FLABASE: TOWARDS THE CREATION OF A FLAMENCO MUSIC KNOWLEDGE BASE

Sergio Oramas[1], Francisco Gómez[2], Emilia Gómez[1], Joaquín Mora[3]
[1]Music Technology Group, Universitat Pompeu Fabra
[2]Technical University of Madrid
[3]Faculty of Psychology, University of Sevilla
{sergio.oramas, emilia.gomez}@upf.edu, fmartin@eui.upm.es, mora@us.es

ABSTRACT

Online information about flamenco music is scattered over different sites and knowledge bases. Unfortunately, there is no common repository that indexes all these data. In this work, information related to flamenco music is gathered from general knowledge bases (e.g., Wikipedia, DBpedia), music encyclopedias (e.g., MusicBrainz), and specialized flamenco websites, and is then integrated into a new knowledge base called FlaBase. As resources from different data sources do not share common identifiers, a process of pair-wise entity resolution has been performed. FlaBase contains information about 1,174 artists, 76 *palos* (flamenco genres), 2,913 albums, 14,078 tracks, and 771 Andalusian locations. It is freely available in RDF and JSON formats. In addition, a method for entity recognition and disambiguation for FlaBase has been created. The system can recognize and disambiguate FlaBase entity references in Spanish texts with an f-measure value of 0.77. We applied it to biographical texts present in Flabase. By using the extracted information, the knowledge base is populated with relevant information and a semantic graph is created connecting the entities of FlaBase. Artists relevance is then computed over the graph and evaluated according to a flamenco expert criteria. Accuracy of results shows a high degree of quality and completeness of the knowledge base.

1. INTRODUCTION

Music context information is now playing a key role in MIR research. Multimodal approaches, semantic approaches, and text-IR approaches have shown important achievements in typical MIR problems, such as music recommendation and discovery, genre classification, or music similarity [17]. Therefore, collecting and storing music context information may be extremely useful for the MIR research community [13]. There are some broad repositories of music context information such as MusicBrainz [1] or Discogs [2]. Although some of these repositories are very complete and accurate, there is still a vast amount of music information out there, which is generally scattered among different sources on the Web. Hence, harvesting and combining that information is a crucial step in the creation of practical and meaningful music knowledge bases. In addition, the creation of genre-specific knowledge bases may be very valuable for research and dissemination purposes, and particularly to non-western music traditions.

In this paper, we propose a methodology for the creation of a genre-specific knowledge base; in particular, a knowledge base of flamenco music. The proposed methodology combines content curation and knowledge extraction processes. First, an important amount of information is gathered from different data sources, which are subsequently combined by applying pair-wise entity resolution. Next, new knowledge is extracted from unstructured harvested texts and employed to populate the knowledge base. For this purpose, an entity linking system has been expressly developed. Finally, the content of the knowledge base is used to compute artist relevance and results are evaluated according to flamenco experts criteria. The content of the knowledge base is freely available and downloadable as data dumps in RDF and JSON formats.

The remainder of the paper is organized as follows. In Section 2, an introduction to flamenco music is presented. In Section 3 some relevant prior work is briefly surveyed. Section 4 describes the structure of the knowledge base. Next, in Section 5 the process of content curation is explained. Section 6 shows the methodology applied for knowledge extraction. In Section 7 artist relevance is computed and some statistics about the content are laid out. Finally, Section 8 concludes the paper and points out for future lines of work.

2. FLAMENCO MUSIC

Several musical traditions contributed to the genesis of flamenco music as we know it today. Among them, the influences of the Jews, Arabs, and Spanish folk music are recognizable, but indubitably the imprint of Andalusian Gypsies' culture is deeply ingrained in flamenco music. Fla-

[1] http://musicbrainz.org
[2] http://www.discogs.com/

menco occurs in a wide range of settings, including festive *juergas* (private parties), *tablaos* (flamenco venues), concerts, and big productions in theaters. In all these settings we find the main components of flamenco music: *cante* or singing, *toque* or guitar playing, and *baile* or dance. According to Gamboa [9], flamenco music grew out of the singing tradition, as a melting process of all the traditions mentioned above, and therefore the role of the singer soon became dominant and fundamental. *Toque* is subordinated to *cante*, especially in more traditional settings, whereas *baile* enjoys more independence from voice.

