MUSICBRAINZ FOR THE WORLD: THE CHILEAN EXPERIENCE

Gabriel Vigliensoni[1], John Ashley Burgoyne[2], and Ichiro Fujinaga[1]

[1] CIRMMT
McGill University
Canada
[gabriel,ich]@music.mcgill.ca

[2] ILLC
University of Amsterdam
The Netherlands
j.a.burgoyne@uva.nl

ABSTRACT

In this paper we present our research in gathering data from several semi-structured collections of cultural heritage—Chilean music-related websites—and uploading the data into an open-source music database, where the data can be easily searched, discovered, and interlinked. This paper also reviews the characteristics of four user-contributed, music metadatabases (MusicBrainz, Discogs, MusicMoz, and FreeDB), and explains why we chose MusicBrainz as the repository for our data. We also explain how we collected data from the five most important sources of Chilean music-related data, and we give details about the context, design, and results of an experiment for artist name comparison to verify which of the artists that we have in our database exist in the MusicBrainz database already. Although it represents a single case study, we believe this information will be of great help to other MIR researchers who are trying to design their own studies of world music.

1. INTRODUCTION

Thousands of commercial and non-commercial websites offer information and metadata about different aspects of music and artists, for example, their recordings, biographies, discographies, video clips, and other resources. However, the data they provide is often disorganized and not interlinked, and the websites disappear frequently. Hence, it is likely that the information collected over years can be lost. It seems sensible to gather all music-related data in a centralized database that can be accessed by several websites or systems. The goal of our project is to take data from several semi-organized collections of Chilean-music cultural heritage—websites and databases which combined represent almost all music that has been composed and performed in Chile—and integrate it into an open-source music database, where information is easy to search and will last for a longer time.

1.1 Music Metadata

Metadata is structured information that identifies, describes, locates, relates, and expresses several, different layers of data about an information resource [3, 5, 8]. It can be of three basic types: *descriptive*, for purposes such as identification and discovery; *structural*, for expressing relations among resources; and *administrative*, for managing resources [8]. Descriptive music metadata commonly provides information about recordings, expressing the song title and length, the artist name, and the release name of a musical object, usually stored in a MP3 ID3 tag. Structural music metadata is used to document relationships within and among digital musical objects to allow navigation, such as the song's order in an album or a playlist, linking song names to video clips, artists to their biographies, and so on. Administrative music metadata can include information such as software and hardware used to digitize the musical resource, system requirements to read the files, use restrictions or license agreements that constrain the use of the resource. Thus, music metadata has been called a digital music commodity because it adds value to the musical objects, mediating the experience of listeners with music and artists, helping them to browse, explore, sort, collect, use, and finally enjoy music [4, 7].

For this project, we collected data about Chilean music from several websites and databases of different scope and size. [1] Centralizing and interlinking these resources should create synergies in the data because one single query will give access to all resources, creating relations that were not established in their previous locations, thus contributing to a larger web of data. Hence, finding the best repository capable in storing and expressing these data relationships was considered significant and crucial for the success of our endeavour.

1.2 Music-Related Metadatabases

Although there are many commercial, professionally reviewed, online music libraries, we focused only on user-built, open-data, music metadatabases. By contrast to commercial libraries, open databases provide user-access to enter data and have similar types of licenses for the use of its data as the Chilean websites. We present now a list of the main services available for free, public use in terms of their

[1] Data from all databases available at
http://www.vigliensoni.com/McGill/BDMC/8_DATA

	Scope	Size	Information resources stored	Relations	IDs	Data quality assurance	Language guidelines	API	Other
FreeDB	CD tracklists	∼2M albums	Album, title, year, genre, time offsets	Disc to genre	FreeDB propriety Disc-ID	Unspecified automatic method	No	No	Limited search engine. Small, fixed set of genres. Two artists can have the same ID.
MusicMoz	Music-related factual data and Internet links	>136K items	Artist's biographies, discographies, profile, reviews, and articles. Music-related links and resources	No relations allowed	Artist name-based URI	MusicMoz's authorized editors	No	No	Many links are no longer available. Unusual ontology of musical categories.
Discogs	Physical discographies and music releases	>2M artists >3M albums	Artist, release, master, label, image	Small set of relations	Artist name-based URI	Community-based high-quality data	English only	RESTful, XML-based API	30s previews. Well designed solution for artist name variation. Marketplace available. Could evolve to a commercial site.
MusicBrainz	Any kind of music release	>600K artists >1M releases >11M tracks	*Core entities* (artist, label, recording, release, release-group, work), entities, and their relationships	*Advanced Relationships* can be expressed for all core and external entities	MB IDs are universally unique identifier (UUID)	Community-based high-quality data	Guidelines for 30 languages	RESTful, XML-based API	MB links data from all other metadatabases. MB's *Advanced Relationships* can be mapped into RDF. MB stores acoustic fingerprints.

