A DATA-CLEANSING FRAMEWORK FOR AGGREGATING ANNOTATED DATASETS FROM MIREX AUTOMATED CHORD ESTIMATION ARCHIVES

Jeff Miller (Presenter), Johan Pauwels, Mark Sandler

Queen Mary University of London

Identification and availability of suitable data sources is a well-known difficulty in music information retrieval research.

This is compounded by:
- Inconsistent presentation formats
- Differences in methodologies
- Differences in complexity of chord vocabularies
- Annotation errors

We were able to repurpose existing data from the MIREX ACE task archives to create a data set of aggregated chord labels by building a framework to apply automated data cleansing and standardization techniques to the MIREX data. We were able to extract a large, labelled chord data set for use in a harmonic modelling study.

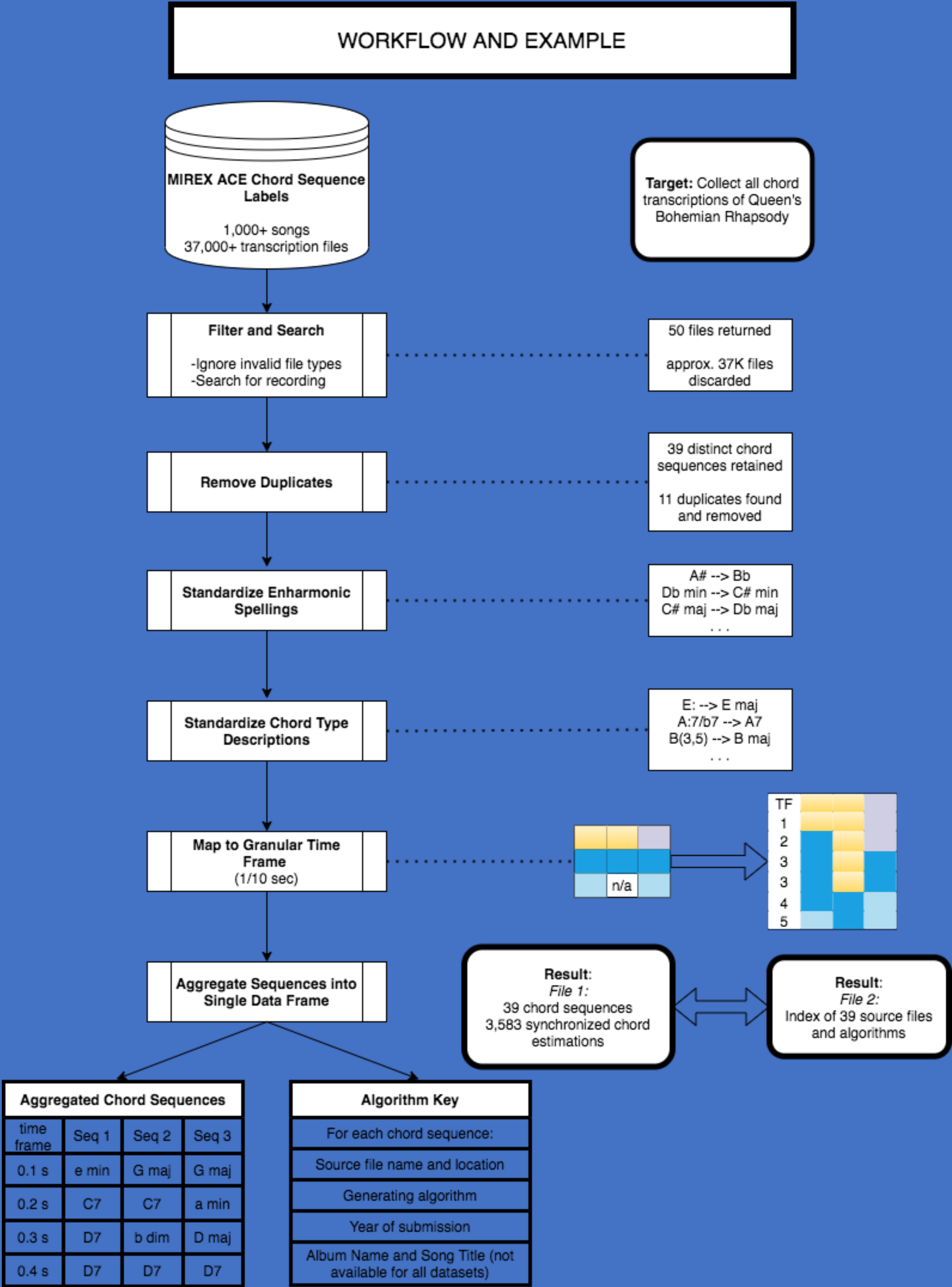Challenges had to be overcome before the MIREX ACE data could be analyzed at scale.
- Standard data cleansing considerations
- Domain-specific logic for text processing of musical chord descriptions.

