MUSICBRAINZ FOR THE WORLD: THE CHILEAN EXPERIENCE

Gabriel Vigliensoni
gabriel@music.mcgill.ca

John Ashley Burgoyne
j.a.burgoyne@uva.nl

Ichiro Fujinaga
ich@music.mcgill.ca

Aim

In this paper we present our research in:

- Gathering data from several semi-structured collections of cultural heritage– Chilean music-related websites
- Store the data into a single, open-source music database, where the data can be easily searched, discovered, and interlinked
- Disambiguate name variations

Websites and Databases of Chilean Music

Several endeavours for creating websites devoted to Chilean music have been developed in the last ten years. These projects have collected data about artists' discographies, biographies, video clips, and album and concert reviews. However, there are two main problems:

- All websites and databases depend on external sources of funding, and so they can be short-lived
- There is no way of creating a common query to retrieve all available data because these resources are neither centralised nor interlinked

Database	Data retrieved
BDCH Base de Datos de la Música Chilena	40,000 songs (33,000 different), 3,300 artists, 3,000 albums, 400 record labels, 80 genres
MusicaPopular.cl	1,500 bands, 1,800 individuals, 1,800 biographies, 40 genres
MUS.cl	500 album reviews, 300 interviews, 600 concerts reviews
PORTALDISC.COM	3,600 album reviews
VCCL	1,600 video clips

Table 1. Approximate collection sizes and data types within the five major Chilean music databases

Music-related Metadatabases

We focused only on user-built, open-data, music metadatabases because they provide user-access to enter data and have similar types of license for the use of its data as the Chilean websites. We were interested in comparing the same set of attributes in the four available metadatabases:


freeDB.org   MUSICMOZ   Discogs   MusicBrainz
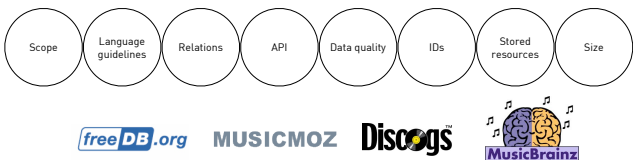
We finally chose MusicBrainz because of its:

- Broadest scope
- API
- UUID for all core entities
- Ability for expressing complex relationships
- Support for different languages
- Acoustic fingerprint capabilities

Data Matching with Musicbrainz

Problems

When comparing the data extracted from the Chilean databases and the entries already in MB we obtained:

- 27, 23, and 21 percent of matches for artists, albums, and songs, respectively
- Large number of false positives due to duplicated names and name variations
- The actual true positive was not always in the first retrieved position
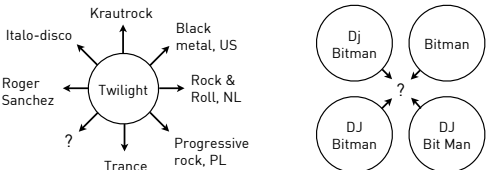
For improving the disambiguation problems, we tried a two-fold approach:

- Creating an advanced query, including several fields at the same time (*CL* as the country code as well as the word *Chile* in the MB artist disambiguation field)
- Determining the best metric and variation threshold when comparing the query string and the results retrieved from MB. We hypothesized that the one with the smallest difference would be the true positive.

Experiment

We designed an experiment to determine the optimum variation threshold that should be accepted to obtain the best precision and recall in artist name comparison:

- We created a random ground-truth subset of artist names (800 entries)
- We manually checked if the artist already existed in MB, by looking at *any* data that might help to disambiguate it.
- We used two different metrics and three variants each for comparing the strings:
  - **L** (Levenshtein) and **J** (Jaro) string distances
  - **N** (no string processing), **A** (ASCII-fied version), and **P** (ASCII-fied, lower-cased, no-spaced version)
- We ran the query for each entry, calculating precision and recall at different thresholds, for each distance and variant

Results

- Precision: **L** ratio offered the best performance, stabilizing its curve close to a normalized value of 0.90. The variations **N, A, P** did not have any statistically significant effect.
- Recall: 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 **L** distance with **P** variant.
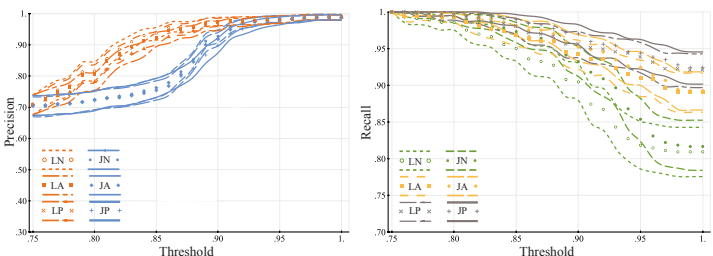

Figure 1. Precision and recall for string comparison between data from Chilean music databases and MB. Error curves generated from a bootstrap sample of 1,000 replications (α=0.05)

Acknowledgements

Centre for Interdisciplinary Research in Music Media and Technology   McGill   DDMAL   DISTRIBUTED DIGITAL MUSIC ARCHIVES & LIBRARIES LAB   Social Sciences and Humanities Research Council of Canada   Conseil de recherches en sciences humaines du Canada   Canada