Table 1. Comparison of free, user-built, open-data, music metadatabases: *FreeDB*, *MusicMoz*, *Discogs*, and *MusicBrainz*.

scope, size, stored information resources, relations among these resources, ID system, data quality and assurance, language guidelines, and API.

FreeDB is a license-free database of CDs track-listings from user-contributed data, originally based on the *Compact Disc Database* (CDDB). [2]

MusicMoz is a user-contributed database that stores music-related, factual data and Internet links. [3]

Discogs is a large, user-populated database of discographies and music releases from physical sources. It provides data of high quality that is ensured by a strict input form mechanism and a large community of users. [4]

MusicBrainz is a large, community-based, user-contributed metadatabase that stores the three aforementioned types of metadata for any kind of music release. Its high-quality database is managed by an open community of non-professionals that negotiate periodically and consistently, with strict standards and routines, about the orientations, developments, style guidelines, and mostly everything on MusicBrainz [4]. [5]

Table 1 shows a comparison of the four metadatabases. It can be seen that, among all these databases, MusicBrainz is the database with the broadest scope, not being restricted to only physical copies, as in the case of Discogs, or CDs, as in FreeDB. Also, MusicBrainz is the only music database capable of storing not only descriptive, but also structural and administrative metadata. This fact is critical for expressing the many relationships among musical resources, as well as for providing efficient ways of managing and interlinking them. For example, the MusicBrainz universal unique identifier-based IDs (MBIDs) practically ensure unique identifiers for all its *core entities* (i.e., artist, label, recording, release, release-group, and work), and they have been already widely used for linking music data in the semantic-web community. Furthermore, the MusicBrainz community decided to develop a set *Advanced Relationships* to describe relations between its core entities, and to publish all the MusicBrainz database, resources and their relationships, as Linked Data. The recently-finished *LinkedBrainz* [6] subproject was intended to map all these relationships into RDF [6]. MusicBrainz also stores acoustic fingerprints (PUIDs) that can be used to search for a resource even if its descriptive metadata is not available. Users, as well as performance right organizations, could benefit from this feature for music rights identification purposes. Moreover, MusicBrainz has also a strict, language-specific, style guidelines for 30 different languages, which explain how data should be formatted in their database. This fact is very important to this project—and also to other similar projects coming from countries and regions where English is not the primary language—because it allows, and forces, the non-English-speaking users to follow the correct standard for their language. Finally, MusicBrainz is the most "open" of the four reviewed metadatabases because it provides methods to link data from other websites and databases to MusicBrainz (e.g., by extracting CD track lists from FreeDB, linking artists' images from Discogs, data from MusicMoz, *Allmusic*, *BBC Music* and *Wikipedia*; or album covers and videos from *Amazon* and *YouTube*, respectively).

[2] http://www.freedb.org/
[3] http://www.musicmoz.org/
[4] http://www.discogs.com/
[5] http://www.musicbrainz.org/
[6] http://wiki.musicbrainz.org/LinkedBrainz

By analyzing all the aforementioned characteristics, we decided to work with the MusicBrainz metadatabase in order to store and make available the corpus of music that we want to work with: Chilean music.

2. COLLECTING DATA ABOUT CHILEAN MUSIC

Over the last ten years, several endeavors for creating websites devoted to Chilean music have been developed. These projects have collected data about artists' discographies, biographies, video clips, and album and concert reviews. Currently, however, there is no way of creating a common query to retrieve all available data for a specific artist, album, or song because these resources are neither centralized nor interlinked, do not have URIs, and the websites depend on scarce sources of funding, and so sometimes they have to shut down.