In the flamenco jargon styles are called *palos*. Criteria adopted to define flamenco *palos* are rhythmic patterns, chord progressions, lyrics and its poetic structure, and geographical origin. In flamenco geographical variation is important to classify *cantes* as often they are associated to a particular region where they were originated or where they are performed with gusto. Rhythm or *compás* is a unique feature of flamenco. Rhythmic patterns based on 12-beat cycles are mainly used. Those patterns can be classed as follows: binary patterns, such as *tangos* or *tientos*; ternary patterns, which are the most common ones, such as *fandangos* or *bulerías*; mixed patterns, where ternary and binary patterns alternate, such as *guajira*; free-form, where there is no a clear underlying rhythm, such as *tonás*. For further information on fundamental aspects of flamenco music, see the book of Fernández [7]. For a comprehensive study of styles, musical forms and history of flamenco the reader is referred to the books of Blas Vega and Ríos Ruiz [3], Navarro and Ropero [12], and Gamboa [9] and the references therein.

3. RELATED WORK

A knowledge base is a centralized repository intended to store both complex structured and unstructured information. Content in a knowledge base can be either curated or extracted, and knowledge bases can be classified according to those criteria [6]. Curated knowledge can be manually gathered by humans or automatically extracted from a structured data source. By contrast, extracted knowledge is produced after the application of an information extraction process over an unstructured data source. There are several well-known general purpose knowledge bases either extracted or curated. The most widely used are DBpedia [3] and Freebase [4], and more recently WikiData [5]. The most relevant extracted knowledge bases are NELL [5] and Open IE [1].

In the music field, one of the most complete and broadly used knowledge bases is MusicBrainz [6], which has been created in a collaborative curated way. However, there is not any extracted and open music knowledge base. Moreover, little effort have been done in the creation of genre-specific knowledge bases. Most relevant initiatives in this direction have been done within the CompMusic project [7]. In this project, one of the main tasks has been the gathering of culture-specific corpora of non-western musical traditions, combining expert information, audio recordings, features, music notation, lyrics, editorial metadata and community information [18]. According to [19], a domain-specific corpora should be designed by satisfying the following criteria: purpose, coverage, completeness, quality and reusability. In [15], the architecture and applications of a system that exploits domain-specific corpora is presented. Another interesting project is Linked Jazz [14], where the application of Linked Open Data (LOD) technology to enhance discovery and visibility of jazz music is studied.

4. FLABASE

FlaBase (Flamenco Knowledge Base) is the acronym of a new knowledge base of flamenco music. Its ultimate aim is to gather all available online editorial, biographical and musicological information related to flamenco music. A first version is just being released. Its content is the result of the curation and extraction processes explained in Sections 5 and 6. FlaBase is stored in RDF and JSON formats, and it is freely available for download [8]. Its RDF version follows the Linked Open Data principles, and it might be queried by setting up a SPARQL endpoint. A JSON version is also available, thus facilitating the use of the content by all the community of researchers and developers. This first release of FlaBase contains information about 1,174 artists, 76 *palos* (flamenco genres), 2,913 albums, 14,078 tracks, and 771 Andalusian locations.

4.1 Ontology Definition

The FlaBase data structure is defined in an ontology schema. One of the advantages of using an ontology as a schema is that it can be easily modified. Thus, our design is a first building block that can be enhanced and redefined in the future. The initial ontology is structured around five main classes: MusicArtist, Album, Track, Palo and Place, and three domain specific classes: *cantaor* (flamenco singer), guitarist (flamenco guitar player), and *bailaor* (flamenco dancer). These three classes were defined because they are the most frequent types of artists in the data. Other instrument players may be instantiated directly from the MusicArtist class. We have tried to reuse as much vocabulary as we could. We re-utilized most of the classes and some properties from the Music Ontology [9], a standard model for publishing music-related data. We selected the classes according to the ones used by the LinkedBrainz project [10], which maps concepts from MusicBrainz to Music Ontology.

