izaro Goienetxea
Iñaki Arrieta
Jon Bagüés
Arantza Cuesta
Pello Leiñena
Darrell Conklin

June 2012

San Sebastian, Spain
www.ccia-kzaa.ehu.es
https://addi.ehu.es/handle/10810/4562

ONTOLOGIES FOR REPRESENTATION OF FOLK SONG METADATA

Izaro Goienetxea[1] **Iñaki Arrieta**[2] **Jon Bagüés**[3]
Arantza Cuesta[2] **Pello Leiñena**[3] **Darrell Conklin**[1,4]
[1]Department of Computer Science and Artificial Intelligence,
University of the Basque Country UPV/EHU, San Sebastián, Spain
[2]Euskomedia Fundazioa, Usurbil, Spain
[3]Eresbil - Basque Music Archive, Renteria, Spain
[4]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

ABSTRACT

The digital management of collections in museums, archives, libraries and galleries is an increasingly important part of cultural heritage studies. This paper describes a representation for folk song metadata, based on the Web Ontology Language (OWL) implementation of the CIDOC Conceptual Reference Model. The OWL representation facilitates encoding and reasoning over a genre ontology, while the CIDOC model enables a representation of complex spatial containment and proximity relations among geographic regions. It is shown how complex queries of folk song metadata, relying on inference and not only retrieval, can be expressed in OWL and solved using a description logic reasoner.

Keywords: ontology, semantic web, Basque folk songs, CIDOC, conceptual reference model

1. INTRODUCTION

The field of cultural heritage studies has emerged in recent years, and the digital management of collections in museums, archives, libraries and galleries has become important. Digital representation of physical objects is increasing, and there is political pressure on memory institutions to make their catalogs accessible [14].

Semantic Web technologies offer a new approach to managing information and processes, by using semantic metadata. This metadata can exist at two different levels. On one level it can describe a document, a web page for example, and on the other level, it can describe entities within the document, like museum objects or songs. The purpose of this technology is to provide context about resources on the Web and tools to use this context, so that machines can make decisions, using structured collections of information and sets of inference rules that they use to conduct automatic reasoning [4]. Semantic Web technologies permit the structuring of metadata using ontologies.

Ontologies have been defined in many ways, but the most cited definition in the knowledge sharing community is Gruber's [9] definition: "An ontology is a description, or a formal specification, of concepts and relationships that exist for an agent or a community of agents, defined in the representational vocabulary, which is used to represent knowledge". Different ontologies have been developed to represent cultural heritage knowledge, such as the CIDOC Conceptual Reference Model (CRM) ontology [7]. The CIDOC CRM has also been converted to OWL [8].

In the field of music information retrieval, formal ontologies have been used to describe data as varied as music events and timelines [12], musical instruments [10], and also folk music performers and records [11]. As with the present work, the latter ontology has been expressed in the CIDOC CRM.

CIDOC CRM facilitates the integration, mediation and interchange of heterogeneous cultural heritage information. It is the culmination of years of standards development work by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM), and it is now an international standard ISO21127. It has been developed by interdisciplinary teams of experts, from fields like computer science, archaeology and museum documentation, and it defines the underlying semantics of document structures used in cultural heritage documentation in terms of a formal ontology. CIDOC CRM is intended for the representation of information required for exchanging and integrating heterogeneous scientific documentation of museum collections.

The work reported in this is paper is in the context of a project of Basque folk song analysis, based on two different Basque folk song catalogs. The first is the Cancionero Vasco, which comprises 1902 songs gathered by Father Donostia, a priest born in 1886. The second is the Cancionero Popular Vasco, which comprises 1175 songs that were collected by Resurrección María de Azkue, a priest born in 1864. Each song contains two types of information: musical data encoding the melody (in MIDI format) and metadata. This metadata represents 26 features, such as the place where the song was collected, the normalized title and genre of the song [5]. All this information has been digitized and organized by Euskomedia Foundation [1], belonging to *Eusko Ikaskuntza - Basque Studies Society*, and Eresbil [2], a Basque musical archive that works on collect-

[1] www.euskomedia.org
[2] www.eresbil.com

ing, preserving, protecting and disseminating the Basque musical heritage.