2.1 Websites and Databases of Chilean Music

During the past few years, Chilean music-related metadata has been accessible through the following five major websites:

Base de datos de la Música Chilena (BDCH) (the Chilean Music Database) is a restricted digital music database developed by the Sociedad Chilena del Derecho de Autor (SCD, the Chilean Society of Authors), the only performance rights organization in Chile. The BDCH provides all associate radio stations across the country with a secure, fast, and easy-to-use website for accessing and exploring the largest repository of Chilean music. Radio stations can legally download and airplay music from the BDCH. The scope of its collection is wide, ranging from rock, pop, and ballad, to classical, experimental, and jazz. Metadata for each song has been manually generated by the artists or producers themselves and cleaned afterwards by expert annotators. Data includes composer, author, interpreter, album, label, genre, and label. [7]

Musicapopular (MP) is a website developed and maintained by music journalists, advised by a musicologist, and funded by several short-term governmental grants. Its database is updated periodically and their authors have a commitment for data accuracy and quality, which makes MP a good source when looking for information about Chilean music artists. It provides artists' biographies and discographies, the evolution of band members over time, birth and death dates for individuals, and start and end dates for bands. All its data is mostly interlinked, but only within the website. Although MP has a genre taxonomy tailored to Chilean music (e.g., *nueva canción chilena*, *música chilota*, or *proyección folclórica*), the mainstream genres, e.g., rock and pop, comprise the bulk of the database. [8]

Mus (MUS) is a website devoted to new album and concert reviews. It was funded by the SCD and short-term governmental funds and was maintained periodically by music journalists. Although the website was shut down on January 2012, its data can still be accessed with the exact URLs of the resources. The site does not provide any search methods.

Portaldisc (PD) is a web portal where Chilean music of many genres is sold. PD shares efforts with the SCD as well as MP, selling most of the catalog belonging to and reviewed in those collections. Its website provides access to audio previews of all songs as well as short album reviews. PD can be searched only by artist name. [9]

Videoclipchileno (VCCL) is a website with a nearly comprehensive collection of Chilean music video clips. It provides not only links to the video clips and their directors, but also contextual information about them. It is maintained periodically by the same group of music journalists that run MP, and they are waiting for new governmental funding to improve the scope and functionality of the site. VCCL's search engine accepts a keyword that is searched across all fields in their database. [10]

Most data retrieved from the aforementioned websites can be represented explicitly with the MusicBrainz metadata schema. Also, by means of creating advanced relationships between resources from different sources, these can be linked and accessed from the MusicBrainz web page or API.

2.2 Data Harvesting and Parsing

Because most of the surveyed websites depend on external sources of funding, they can be short-lived and their data—and people's work behind it—can be lost. For that reason, we decided to harvest all data from the available websites.

Although the nature of the data in all databases was different, it can be combined or its overlap can be used to extract more accurate data. For example, the album reviews can provide the track list, and so they can be used to obtain the actual list of songs for an specific album, or to help to disambiguate any difference in the track lists. Table 2 shows the data types and approximate number of entries we extracted by scraping the websites. [11]

There were some problems when scraping, parsing, and storing the data from the websites.The first problem was handling all non-ASCII characters. These characters typically arise because most of the words are written in Spanish, but also sometimes because of special characters that artists use for their names or for the names of their songs.

[7] http://bdch.musica.cl/web_bdch/
[8] http://www.musicapopular.cl/
[8] http://www.mus.cl/
[9] http://www.portaldisc.cl/
[10] http://www.vccl.tv/
[11] Code and scripts available at http://www.github.com/vigliensoni/bbdd_much/

Database	Data retrieved
BDCH	40,000 songs
	33,000 different songs
	3,300 artists
	3,000 albums
	400 record labels
	80 genres
MP	1,500 bands
	1,800 individuals
	1,800 biographies
	40 genres
MUS	500 albums
	300 interviews
	600 concerts
PD	3,600 album reviews
VCCL	1,600 video clips

Table 2. Approximate collection sizes and data types within the five major Chilean music databases.

We ended up using Unicode for representing all data internally. Another problem we faced was the many variations that a resource name can have across repositories, releases, or even within the same website. For example, different people can use alternative forms of an artist name (e.g., "Dj Bitman", "DJ Bitman", "Dj Bit Man", and so on). Also, artists themselves can use variations of their names across several releases (e.g., "Bitman" and "DJ Bitman"). Finally, many slight variations of the same resource name can exist across different repositories due to human error when entering the data. Resolving these inconsistencies can be a tricky problem, for example, as in recent work from Angeles et al. [1] where they combined a metadata manager software with fingerprinting-based querying and still obtained a low rate of consistency between resource names among different databases. Libraries' practice of using authority files for identification and disambiguation of catalog names is able to deal with these inconsistencies, however, most Chilean artist names are still not cataloged in institutional databases. This project tries to collect data from several sources, disambiguate name variations, and store the data into a single searchable database. In Section 3.1 we will detail a string matching-based experiment that helped us take a step toward solving this problem.