[3] http://dbpedia.org
[4] http://www.freebase.com
[5] http://www.wikidata.com
[6] http://musicbrainz.org
[7] http://compusic.upf.edu
[8] http://mtg.upf.edu/download/datasets/flabase
[9] http://musicontology.com
[10] https://wiki.musicbrainz.org/LinkedBrainz

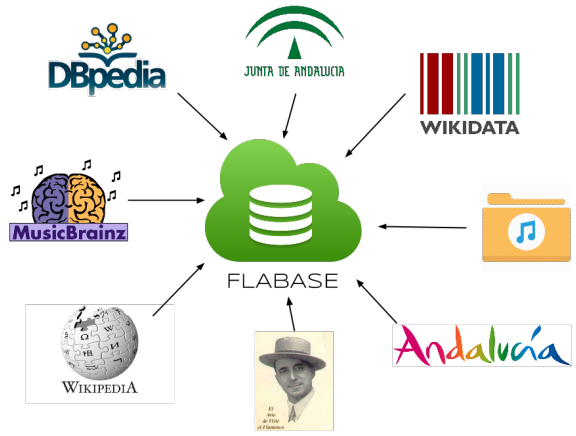

Figure 1. Selected data sources

5. CONTENT CURATION

The first step towards building a domain-specific knowledge base is to gather all possible content from available data sources. This implies at least two problems, namely, the selection of sources, and the matching between entities from different sources. In what follows we enumerate the involved data sources and describe the methodology applied to entity resolution.

5.1 Data Acquisition

Our aim is to gather an important amount of information about musical entities, including textual descriptions and available metadata. A schema of the selected data sources is shown in Figure 1. We started by looking at Wikipedia [11], the free and multilingual Internet encyclopedia. It is the Internet's largest and most popular general reference work. Each Wikipedia article may have a set of associated categories. Categories are intended to group together pages on similar subjects and are structured in a taxonomical way. To find Wikipedia articles related to flamenco music, we first looked for flamenco categories. The taxonomy of categories can be explored by querying DBpedia, a knowledge base with structured content extracted from Wikipedia. In particularly, we employed the SPARQL endpoint of the Spanish DBpedia [12]. We queried for categories related to the flamenco category in the taxonomy. At the end, we obtained 17 different categories (e.g., *cantaores de flamenco, guitarristas de flamenco*).

By querying again DBpedia, we gathered all DBpedia resources related to one of these categories. We obtained a total number of 438 resources in Spanish, of which 281 were also in English. Each DBpedia resource is associated with a Wikipedia article. Text and HTML code were then extracted from Wikipedia articles in English and Spanish by using the WikiMedia API. Next, we classified the extracted articles according to the ontology schema defined in our knowledge base (Section 4.1). For this purpose, we exploited classification information provided by DBpedia (DBpedia ontology and Wikipedia categories). At the end, from all gathered resources, we only kept those related to artists and *palos*, totalling 291 artists and 56 *palos*.

However, the amount of information present in Wikipedia related to flamenco music is somewhat scarce. Therefore, we decided to expand our knowledge base with information from two different websites. First, *Andalucia.org*, the touristic web from the Andalusia Government [13]. It contains 422 artist biographies in English and Spanish, and the description of 76 *palos* also in both languages. Second, a website called *El arte de vivir el flamenco* [14], which includes 749 artist biographies among *cantaores*, *bailaores* and guitarists. Both webs were crawled and their content stored in our knowledge base.

MusicBrainz is one of the biggest and more reliable open music databases, which provides an unambiguous form of music identification. Therefore, we turned to it in order to fill our knowledge base with information about flamenco album releases and recordings. Artists present in FlaBase were intended to be mapped with MusicBrainz artists. For every match, all content related to releases and recordings was gathered. After doing so, we obtained a total number of 814 releases and 9,942 recordings.

The information gathered from MusicBrainz is a little part of the actual flamenco discography. Therefore, to complement it we used a flamenco recordings database gathered by Rafael Infante and available at CICA website [15] (Computing and Scientific Center of Andalusia). This database has information about releases from the early time of recordings until present time, counting 2,099 releases and 4,136 songs. For every song entry, a *cantaor* name is provided, and most of the times also guitarist and *palo*, which is a very valuable information to define flamenco recordings.

Finally, we supplied our knowledge base with information related to Andalusian towns and provinces. We gathered this information from the official database SIMA [16] (Multi-territorial System of Information of Andalusia).

5.2 Entity Resolution

Entity Resolution (ER) is the problem of extracting, matching and resolving entity mentions in structured and unstructured data [10]. There are several approaches to tackle the ER problem. For the scope of this research, we selected a pair-wise classification approach based on string similarity between entity labels.