The main objective of the project is to organize the metadata of songs in a hierarchical way, structuring the song collections. This structure will allow searching in the catalog, grouping songs depending on their features, and making complex queries in the catalog that are impossible to make with a flat structure. This ontology will be used directly in the analysis of musical properties of the Basque songs, using automated pattern discovery [6]. This consists in discovering patterns that are distinctive and maximally general in a genre and using these patterns to classify songs of unknown genres.

Folk song collections have several interesting peculiarities that must be handled in any formal representation. Songs can have different geographic roles: the toponymic information can make reference to the place where a song was collected or to the place where the person who transmitted the song was from. Songs may have missing toponymic information. More than one place can exist under the same name. All of these considerations are handled by the CIDOC CRM solution.

Queries that can be done using the hierarchical structure can be simple, like "get all the songs collected in Gipuzkoa (territory)", but more complex queries such as "get all songs collected in Gipuzkoa or any of its contained regions" require formal inference. In Table 1 a list of use cases that can be implemented using the structure presented in this paper is shown. All of these use cases are handled by the OWL reasoner with the ontology.

Get all the songs collected in territory Gipuzkoa
Get the names of all the territories
Get all the urban areas in territory Bizkaia
Get all the songs collected within Gipuzkoa (territory)
Get all the song classified as dances in Bizkaia
Get all songs with the name "Agur,_adiskidea"
Get all the songs from territories that border with Bizkaia

Table 1. Use case examples for folk music metadata queries.

2. ONTOLOGY DEVELOPMENT

This section describes the tools used to build the ontologies that structure metadata of the song catalog. The methods for representation of folk song metadata in the CIDOC CRM are described. Finally it is shown how description logic queries can be used to retrieve songs from the ontology.

2.1 Development tools

The most important development tool used in this work is OWL 2 Web Ontology Language. It is the recommended markup language used to define ontologies for use in the Semantic Web and it is compatible with the OWL 1 standard of 2004. It is designed to be used by applications that process content of information instead of just presenting it. It can describe the meaning of terms in vocabularies and represent the relations between these terms, namely ontologies [15]. There is a Java open source API, for creating and manipulating OWL ontologies, named OWL API [1]. It includes various interfaces for accessing OWL reasoners, that are available for download in order to provide implementations to the interfaces.

OWL 2 builds on XML and provides machine readable semantic annotations. It can express classes, object properties (roles), individuals (instances) and data properties. Classes represent concepts, individuals represent instances of classes, data properties represent information about individuals and object properties are representations of relations between classes or between individuals. OWL 2 is a revision of the DAML+OIL web ontology language, and it has more tools for expressing semantics than XML, RDF and RDF-S.

Ontology visualization and the quality of interaction provided are very important, since the concepts expressed in an ontology are very specific and the analysis of individual relations is complex [13]. Protégé [2] has ontology visualization techniques implemented. It provides the possibility of using different OWL semantic reasoners as a plug-in. Semantic reasoners are pieces of software able to infer logical consequences from a set of asserted facts of axioms (an ontology), and two different reasoners are available on the current version of Protégé; FaCT++ and HermiT.

2.2 CIDOC CRM

The CIDOC CRM has some very useful classes and roles to build the ontologies for the metadata of folk songs. Places are defined as instances of class E53_Place, and as more than one place can exist with the same name, place names are defined as instances of class E44_Place_Appellation.

Names (appellations) and places are connected by the role P1_is_identified_by, which relates place (subclass of E1_CRM_Entity) instances to place appellation (subclass of E41_Appellation) instances, making possible the existence of two or more different toponyms with the same name.

As CIDOC CRM was initially created for museum collections, it does not have a class for songs, therefore a new class song has been added as a subclass of E73_Information_Object, with scope "identifiable immaterial items that have an objectively recognizable structure". Different song instances can have the same song name, so the class E35_Title is used to represent the song titles which can then be added to instances using the role P1_is_identified_by.

