Discovering Metadata Inconsistencies

Via Fingerprinting Querying and Comparing Local and Remote Metadata

Bruno Angeles Cory McKay Ichiro Fujinaga

CIRMMT • Schulich School of Music • McGill University

INTRODUCTION

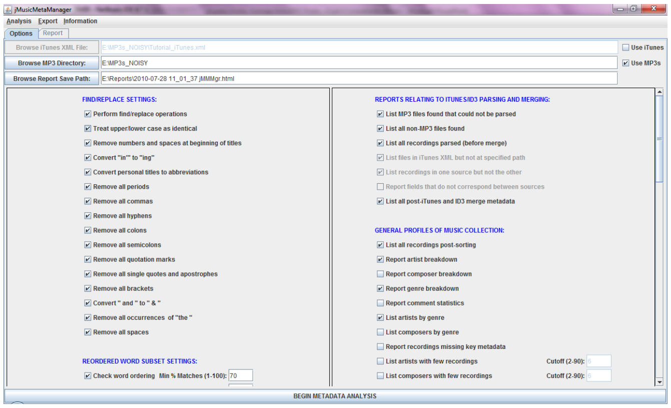
- Difficult to keep metadata consistent in large libraries.
- Existing metadata repositories are very noisy: different contributors, approaches, etc.
- We combine **metadata management software**, **acoustic fingerprinting**, and the **querying of a metadata database** to discover **errors and inconsistencies** in a local music library.
- We compare a library of **manually-maintained music files** (Codaich) with a collection of **uncurated music files** acquired from file sharing services (the reference library).

- Sample **metadata repositories**: MusicBrainz, Discogs, Last.fm, Allmusic, etc.
- Musical **metadata management software**: MusicBrainz Picard, MediaMonkey, jMusicMetaManager, Mp3tag, GNAT, etc.
- Acoustic fingerprinting: associate recordings with a unique key.
- We chose AmpliFIND's **PUID** system (e.g., 1246081f-096f-da6b-a7a6-82ade5ee041c).
- Querying done on a **MusicBrainz** server hosted at McGill University.
- We added PUID-based MusicBrainz querying to **jMusicMetaManager** and improved its support of ID3 tags.

- Experiment performed: we found the percentage of metadata fields (**artist**, **title**, **album**, and **all three**) that were identical between each of our libraries and the MusicBrainz metadata server.

JMUSICMETAMANAGER

- JAVA application that recognizes **metadata inconsistencies and errors** (screenshot below).
- Free, open-source, cross-platform; part of the **jMIR** software suite.
- Handles **multiple valid spellings** for entries: **Стравинский → Stravinsky / Stravinski**.
- Calculates Levenshtein distances between pairs of entries, uses threshold.
- Can remove articles and punctuation, consider abbreviations and word subsets, etc.
- Generates **HTML reports** (see new MusicBrainz Report snippet below).
- Supports ID3 tags and the iTunes XML format.



Browse iTunes XML File: E:\MP3s_NOISY\Tutorial_iTunes.xml □ Use iTunes

Browse MP3 Directory: E:\MP3s_NOISY ☑ Use MP3s

Browse Report Save Path: E:\Reports\2010-07-28 11_01_37 jMMMgr.html

FIND/REPLACE SETTINGS:

☑ Perform find/replace operations
☑ Treat upper/lower case as identical
☑ Remove numbers and spaces at beginning of titles
☑ Convert "in'" to "ing"
☑ Convert personal titles to abbreviations
☑ Remove all periods
☑ Remove all commas
☑ Remove all hyphens
☑ Remove all colons
☑ Remove all semicolons
☑ Remove all quotation marks
☑ Remove all single quotes and apostrophes
☑ Remove all brackets
☑ Convert " and " to " & "
☑ Remove all occurrences of "the "
☑ Remove all spaces

REORDERED WORD SUBSET SETTINGS:

☑ Check word ordering Min % Matches (1-100): 70

REPORTS RELATING TO ITUNES/ID3 PARSING AND MERGING:

☑ List MP3 files found that could not be parsed
☑ List all non-MP3 files found
☑ List all recordings parsed (before merge)
☑ List files in iTunes XML but not at specified path
☑ List recordings in one source but not the other
□ Report fields that do not correspond between sources
☑ List all post-iTunes and ID3 merge metadata

GENERAL PROFILES OF MUSIC COLLECTION:

☑ List all recordings post-sorting
☑ Report artist breakdown
□ Report composer breakdown
☑ Report genre breakdown
□ Report comment statistics
☑ List artists by genre
□ List composers by genre
□ Report recordings missing key metadata
□ List artists with few recordings Cutoff (2-90): 6
□ List composers with few recordings Cutoff (2-90): 6

