GRAIL: A MUSIC IDENTITY SPACE COLLECTION AND API

Michael Barone[1], Kurt Dacosta[1], Gabriel Vigliensoni[2], Matthew Woolhouse[1]
Digital Music Lab, McMaster University[1], Hamilton, Canada
CIRMMT, McGill University[2], Montreal, Canada
baronem@mcmaster.ca, dacostak@mcmaster.ca, gabriel@music.mcgill.ca, woolhouse@mcmaster.ca

ABSTRACT

Services such as MusicBrainz [1], Echonest [2], Last.fm [3], and Mus-ixMatch [4] provide valuable access to their growing corpora of music information through the use of pubicly available application programming interfaces (APIs). Although useful to the research community, information collected, and organized by these services can be under-utilized due to independent identity space (ID) generation practices. As a result, there is a scarcity of accurate ID matches at the track level. We introduce GRAIL, an API that links track IDs from diverse online music services using adaptive matching criteria. GRAIL will hopefully increase traffic to these services by enabling developers and researchers to innovate through combining music APIs with relative ease. To our knowledge, there are no organized academic efforts attempting to link track IDs of digital-music services. We describe our scalable architecture and data verification process, as well as explore the challenges in collating digital-music information from disjoint resources.

1. INTRODUCTION

As the Web 2.0 continues to expand and share its rich data sources, unique opportunities for population-level research become increasingly possible for the music-research community. APIs can be a useful resource for collecting data relevant to major music-research topics. While accessing an API is straightforward, many research questions require utilizing more than one data source to innovate and discover. For example, linguists may be interested in relating and comparing patterns within lyrics to user geography using MusixMatch and MusicBrainz data. Cognitive psychologists may wish to examine how consistent audio feature extraction algorithms are between services. And social scientists may wish to investigate global listening trends and social-networks analysis through Last.fm, and Spotify [5].

However, despite this and other research [1] arguably there is a scarcity of interdisciplinary, population-level analysis of music consumption due to a lack of trustworthy linkages between ID spaces of music information-related web-services. The motivation for GRAIL is to facilitate ID space linkage, enabling complex research to be undertaken rapidly.

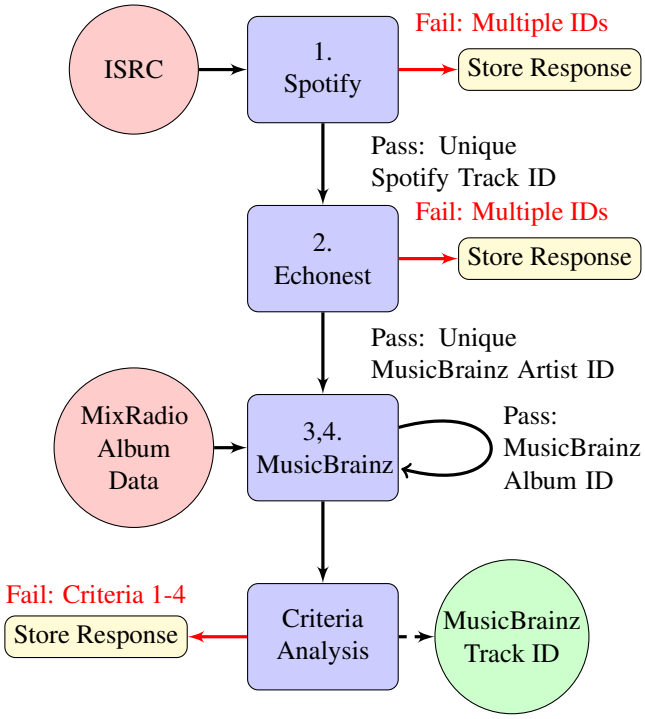

Figure 1. GRAIL ID Space Linkage Workflow

2. DATA LINKING

As part of a 5-year data sharing agreement with MixRadio, ca. 27 million International Standard Recording Codes (ISRCs), linked to MixRadio track IDs, were accessed by our team for research and analysis. ISRCs are usually issued during the mastering process and are considered to be a reliable, and atomic track-level identifier. Tracks retain their ISRC identifier so long as the recording hasn't been edited, remixed, or remastered [6].

[1] https://musicbrainz.org/
[2] http://developer.echonest.com/
[3] www.last.fm/api
[4] https://developer.musixmatch.com/

[5] https://developer.spotify.com/
[6] http://www.isrcmusiccodes.com/

We match track IDs of a number of music APIs through a combination of ISRC linkages with data provided by APIs. Data verification is handled through strict matching criteria, including: i) checks of album-track cardinality, ii) track ordering, and iii) string matching of artist, album, and track titles. ISRCs are matched to Spotify, Echonest, MusicBrainz, MixRadio, Rdio, and MusixMatch IDs at the track, album, and artist levels; our matching process is described below.