The first issue after gathering the data is to decide whether two entities from different sources are referring to the same one. Therefore, given two sets of entities A and B, the objective is to define an injective and non-surjective mapping function f between A and B that decides whether an entity $a \in A$ is the same as an entity $b \in B$. To do that, a string similarity metric $sim(a, b)$ based on the Ratcliff-Obershelp algorithm [16] has been defined. It measures

[11] http://www.wikipedia.org
[12] http://es.dbpedia.org
[13] http://andalucia.org
[14] http://www.elartedevivirelflamenco.com/
[15] http://flun.cica.es/index.php/grabaciones
[16] http://www.juntadeandalucia.es/institutodeestadisticaycartografia/sima

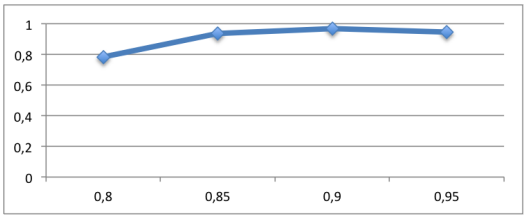

Figure 2. F-measure for different values of θ

the similarity between two entity labels and outputs a value between 0 and 1. We consider that a and b are the same entity if their similarity is bigger than a parameter θ. If there are two entities $b, c \in B$ that satisfy that $sim(a, b) \geq \theta$ and $sim(a, c) \geq \theta$, we consider only the mapping with the highest score. To determine the value of θ, we tested the method with several θ values over an annotated dataset of entity pairs. To create this dataset, the 291 artists gathered from Wikipedia were manually mapped to the 422 artists gathered from Andalucia.org, obtaining a total amount of 120 pair matches. As it is shown in Figure 2 the best F-measure (0,97) was obtained with $\theta = 0.9$. Finally, we applied the described method with $\theta = 0.9$ to all gathered entities from the three data sources. Thanks to the entity resolution process, we reduced the initial set of 1,462 artists and 132 *palos* to a set of 1,174 artists and 76 *palos*.

Once we had our artist entities resolved, we began to gather their related discographic information. First, we tried to find out the MusicBrainz ID of the gathered artists. Depending on the information about the entity, two different process were applied. First, every Wikipedia page, and its equivalent DBpedia resource, has a correspondent entity defined in Wikidata. Wikidata is a free linked database which acts as a structured data storage of Wikipedia. There are several properties in Wikidata that may link Wikidata items with MusicBrainz items. Thus, the equivalent Wikidata resource of a Wikipedia artist page may have a link to its corresponding MusicBrainz artist ID. Therefore, we looked for these relations and mapped all possible entities. For those artists without a direct link to MusicBrainz, we queried the MusicBrainz API by using the artist labels, and then applied our entity resolution method to the obtained results.

Finally, to integrate the discography database of CICA into our knowledge base, we applied the entity resolution method to the fields *cantaor*, guitarist and *palo* of each recording entry in the database. From the set of 202 *cantaores* and 157 guitarists names present in the recording entries, a total number of 78 *cantaores* and 44 guitarists were mapped to our knowledge base. The number of mapped artists was low due to differences between the way of labeling an artist. An artist name may be written using one or two surnames, or using a nickname. In the case of *palos*, there were 162 different *palos* in the database, 54 of which were mapped with the 76 of our knowledge base. These 54 *palos* correspond to an 80% of *palo* assignments present in the recording entries.

6. KNOWLEDGE EXTRACTION

Once the process of data acquisition is finished, the knowledge base is ready for use. However, there is an important amount of knowledge present in the data that has not been fully exploited. Texts gathered contain a huge epistemic potential that remains implicit. Consequently, to enhance the amount of structured data in FlaBase, a process of knowledge extraction has been carried out. This implicit knowledge may vary from biographical data, such as place and date of birth, to more complex semantic relations involving different entities. Three tasks play a key role in the process of knowledge extraction from non-structured text: named entity recognition (NER), named entity disambiguation (NED), and relation extraction (RE) [20]. In this research, we focus on the two first tasks. In what follows, a system for entity recognition and disambiguation is described and evaluated. Lastly, an information extraction process is applied to populate the knowledge base.

6.1 Named entity recognition and disambiguation

To extract implicit knowledge from a text, the first step is to semantically annotate it by identifying entity mentions. Named entity recognition is a task that seeks and classify words in text into pre-defined categories (e.g., person, organization, or place). Named entity disambiguation, also called entity linking, aims to determine what is actually a named entity present in a text. It generally does so by identifying it in a knowledge base of reference. NED can be addressed directly from the text, or applied to the output of a NER system. We propose a method that employs a combination of both approaches, depending on the category of the entity. For NER, we used the Stanford NER system [8], implemented in the library Stanford Core NLP [17] and trained on Spanish texts. For NED we tried two different approaches. First, we looked for exact string matches between FlaBase entity labels and word n-grams extracted from the text. Second, we searched for exact string matches between FlaBase entity labels and the output of the NER system. In fact, we tried several combinations of both approaches until we obtained the most satisfactory one.