3. DATA MATCHING WITH MUSICBRAINZ

After we had consolidated all data, we wanted to know how many entries were (and were not) already in the MusicBrainz database. Using the MusicBrainz web services, we proceeded to compare our data with MusicBrainz and obtained 27, 23, and 21 percent of matches for artists, albums, and songs, respectively. However, among the artists, we retrieved a large number that were false positives (i.e., wrongly recognized as the queried artist when they did not correspond with the actual Chilean artist). In order to reduce these inconsistencies, we decided to add constraints on the resulting resources. The country name alone seemed a good predictor for fixing this problem, but we were discouraged because currently only 22 percent of the total number of artists in MusicBrainz have a country name assigned. Even worse, among them there are only 200 artists with a *CL* country-code value, the ISO 3166-1 Alpha-2 code for Chile. A second problem we found was that sometimes the actual, true positive result retrieved by the Lucene-based MusicBrainz search server was not the one in the first position or the one with the highest score. We realized that we would need to iterate over all retrieved results and compare the strings in order to see which one was the proper match.

To improve the number of true positives for our query, we designed and ran an experiment considering the advanced search method that MusicBrainz provides and the two aforementioned issues. The objective of the experiment was to determine an ideal threshold that allows us to have the largest precision and recall for the artist names. That led us to three questions:

- How many artists have an exact match (i.e., they are already in the MusicBrainz database)?
- How many artists do not match (i.e., they are not in the MusicBrainz database)?
- How many artists match partially? Among these we need to see what is the best threshold to obtain the largest precision and recall.

3.1 Approaches to Artist Name Matching

Our approach for the experiment was two-fold. On the one hand, we created a query that consisted of the artist name plus the country code; we also looked for any comment with the word *Chile* in the annotation field for artist disambiguation (i.e., MusicBrainz's own method for disambiguating similar artist names in its database). On the other hand, we hypothesized that for selecting the proper string from all the retrieved results, we could rely on measuring the string difference between the query string and each one of the retrieved results: the one with the smallest difference would be the true positive. Thus, to handle all nuances or variations of the strings due to special Spanish characters or typographical errors when the artist name was entered, we implemented two string metrics with three variations each:

Levenshtein distance (L) permitted us to calculate the cost of the best sequence of edits to convert one string into the other. We used it in its ratio form to consider the number of letters of the query, so we obtained a normalized value, where 1 represents an exact match.

Jaro metric (J) also allowed us to calculate a normalized value that represents the number of edits needed to convert one string into the other, but it further weights positively or negatively if source characters are or are not present in the target string [2].

Normal string (N) was a direct comparison of the original strings.

ASCII-fied string (A) allowed us to compare ASCII-fied versions of the strings where all tilde and accents were removed from the strings.

ASCII-fied, lower-cased, no-space string (P) allowed us to compare strings where all characters were ASCII-fied and lower-cased, and from which the spaces were also removed.

Hence, we ended up with the *LN*, *LA*, *LP*, *JN*, *JA*, and *JP* variations of the similarity of the artist name query string against each one of the retrieved artist names by the MusicBrainz web service. It should be noted that in all cases the Lucene special characters were escaped before doing the query, and so if an artist name had any of these characters, it was not considered.

3.2 Experimental Procedure

For the experiment, we designed a testing subset with artist names randomly chosen from the ones in our dataset. This subset was created among those artists with string distances between the range of 0.75 and 1.00. The size of the dataset was 800 entries, which represents roughly a quarter of the total number of artists in our dataset. We manually searched to see if the artists in the dataset already existed in the MusicBrainz database. This process was long and complex because there were many false positives among those artists with common and short names, such as *Rachel*, *Quorum*, *Criminal*, *Twilight*, and many others. The only way to determine if the query returned a true or false positive was searching in the artist's country, releases, relationships, or works, and seeing if anything there matched some data in our database. We then ran the query for each entry, chose the items with the largest metric values, and identified true and false positives, and false negatives. We then calculated precision and recall for the whole subset at different thresholds, and calculated the error for each metric and threshold using the bootstrapping technique. Figure 1 shows the precision and recall for each threshold, metric, and variant. Error curves at α = 0.05 generated from a bootstrap sample of 1,000 replications of the original sample are also shown. These plots have to be understood as a series of two complementary runs. The plots in the left refer to precision (upper left) and recall (lower left) for the comparison of strings only. In this case, all retrieved artists with the same name were considered as a true positive, even if they are not in fact the same artist. This approach was taken because none of the six metrics can know in advance whether the retrieved name denotes the specific person that we are looking for. We can observe that if we were to be too strict with the string distance threshold, the precision would be high but the recall would be low, as expected.