2.1 Metadata Ingestion

In Step 1 of Figure 1, Spotify Track IDs are retrieved using ISRCs via the Spotify API. If a unique match is generated, the ISRC and its corresponding Spotify Track IDs are ingested into GRAIL. If several Spotify Track IDs are retrieved, the response is stored for later processing. Second, MusicBrainz Artist IDs are retrieved using Spotify Track IDs via the Echonest API. If a unique match is generated, MusicBrainz Artist IDs are ingested into GRAIL and linked to the corresponding ISRCs. Third, MusicBrainz Album names (strings) are retrieved using MusicBrainz Artist IDs and MixRadio Album names via the MusicBrainz API. The returned MusicBrainz Album ID is ingested into GRAIL only if there is a unique match, and there is agreement in album-track cardinality between MixRadio and MusicBrainz.

2.2 Track Matching Criteria

MixRadio metadata, including track names and ordering are used as the basis by which MusicBrainz IDs are linked to ISRCs. In Step 4, we input successfully linked album IDs back into MusicBrainz to return a series of ordered MusicBrainz Track IDs and verify that track-to-album matches are accurate using 4 criterion levels (Table 2). Criteria 1 and 2 have been ingested into GRAIL, whereas Criteria 3 and 4 have been stored to determine accuracy. For all criteria, tracks maintain case insensitive string matches. Additional crtieria conditions are described in Table 1. These criteria consider track ordering and string cleaning prior to the matching process. Each criterion loosens its restrictions in a step-wise fashion. Criterion 1 contains the strictest requirements, criterion 4 contains the loosest.

By using the aforementioned procedure and matching criteria, more than 11 million ISRCs have been linked to Spotify Track IDs; half a million MusicBrainz Track IDs have been linked to Spotify Track IDs and ISRCs. Current results for music IDs across APIs are shown in Table 2.

3. API DETAILS

Our intention is for the obtained ID linkages to be available as a REST API through the url: *www.digit-almusiclab.org/-grail/*. Users will be able to query the API by track, album, artist names or IDs from a documented list of ID spaces. GRAIL's response will be formatted in JSON or XML. A sample request will have the form: *http://digitalmusiclab-.org/grail/track/id=isrc:**ISRC**&api_key=**API_KEY**&inc=musicbrainz*.

Criteria	Cardinality	Ordering	String Match
1.	1	1	1
2.	1	0	1
3.	1	1	0
4.	1	0	0

Table 1. MusicBrainz track matching criteria. Cardinality and ordering with a value of 1 represents True. String Matching of 1 represents exact matches, whereas 0 represents fuzzy matching. Criterion 1 is considered the strongest match, criterion 4 the weakest.

API ID Space	# Linked to GRAIL
Spotify Track ID	11,454,349
Echonest Track ID	7,340,920
MusicBrainz Track IDs	465,747
MusicBrainz Album ID	163,242
MusicBrainz Release Group ID	96,696
Spotify Artist ID	776,620
Rdio Artist ID	707,620
MusicBrainz Artist ID	203,923
MusixMatch Artist ID	73,459

Table 2. Number of IDs successfully linked to GRAIL using Criteria 1 and 2.

GRAIL will require an API key to access its data through free registration. GRAIL's data will be available only for academic research. Our plan is for API keys to have reasonable rate limits based on server restrictions. In all likelihood, the terms will state that this API is for and by the creative commons, and is designed for research with music information retrieval. Users will not be able to use GRAIL for profit without direct permission of the services used in the application.

4. FUTURE WORK

Future work entails continued development of data-cleaning, the analysis of Criteria 3 and 4, and implementation of new criteria. A common problem in track matching involves albums where track cardinality is in disagreement. In these cases, determining a good match is problematic because groundtruth metadata is unclear. Linking these tracks based on cardinality could be inaccurate, and requires continued refinement of matching procedures going forward. Lastly, verification across multiple APIs will be necessary. Comparing the accuracy of metadata across APIs (such as comparing Spotify to Last.fm track information) is a crucial next step that will highlight the level of agreement across the digital-music industry with respect to the information they provide to customers and researchers.

5. REFERENCES

[1] B. Thierry, D. Ellis, B. Whitman, & P. Lamere "The million song dataset," *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pp. 591–596, 2011.