For the scope of this research, we focused on Spanish texts, as flamenco texts are mostly written in Spanish. Although there are many entity linking tools available, we decided to develop ours because state-of-the-art systems (e.g., Tag-me or Babelfy) are well-tuned for English texts, but do not perform well on Spanish texts, and even less with music texts of a specific domain. In addition, we wanted to have a system able to map entities to our knowledge base. Therefore, we developed a system able to detect and disambiguate three categories of entities: person, *palo* and location. Three different approaches were defined by combining NER and NED in different ways according to the category. First, directly applying NED to text. Second, disambiguating location and person entities from the

[17] `http://nlp.stanford.edu/software/corenlp.shtml`

Approach	Precision	Recall	F-measure
1) NED	0.829	**0.694**	0.756
2) NED + NER to PERS & LOC	0.739	0.347	0.472
3) NED + NER to LOC	**0.892**	0.674	**0.767**

Table 1. Precision, Recall and F-measure of NER+NED

NER output, and *palo* directly from text. Third, only disambiguating location entities from the NER output, and location and *palo* directly from text.

To determine which approach performs better, three artist biographies coming from three different data sources were manually annotated, having a total number of 49 annotated entities. We followed an evaluation methodology similar to the one used in KBP2014 Entity Linking Task [18] . Results on the different approaches are shown in Table 1. We observe that applying NER to entities of the person category before NED worsens performance significantly, as recall suddenly decrease by half. After manually analysing false negatives, we observed that this is caused because many artist names have definite articles between name and surname (e.g., *de, del*), and this is not recognized by the NER system. In addition, many artists have a nickname that is not interpreted as a person entity by the NER system. The best approach is the third (NED + NER to LOC), which is slightly better than the first (only NED) in terms of precision. This is due to the fact that many artists have a town name as a surname or as part of his nickname. Therefore, applying NED directly to text is misclassifying person entities as location entities. Thus, by adding a previous step of NER to location entities we have increased overall performance, as it can be seen on the F-measure values.

6.2 Knowledge base population

Biographical texts coming from different data sources have been stored in FlaBase. These texts are full of relevant information about FlaBase entities, but in an unstructured way. Thus, a process of information extraction is necessary to transform the unstructured information into structured knowledge. For the scope of this research, we focused on extracting two specific data: birth year and birth place, as they can be very relevant for anthropologic studies. We observed that this information is often in the very first sentences of the artist biographies, and always near the word *nació* (Spanish translation of "was born"). Therefore, to extract this information, we looked for this word in the first 250 characters of every biographical text. If it is found, we apply our entity linking method to this piece of text. If a location entity is found near the word "nació", we assume that this entity is the place of birth of the biography subject. In addition, by using regular expressions, we look for the presence of a year expression in the neighborhood. If it is found, we assume it as the year of birth. If more than one year is found, we select the one with the smaller value.

To evaluate our approach, we tested the extraction of birth places in all texts coming from the web Andalucia.org (442 artists). We chose this subset because Andalucia.org also provides specific information about artist origin that had been previously crawled and stored in FlaBase. However, we observed that in many occasions the artist origin provided by the data source was wrong. Therefore, we decided to manually annotate the province of precedence of these 442 artists for building ground truth data. After the application of the extraction process on the annotated test set, we obtained a precision value of 0,922 and a recall of 0,648. Therefore, we can state that our method is extracting biographic information with very high precision and quite reasonable recall. We finally applied the extraction process to all artist entities with biographical texts coming from any of the two flamenco crawled websites. Thus, from a total number of 1,123 artists coming from these data sources (95% of the artists in the knowledge base), 743 birth places and 879 birth years were extracted.

[18] http://nlp.cs.rpi.edu/kbp/2014/

7. LOOKING AT THE DATA

7.1 Artist Relevance

We assume that an entity mention inside an artist biography means a semantic relation between the biography subject and the mentioned entity. Based on this assumption, we build a semantic graph by applying the following steps. First, each artist of the knowledge base is added to the graph as a node. Second, entity linking is applied to artist's biographical texts. For every linked entity, a new node is created in the graph (only if it was not previously created). Next, an edge is added by connecting the artist entity node with the linked entity node. This way, a directed graph connecting the entities of FlaBase is finally obtained. Entities identified in a text can be seen as hyperlinks. Hence, algorithms to measure the relevance of nodes in a network of hyperlinks can be applied to our semantic graph [2]. In order to measure artist relevance, we applied PageRank [4] and HITS [11] algorithms to the obtained graph.