To relate these classes CIDOC CRM has roles defined as object properties. The roles used to build the ontology are presented in the Table 2 below. Roles P89_falls_within and its inverse P89i_contains manage regions containment relations, permitting regions with different level of specificity referenced by songs in the catalog. These are transi-

tive roles, allowing to have containment relations between toponyms of different specificity levels. Role P122_borders_with represents borders between different regions.

Figure 1. Map of Euskal Herria territories.

Songs in the catalogs can be related to toponyms with two possible roles; a song can be collected in a place or it can be transmitted by someone from a certain place. CIDOC CRM does not have object properties able to represent these roles, and two new object properties (and their inverses) have been added: collectedIn and informantFrom.

2.3 Toponymic information

All the pieces in the Basque folk song catalogs have been collected in *Euskal Herria*, a region that consists of 7 territories, as shown in Figure 1. Three of these territories are located in French territory, and the other four are located in Spanish territory.

These toponyms have three levels of specificity; territory level, municipality level and urban area level, and this specificity is represented in the ontology with a new class for each level. The toponyms of Euskal Herria are represented as instances of the class E53_Place. Each instance of the E53_Place class is also an instance of one of these three classes, according to the specificity of the toponym. Songs in the catalogs can make reference to any of these three levels, or they can have an empty toponym information field.

In Figure 2 an example of a toponym definition can be seen. muni-ELORRIO is an instance of E53_Place to specify that it is a toponym, and it also is an instance of municipio to specify that it is in the municipality level. As presented in Section 2, the developed ontology uses some CIDOC CRM roles to relate entities between them. All the toponym instances have containment relations (role P89_falls_within) to relate the urban areas with their municipalities and municipalities with their territories. This is represented in Figure 2 with the relations between instances terr-BIZKAIA and muni-ELORRIO and between muni-ELORRIO and nucl-ELORRIO. The borders between territories are defined using role P122_borders_with, and each song has an instance of object property collectedIn or informantFrom to relate it to a toponym. The CIDOC CRM role P1_is_identified_by is used to relate place appellation instances to place instances. This is shown in Figure 2, where the place appellation ELORRIO is related to municipality instance muni-ELORRIO and urban area instance nucl-ELORRIO, two different entities but with the same appellation. The collectedIn role, that relates the songs to places where they were collected, is also shown in Figure 2, with the relation between the song 1539 and the toponym nucl-ELORRIO.

2.4 Genres

The genres sub-ontology represents a hierarchy for all the genres referenced in the two song catalogs. The hierarchy has 51 genre classes and it has been created by musicologists from Eresbil in order to have a common classification hierarchy for the two song catalogs, since they use different nomenclatures for the genre names. Each song is represented as an instance of its corresponding genre class.

Figure 3 shows a fragment of the subclasses created for the class Canciones de entretenimiento.

In Figure 4 an example of a combination of toponym and genre information can be seen. The instances 2337 and 2338 are songs collected in an urban area called *Lekaroz* from the municipality *Baztan*. Both have the same title, *Mila zortziehun eta berrogei*, and are religious songs. In Table 3 the definition of some of the entities of Figure 4 is shown.

2.5 Queries

The ontology structure presented above permits complex queries like the ones presented on Table 4. Some of these queries are simple, like the first one on Table 4 “get all the songs collected in territory Gipuzkoa”. The use of a reasoner is not necessary to answer that query, since it just returns the songs that are annotated to be collected in Gipuzkoa. On the other hand, to solve the second query, “get all the songs collected within territory Gipuzkoa”, the use of a reasoner is mandatory, as it must gather all the songs collected in all the toponyms that fall within Gipuzkoa. There are more complex queries that are interesting for song catalogs like the ones used in this project, like “get all the song classified as dances in Bizkaia”. The reasoner uses information about toponym and genre subontologies to answer this query, returning songs instantiating the query from both catalogs.

3. DISCUSSION AND CONCLUSIONS

This paper presented a method to encode and structure metadata of folk song collections. This method uses and extends the CIDOC CRM, which is used by many researchers in the museum documentation field. The roles and classes provided by CIDOC CRM have been useful in dealing with the particularities of the Basque folk song collections. The CIDOC CRM extension with song and genre types could be used for encoding metadata of other folk tune collections.