In terms of precision, it can be seen that Levenshtein string distance (L) in its variants for the normal string (N), its ASCII-fied version (A), and the ASCII-fied, lower-cased, and no-spaced string version (P) performs better than the Jaro (J) distance for thresholds between the range of 0.80 and 0.90. The three different variants (N, A, P) do not make statistically significant differences in the results for any of the methods. However, in terms of recall, it can be seen that the P versions of both L and J are the ones with the best recall, especially when the thresholds become higher. There is no statistically significant difference between the performance of P and A, but at high thresholds, N statistically significantly underperforms both P and A. We can conclude that for the best results, the queried string should at least be ASCII-fied and the Levenshtein distance should be used. The threshold point to obtain the best precision while still having a good recall can be found around 0.90.

The two plots in the right of Figure 1 refer to the same experiment and dataset but evaluated with respect to their real-life context. In other words, in this second case, if the query returned an artist with the same name but it referred to another artist in real life, the returned artist was considered as a false instead of a true positive. Under these conditions, we reached a ceiling of precision at about 60 percent, which establishes that some kind of verification process is essential when using name matching in a real-world application.

4. CONCLUSIONS AND FUTURE WORK

We have done a review of user-contributed, music metadata libraries, and we have shown why we have chosen MusicBrainz as the metadatabase that we will use. We have also shown the five websites that we used for collecting Chilean music-related metadata and given details about the data that each website provides.

When we tried to combine our database with MusicBrainz, we realized that there was a large amount of false-positive noise in the results for each query. We tried several types of queries and determined that an advanced search method, including several fields at the same time, was required, and also that it was necessary to iterate over all results and perform a string comparison between the query and the retrieved query to look for the true result. Hence, we developed an experiment whereby we created a random ground-truth subset of artist names from our database. We then queried these same entries with MusicBrainz, compared the retrieved results with the ground truth, and then tried to define the optimum variation threshold that should be accepted between the queried and the retrieved string in order to obtain the best precision and recall. We compared two different metrics with three variations each. For precision, the Levenshtein ratio offered the best performance, stabilizing its curve close to a normalized value of 0.90, where 1.00 means that the two strings are identical. The three variations did not matter. For recall, however, we established that the strings should, at least, be ASCII-fied to obtain a better recall. Overall, the best string comparison threshold on a normalized scale is close to 0.90, and the method used should be Levenshtein distance with an ASCII-fied, lower-cased, no-spaced version of the strings.

Short-term future work for this project will be to apply these experimentally obtained values for the string comparison across albums and songs. However, it is expected that the results will be much better because we will know