We built an ordered list with the top-10 entities of the different artist categories (*cantaor*, guitarist and *bailaor*) for the two algorithms. For evaluation purposes, we asked a flamenco expert to build a list of top-10 artists for each category according to his knowledge and the available bibliography. The concept of artist relevance is somehow subjective and there is no unified or consensual criteria for flamenco experts about who the most relevant artists are. Despite that, there is a high level of agreement among them on certain artists that should be on such a hypothetical list. Thus, the expert provided us with this list of hypothetical top-10 artists by category and we considered it as ground truth. We define precision as the number of identified artists in the resulting list that are also present in the ground truth list divided by the length of the list. We evaluated the output of the two algorithms by calculating precision over the entire list (top-10), and over the first five elements (top-5) (see Table 3). We observed that PageRank results (see Table 2) show the greatest agreement with the flamenco expert. High values of precision, specially for the top-5 list, indicates that the content gathered in FlaBase is

highly complete and accurate (see Table 3).

Cantaor	Guitarist	*Bailaor*
Antonio Mairena	Paco de Lucía	Antonio Ruiz Soler
Manolo Caracol	Ramón Montoya	Rosario
La Niña de los Peines	Niño Ricardo	Antonio Gades
Antonio Chacón	Manolo Sanlúcar	Mario Maya
Camarón de la Isla	Sabicas	Carmen Amaya

Table 2. PageRank Top-5 artists by category

	Top-5	Top-10
PageRank	0.933	0.633
HITS Authority	0.6	0.4

Table 3. Precision values

7.2 Statistics

For the sake of completeness, some statistics on the data stored in FlaBase were calculated. Data shown in Figure 3 was produced out of the entity resolution process, while data shown in Figures 4 and 5 was calculated according to the populated data. In Figure 3 it is shown that the most representative *palos* are represented in the knowledge base, with a higher predominance of fandangos. We can observe in Figure 4 that most flamenco artists are from the Andalusian provinces of Seville and Cadiz. Finally, in Figure 5 we observe a higher number of artists in the data were born from the 30's to the 80's of the 20th century.

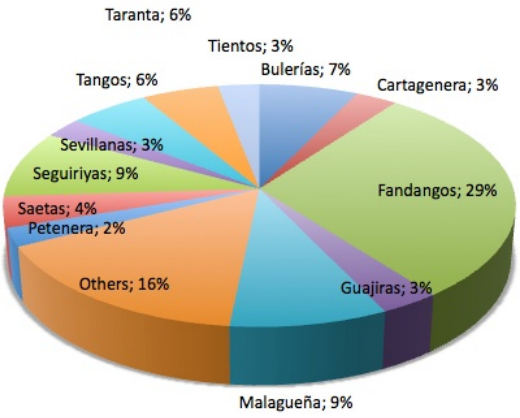

Figure 3. Songs by *palo*

8. CONCLUSIONS AND FUTURE WORK

A new knowledge base that contains information about flamenco music has been created and released. A process of automatic knowledge curation has been applied to combine information coming from different data sources. In addition, the knowledge base has been enriched with content extracted directly from texts by using a custom entity linking system. Using FlaBase data, artist relevance has been computed and compared to the flamenco experts' judgment. Precision values obtained reveals a high degree of coverage and a good quality of the knowledge base content.