Role	Domain	Range	Example
P1_is_identified_by	E1_CRM_Entity	E41_Appellation	muni-LEZO P1_is_identified_by LEZO
P1i_identifies	E41_Appellation	E1_CRM_Entity	LEZO P1i_identifies nucl-LEZO
P1i_identifies	E35_Title	E1_CRM_Entity	Allegretto P1i_identifies 3316
P89_falls_within	E53_Place	E53_Place	muni-LEZO P89_falls_within terr-GIPUZKOA
P89i_contains	E53_Place	E53_Place	terr-GIPUZKOA P89i_contains muni-LEZO
P122_borders_with	E53_Place	E53_Place	terr-BIZKAIA P122_borders_with terr-GIPUZKOA
collectedIn	song	E53_Place	1539 collectedIn nucl-ELORRIO
informantFrom	song	E53_Place	999 informantFrom nucl-ARETXABALETA

Table 2. Roles used in the project. The top roles are CIDOC CRM roles, the last two are custom roles added for folk song metadata representation.

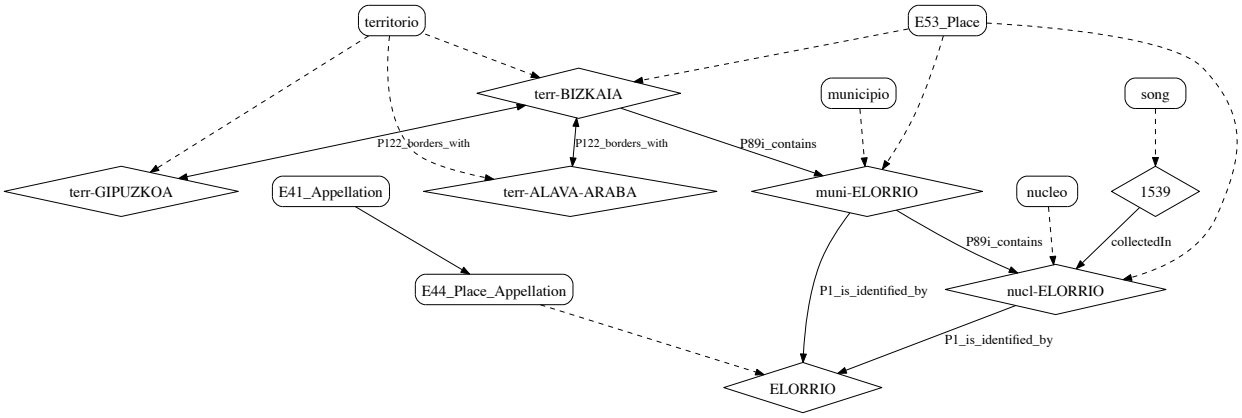


Figure 2. Example of a toponym definition. Instances are diamonds and classes are in boxes. The dashed arrows represent class-instance relations, while solid arrows represent relations between instances and also subsumption relations.

The formal encoding of the toponymic information and containment relations among regions brings up some interesting ideas in visualization. The work of Aarden and Huron [3] could be applied and extended to present visualizations of genre distributions in particular regions with the ability to focus further on subregions. In fact any feature of subgroup of songs, for example a distinctive pattern [6] or any other global melodic feature, could be visualized with reference to the toponym ontology.

Future work with ontology development will consider the representation of melodic and rhythmic content of songs, for example the representation of global features, distinctive patterns, and tune families. The ontology will also be used directly for data mining and pattern discovery, with patterns distinctive to particular genres or regions represented directly as description logic concepts and integrated into the ontology.

4. ACKNOWLEDGMENTS

This research was partially supported by a grant *Análisis Computacional de la Música Folclórica Vasca* (2011-2012) from the Diputación Foral de Gipuzkoa, Spain. Thanks to K. Neubarth for valuable comments on the manuscript, and to J. Wissmann for consultations on Protégé and OWL.

