Connecting Crowdsourced Audio Recording Metadata with MARC Records

Brian Rennick
Brigham Young University
June 22, 2019
ALCTS CaMMS Catalog Management Interest Group
ALA Annual Washington DC


LET IT BE


SONY
TASCAM
00:01:36
SS-CDR200
SONY

Discogs

Search artists, albums and more...

Explore Marketplace Cor

The Rise and Fall of Ziggy Stardust and
David Bowie

All Release 11,293,961 **Master** Artist 6,113,784 Label 1,337,389

Find Music on Discogs

201 – 250 of 1,564,772 Prev Next

Genre

Rock	503,996
Electronic	469,698
Pop	289,530
Folk, World, & Country	161,782
Jazz	124,509

All

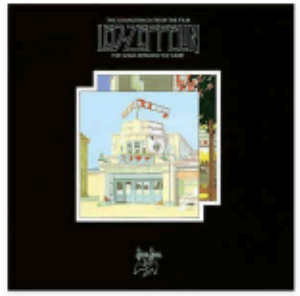
Style

House	76,161
Pop Rock	72,872
Punk	59,299
Vocal	53,428
Techno	53,212

All

Format

Vinyl	909,344
Album	512,990
CD	442,520
LP	382,740



The Soundtrack From...
Led Zeppelin

Waiting For The Sun
The Doors

Innervisions
Stevie Wonder

With The Beatles
The Beatles

Jazz
Queen

Station To Station
David Bowie

More Images

David Bowie – ChangesOneBowie

Genre: Rock

Style: Pop Rock, Soul, Glam

Year: 1976

Notes: This 11 track collection, billed on stickers attached to some editions as Hits", contains album versions for five tracks though a different single v previously been released.

Two of the tracks had previously been released as "b" sides, one of thes was also issued as an "A" side to promote this compilation.

Of the RCA editions, early UK editions contain an alternative 1973 recor Only Dancing", this soon changed to the recording from the 1972 single

Tracklist

Space Oddity
John, I'm Only Dancing
Changes
Ziggy Stardust
Suffragette City
The Jean Genie
Diamond Dogs
Rebel Rebel
Young Americans
Fame
Golden Years

Discogs Fields Used for Matching

```
Title:   ChangesOneBowie
Format:  (LP, Comp)
Label:   RCA Victor
Cat#:    APL1-1732
Country: US
Year:    1976
```

MARC Fields Used for Matching

```
028 02 $a APL1-1732
       $b RCA Victor
245 10 $a Changesonebowie /
       $c David Bowie.
260 ## $a New York, N.Y. :
       $b RCA Victor,
       $c [1976]
650 #0 $a Rock music
       $y 1971-1980.
655 #0 $a Popular music
       $y 1971-1980.
```

Discogs Data Files

- Monthly dump
- Artists, labels, masters, releases
- 45 GB XML releases file
- 14 million releases

Discogs Data Transformation

- Python script
- Filtered out unnecessary tags
- Removed recordings after 1979 and all classical music
- Tab-separated values (TSV) file with 1.2 million records
- Used OpenRefine to identify problem records
- Converted to SQL database

OpenRefine Clustering

- "... finding groups of different values that might be alternative representations of the same thing"
  - "New York" and "new york"
  - "Gödel" and "Godel"
- Key collision methods
- Multiple fingerprinting algorithms

See the OpenRefine wiki for details.

Custom Fingerprinting to Match Records

```
4
5   def filter_stop_words(v):
6       stop_words = ['and', 'are', 'for', 'from', 'its', 'ive', 'las', 'los', 'music',
7                     'records', 'song', 'songs', 'sound', 'sounds', 'the', 'too',
8                     'you', 'your', 'was', 'were', 'with']
9
10      # Assume words less than three characters are stop words.
11      if len(v) < 3: return False
12
13      if(v in stop_words):
14          return False
15      else:
16          return True
17
```

```
18
19  def create_fingerprint(input_string):
20      # Replace non-english characters with english
21      clean_string = unicodedata.normalize('NFKD',
22          input_string).encode('ascii', 'ignore').decode('utf8')
23
24      # Make all words lower case
25      clean_string = clean_string.lower()
26
27      # Remove punctuation
28      trans_table = str.maketrans(string.punctuation, ' '*len(string.punctuation))
29      clean_string = clean_string.translate(trans_table)
30
31      # Convert string to a list
32      clean_list =  clean_string.split()
33
34      # Remove duplicates by converting to a set
35      clean_list = sorted(set(clean_list))
36
37      clean_list = filter(filter_stop_words, clean_list)
38
39      fingerprint = ';'.join(clean_list)
40      return fingerprint
41
```

Method

- Loop through each MARC record.
- For each of the MARC records, query the Discogs SQL database using the record label catalog number as the key.
- Loop through each of the SQL query results to find the best match.

Finding the Best Match

- *Title* fingerprints or *Artist* fingerprints must match.
- *Genre* and *Style* fields must be complete.
- Prioritize by *Country*.
- Write to the error log if no match is found.

Thank you!

brian_rennick@byu.edu
@rennickb