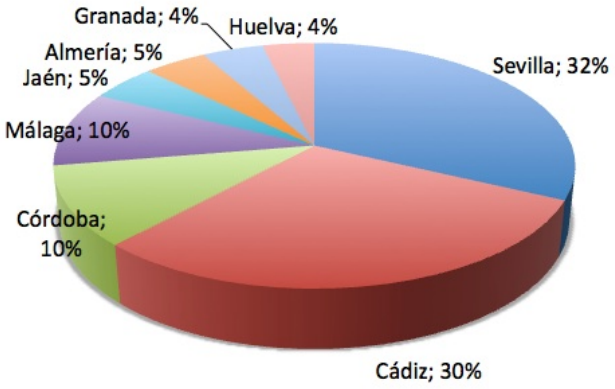

Figure 4. Artists by province of birth

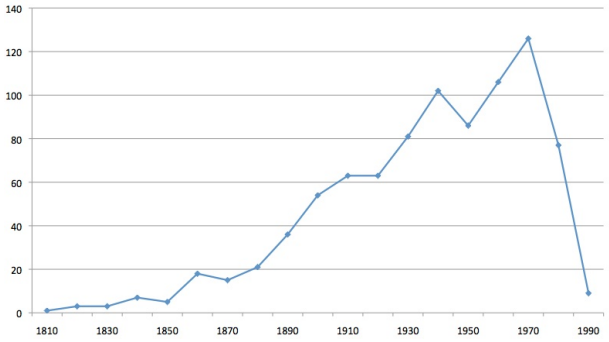

Figure 5. Artists by decade of birth

There are still many avenues to be explored for future work. More websites can be exploited to increase coverage. The entity resolution step might be improved by increasing the amount of entity labels used, or by applying learning algorithms. A SPARQL endpoint might be created, letting users query FlaBase directly. In addition, implementing a collaborative environment for knowledge management would lead to an improvement in terms of completeness and data accuracy, as content might be added, checked and corrected directly by a community of users.

9. ACKNOWLEDGMENTS

This work was funded by the COFLA2 research project (Proyectos de Excelencia de la Junta de Andaluca, FEDER P12-TIC-1362) and the SIGMUS research project (TIN2012-36650). We thank Rafael Infante and José Ruiz Fuentes for the provided content.

10. REFERENCES

[1] Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open Information Extraction from the Web. In *International Joint Conferences on Artificial Intelligence*, pages 2670–2676, 2007.

[2] Francesco Bellomi and Roberto Bonato. Network

Analysis for Wikipedia. *Proceedings of Wikimania*, 2005.

[3] Jose Blas Vega and Manuel Ríos Ruiz. *Diccionario enciclopédico ilustrado del flamenco*. Cinterco, Madrid, 1988.

[4] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 30:107–117, 1998.

[5] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. Toward an Architecture for Never-Ending Language Learning. *In AAAI*, 2010.

[6] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Open question answering over curated and extracted knowledge bases. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, pages 1156–1165, 2014.

[7] Lola Fernández. *Teoría musical del flamenco*. Acordes Concert, Madrid, 2004.

[8] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, 2005.

[9] J. M. Gamboa. *Una historia del flamenco*. Espasa-Calpe, Madrid, 2005.

[10] Lise Getoor. Entity Resolution: Theory, Practice & Open Challenges. *Tutorial at AAAI-12*, pages 2018–2019, 2012.

[11] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46:604–632, 1999.

[12] J.L. Navarro and M. Ropero. *Historia del flamenco*. Ed. Tartessos, Sevilla, 1995.

[13] Sergio Oramas. Harvesting and Structuring Social Data in Music Information Retrieval. *Extended Semantic Web Conference (ESWC). Lecture Notes in Computer Science*, 8465:817–826, 2014.

[14] M Cristina Pattuelli, Matt Miller, Leanora Lange, Sean Fitzell, and Carolyn Li-Madeo. Crafting Linked Open Data for Cultural Heritage: Mapping and Curation Tools for the Linked Jazz Project. *Code4Lib Journal*, page 4, 2013.

[15] Alastair Porter, Mohamed Sordo, and Xavier Serra. Dunya: A System for Browsing Audio Music Collections Exploiting Cultural Context. *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, 2013.

[16] John W Ratcliff and David Metzener. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13:46–72, 1988.

[17] Markus Schedl, Emilia Gómez, and Julián Urbano. Music Information Retrieval: Recent Developments and Applications. *Foundations and Trends® in Information Retrieval*, 8(2-3):127–261, 2014.

[18] Xavier Serra. Data gathering for a culture specific approach in MIR. *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, page 867, 2012.

[19] Xavier Serra. Creating Research Corpora for the Computational Study of Music : the case of the CompMusic Project. *53rd International Conference: Semantic Audio (January 2014)*, pages 1–9, 2014.

[20] Ricardo Usbeck, Axel-cyrille Ngonga Ngomo, R Michael, Daniel Gerber, Sandro Athaide Coelho, and Andreas Both. AGDISTIS - Graph-Based Disambiguation of Named Entities using Linked Data. *The Semantic Web – ISWC 2014*, 2014.