Codaich

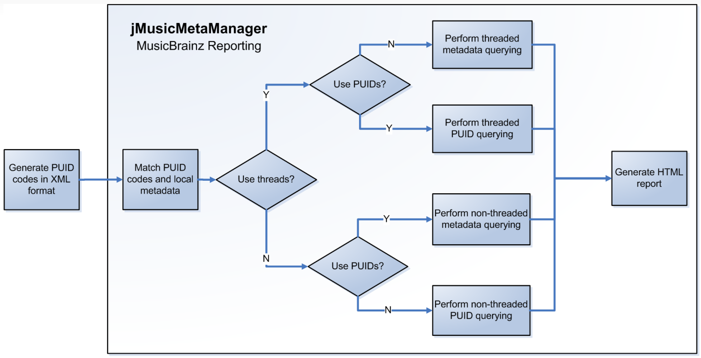
- Curated audio research dataset with **32,328 recordings** (3,000+ artists, 57 musical genres, 19 metadata fields).
- Four sections: **Classical, Jazz, Popular, World**.
- Metadata was cleaned manually and with jMusicMetaManager.

Reference Library

- For contextual comparison with Codaich.
- Unprocessed files collected from file sharing systems.
- **1,363 recordings** (446 artists, 70 genres).
- Files without ID3 metadata: used file names to assign metadata.

METHOD

- We compared metadata fields (**artist**, **title**, **album**, and **all three**) of both Codaich and the reference library with MusicBrainz metadata.
- We analyzed the results by genre: **Classical**, **Jazz**, **Popular**, and **World**.
- To show the advantage of manually-maintained libraries, we also calculated the difference between each library's rate of agreement with the MusicBrainz server.
- Implemented threaded querying to overcome the 1 query per second MusicBrainz limit .



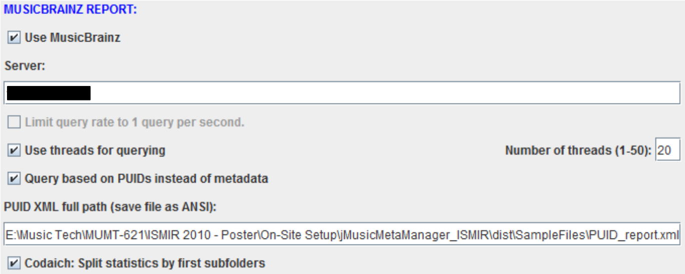
jMusicMetaManager
MusicBrainz Reporting

Generate PUID codes in XML format → Match PUID codes and local metadata → Use threads?

Use threads? Y → Use PUIDs? N → Perform threaded metadata querying; Y → Perform threaded PUID querying

Use threads? N → Use PUIDs? Y → Perform non-threaded metadata querying; N → Perform non-threaded PUID querying

→ Generate HTML report



MUSICBRAINZ REPORT:

☑ Use MusicBrainz

Server: [redacted]

□ Limit query rate to 1 query per second.

☑ Use threads for querying Number of threads (1-50): 20

☑ Query based on PUIDs instead of metadata

PUID XML full path (save file as ANSI): E:\Music Tech\MUMT-621\ISMIR 2010 - Poster\On-Site Setup\jMusicMetaManager_ISMIR\dist\SampleFiles\PUID_report.xml

☑ Codaich: Split statistics by first subfolders

RESULTS AND DISCUSSION

Codaich	Recordings	Artist	Album	Title	All three
Classical	1,476	3%	2%	6%	0%
Jazz	3,179	70%	25%	**64%**	12%
Popular	16,206	**84%**	**52%**	61%	**32%**
World	1,640	58%	29%	46%	11%

Reference Library	Recordings	Artist	Album	Title	All three
Classical	285	17%	0%	5%	0%
Jazz	181	43%	14%	39%	4%
Popular	481	**79%**	**19%**	**51%**	**10%**
World	115	57%	12%	41%	3%

Difference	Artist	Album	Title	All three
Classical	-14%	2%	1%	0%
Jazz	**27%**	11%	**25%**	8%
Popular	5%	**33%**	11%	**22%**
World	2%	17%	6%	9%

Tables above: Percentages indicate the agreement between the test libraries and MusicBrainz, and the difference between the first two tables. This provides a measure of improvement relative to manual maintenance.

- Highest agreement was in Popular music (album, artist, and all three), followed closely by Jazz, possibly because community-based metadata services are driven by musical genres familiar to tech-savvy, young contributors.
- **Codaich:** highest result for titles was in Jazz, possibly due to the curator's knowledge.
- **Reference library:** some ID3v1 tags with 30-character limit led to errors.
- **Difference** shows that manual maintenance improved agreement, except in the case of Classical artists (in Codaich the artist field is used for performer, not composer) and all three fields in Classical (because of challenges such as key, opus number, long subtitles, etc.)
- 2 groups: Classical & World (lowest agreement) vs Jazz & Popular (highest agreement).

Conclusion

- Manual maintenance provides greater agreement with MusicBrainz than unprocessed data.
- Fingerprinting-based querying is particularly useful for Jazz & Popular.
- Must be careful with Classical because the metadata server might not be correct.
- The capabilities of jMusicMetaManager have been enhanced by adding fingerprinting queries.

McGill

Schulich School of Music
École de musique Schulich

CIRMMT Centre for Interdisciplinary Research in Music Media and Technology