5. REFERENCES

[1] http://owlapi.sourceforge.net/documentation.html.

[2] http://protege.stanford.edu.

[3] B. Aarden and D. Huron. Mapping European folksong: Geographical localization of musical features. *Computing in Musicology*, 12:169–183, 2001.

[4] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[5] D. Conklin. Computational analysis of Basque song collections. *Euskonews*, 2011. online at http://www.euskonews.com/0594zbk.

[6] D. Conklin and C. Anagnostopoulou. Comparative pattern analysis of cretan folk songs. *Journal of New Music Research*, 40(2):119–125, 2011.

[7] N. Crofts, M. Doerr, T. Gill, S. Stead, and M. Stiff, editors. *Definition of the CIDOC Conceptual Reference Model*. ICOM/CIDOC CRM Special Interest Group, Oct. 2006.

[8] G. Görz, M. Oischinger, and B. Schiemann. An implementation of the CIDOC conceptual reference model

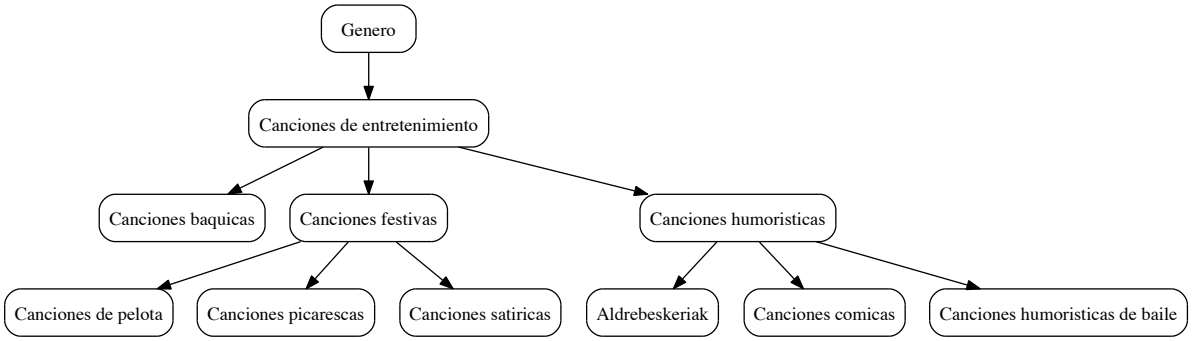


Figure 3. Detail of a fragment of the genre sub-ontology, showing a major subclass (Canciones de entretenimiento) with three minor subclasses. Arrows represent subsumption relations.