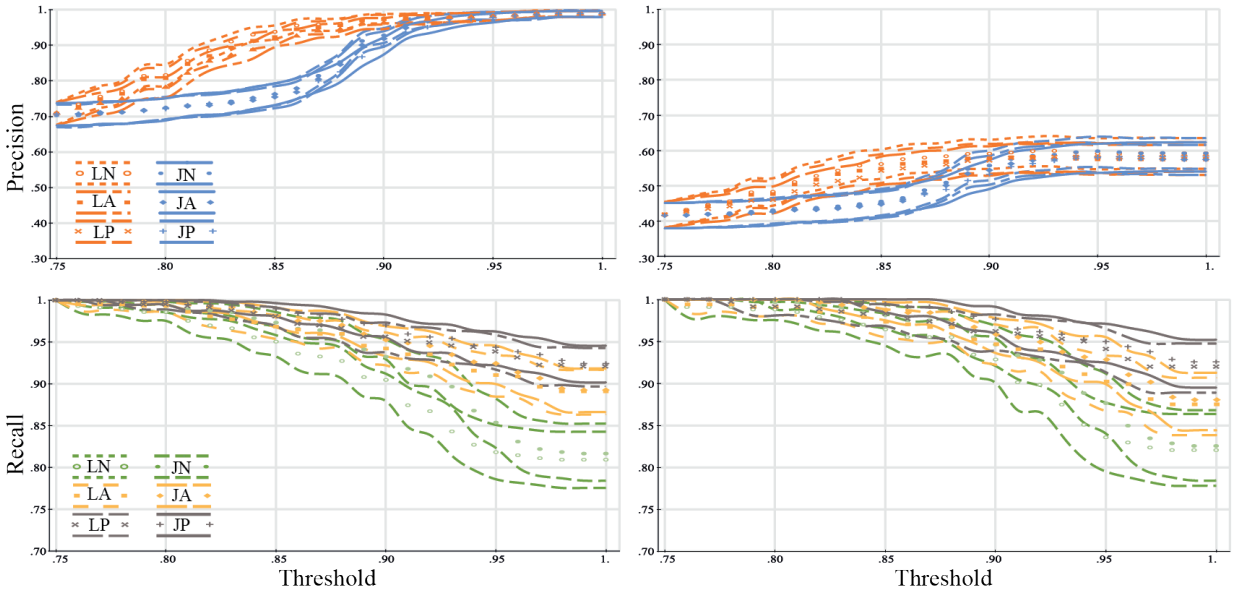

Figure 1. Precision and recall for artist names string comparison between the consolidated data from Chilean music databases and MusicBrainz. While upper and lower left plots show the results for the raw string comparison using Levenshtein (L) and Jaro (J) string distances and three variants (N, A, P), the ones in the right show the results in real-life context, where false positives were discarded. Error curves show upper and lower limits for 1,000 populations replicated from the original sample using bootstrap at $\alpha = 0.05$.

beforehand if the album's artist or song's artist is already in the MusicBrainz database. For future work in the mid-term, we will enter all data we collect into the MusicBrainz database. The MusicBrainz API does not allow one to automate this kind of process, but we experimented using a POST method to fill the forms automatically, and a user would simply need to review and submit the data to the database. By these means, all the data we collect will be available from MusicBrainz. As a second mid-term project, and by asking permission to the SCD, it would also be possible to use the audio, thereby allowing audio analysis over the whole corpus of music. This kind of analysis would be beneficial for everyone in the music-industry chain. For example, running structural music analysis to know where the choruses are in a given song, or knowing the overall tempo of a song, would be of benefit for the radio stations that use the BDCH, or the general public interested in creating a remix of a given song. In the long-term, the experience gained and the techniques developed will be applicable to other countries and cultures, to enable them to merge their metadata to global central databases.

5. ACKNOWLEDGEMENTS

This research was supported by the Social Sciences and Humanities Research Council, and by BecasChile Bicentenario, CONICYT (Comisión Nacional de Ciencia y Tecnología), Gobierno de Chile. The authors would like to thank Alastair Porter for sharing valuable knowledge about the MusicBrainz web services and API, and to the *Sociedad Chilena del Derecho de Autor* (SCD) for granting us access to their database.

6. REFERENCES

[1] Angeles, B., C. McKay, and I. Fujinaga. 2010. Discovering metadata inconsistencies. *Proceedings of the International Society for Music Information Retrieval Conference*. 195–200.

[2] Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. 2003. Adaptive name matching in information integration. *Intelligent Systems, IEEE* 18 (5): 16–23.

[3] Dublin Core Metadata Initiative. Retrieved April 6 2013, from http://dublincore.org/metadata-basics/.

[4] J. Hemerly. 2011. Making metadata: The case of MusicBrainz. *SSRN eLibrary* 1982823.

[5] International Federation of Library Associations and Institutions. *Digital Libraries: Metadata Resources*. Retrieved April 6 2013, from http://archive.ifla.org/II/metadata.htm.

[6] Jacobson, K., S. Dixon, and M. Sandler. 2010. LinkedBrainz: providing the MusicBrainz Next Generation Schema as Linked Data. *Late-breaking demo session at the 11th International Society for Music Information Retrieval Conference*.

[7] J. W. Morris. 2012. Making music behave: Metadata and the digital music commodity. *New Media & Society* 14 (5):850–66

[8] National Information Standards Organization. 2004. *Understanding metadata*. NISO Press, Bethesda, MD, USA.