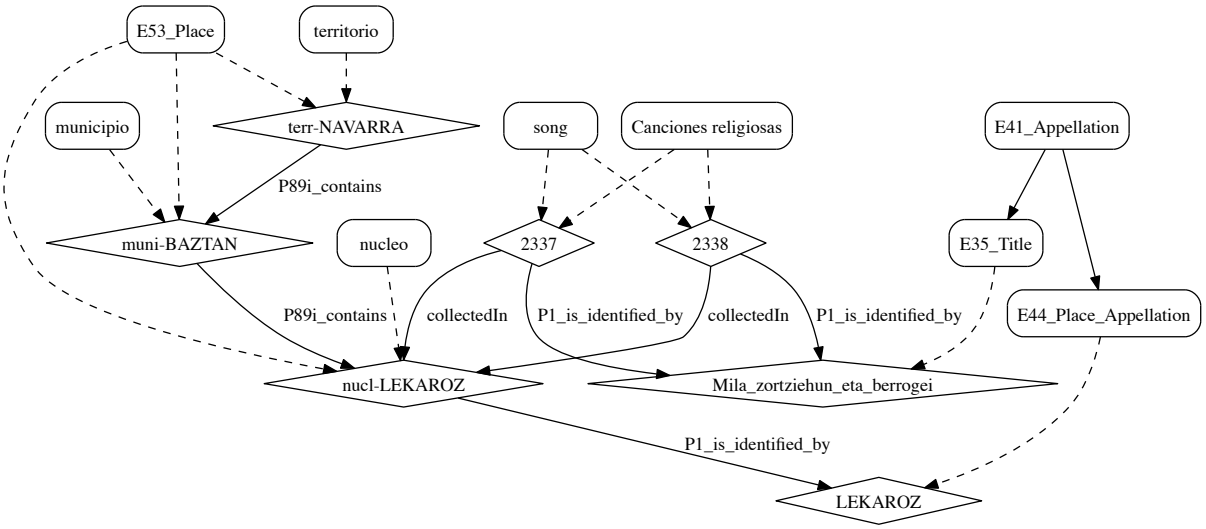


Figure 4. Fragment of a combination of genre and toponym classes. The dashed arrows represent class-instance relations while solid arrows represent relations between instances and also subsumption relations.

Entity	Type	Definition
Canciones Religiosas	class	SubClassOf Canciones de Creencias
2338	individual	Type song and Type Canciones Religiosas
muni-BAZTAN	individual	Type municipio and Type E53_Place
LEKAROZ	individual	Type E44_Place_Appellation
E35_Title	class	SubClassOf E41_Appellation
territorio	class	SubClassOf Region

Table 3. Definition of some of the entities in Figure 4.

Use case	DL Query	Results
Get all the songs collected in territory Gipuzkoa	collectedIn some terr-GIPUZKOA	3 songs
Get all the songs collected within territory Gipuzkoa	collectedIn some (P89_falls_within some terr-GIPUZKOA or collectedIn some terr-GIPUZKOA)	388 songs
Get all the urban areas in Bizkaia (territory)	P89_falls_within value terr-BIZKAIA and nucleo	642 urban areas
Get all the song classified as dances in Bizkaia	collectedIn some (P89_falls_within some terr-BIZKAIA) or collectedIn some terr-BIZKAIA) and Danzas	68 songs
Get all songs with name "Agur,_adiskidea"	P1_is_identified_by some Agur,_adiskidea	6 songs
Get the names of all the territories	P1i_identifies some territorio	7 territory names
Get the names of the territories that border with Gipuzkoa	P1i_identifies some (P122_borders_with value terr-GIPUZKOA)	4 territory names
Get all the songs from territories that border with Bizkaia	collectedIn some (P89_falls_within some (P122_borders_with value terr-BIZKAIA)) or (collectedIn some (P122_borders_with value terr-BIZKAIA))	451 songs

Table 4. Solved query examples

(4.2.4) in OWL-DL. In *Proceedings CIDOC 2008 — The Digital Curation of Cultural Heritage. Athens, Benaki Museum, 15.–18.09.2008*, Athens, September 2008. ICOM CIDOC.

[9] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. In *International Journal of Human-Computer Studies*, pages 907–928. Kluwer Academic Publishers, 1993.

[10] S. Kolozali, M. Barthet, G. Fazekas, and M. Sandler. Knowledge representation issues in musical instrument ontology design. In A. Klapuri and C. Leider, editors, *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011*, pages 465–470, 2011.

[11] M. Marolt, J. Vratanar, and G. Strle. Ethnomuse: Archiving folk music and dance culture. In *EUROCON 2009. IEEE*, pages 322 –326, May 2009.

[12] Y. Raimond, S. A. Abdallah, M. B. Sandler, and F. Giasson. The music ontology. In S. Dixon, D. Bainbridge, and R. Typke, editors, *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007*, pages 417–422. Austrian Computer Society, 2007.

[13] I. C. Siqueira da Silva and C. M. Dal Sasso Freitas. Using visualization for exploring relationships between concepts in ontologies. In *Proceedings of the 2011 15th International Conference on Information Visualisation*, IV '11, pages 317–322, Washington, DC, USA, 2011. IEEE Computer Society.

[14] S. Staab, R. Studer, and M. Doerr. *Ontologies for Cultural Heritage*, pages 463–486. Springer Berlin Heidelberg, 2009.

[15] W3C OWL Working Group. *OWL 2 Web Ontology Language: Document Overview*. W3C Recommendation, 27 October 2009. Available at url http://www.w3.org/TR/owl2-overview/.