The standardization of both presentation format and musical descriptions of the annotations made it possible to perform meaningful statistical analysis of:
- Differing chord vocabularies
- Chord distributions
- Discrepancies between various algorithms as they attempted to describe each chord

Automating the process made it possible to quickly modify parameters and reproduce data sets while maintaining data consistency and reducing process errors.

The corresponding author can be contacted at
j.k.miller@qmul.ac.uk

WORKFLOW AND EXAMPLE

MIREX ACE Chord Sequence Labels
1,000+ songs
37,000+ transcription files

Target: Collect all chord transcriptions of Queen's Bohemian Rhapsody

Filter and Search
- Ignore invalid file types
- Search for recording

50 files returned

approx. 37K files discarded

Remove Duplicates

39 distinct chord sequences retained

11 duplicates found and removed

Standardize Enharmonic Spellings

A# --> Bb
Db min --> C# min
C# maj --> Db maj
...

Standardize Chord Type Descriptions

E: --> E maj
A:7/b7 --> A7
B(3,5) --> B maj
...

Map to Granular Time Frame (1/10 sec)

n/a

Aggregate Sequences into Single Data Frame

Result:
File 1:
39 chord sequences
3,583 synchronized chord estimations

Result:
File 2:
Index of 39 source files and algorithms

Aggregated Chord Sequences

time frame	Seq 1	Seq 2	Seq 3
0.1 s	e min	G maj	G maj
0.2 s	C7	C7	a min
0.3 s	D7	b dim	D maj
0.4 s	D7	D7	D7

Algorithm Key

For each chord sequence
Source file name and location
Generating algorithm
Year of submission
Album Name and Song Title (not available for all datasets)

Examples of incompatibilities that can confuse direct chord comparison

Three source sequences with different start/end times and non-aligned chord estimations

start time	end time	chord value
0	0.459	N
0.459	4.122	Bb:maj6
4.122	4.911	C:7
4.911	5.304	Bb:maj6

start time	end time	chord value
0	0.485	N
0.485	4.195	G:min7
4.195	7.64	C:7
7.64	11.395	F:7

start time	end time	chord value
0	0.04643991	N
0.04643991	4.13315193	G:min7
4.13315193	7.56970522	G:min6
7.56970522	8.87002268	F:7

Inclusion of inversions in some estimations will result in different spellings for what is essentially the same chord

start time	end time	chord value
89.999	91.243	F:min
91.243	91.723	F:min/7
91.723	92.531	F:min/b7
92.531	93.383	F:min/6

Unusual enharmonic spellings can confuse comparison algorithms

start time	end time	chord value
206.305	206.705	Bb:maj
206.705	209.925	D#:maj
209.925	210.725	G#:maj
210.725	211.525	D#:maj

Example output:

3 seconds of time-aligned chord estimations from Queen's 'Bohemian Rhapsody'

timeframe	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
7	C:7	C:7	G:min6	NA:	NA:	NA:	NA:	C:maj	C:maj	G:min6	NA:	Eb:maj	NA:	NA:	G:maj	C:7	C:7	NA:
7.1	C:7	C:7	G:min6	NA:	NA:	NA:	NA:	C:maj	F:7	G:min6	NA:	Eb:maj	NA:	NA:	G:maj	C:7	C:7	NA:
7.2	C:7	C:7	G:min6	NA:	NA:	NA:	NA:	C:maj	C:maj	G:min6	NA:	Eb:maj	NA:	NA:	G:maj	C:7	Eb:maj7	NA:
7.3	C:7	C:7	G:min6	NA:	NA:	NA:	NA:	C:maj	F:7	G:min6	NA:	Eb:maj	NA:	NA:	G:maj	C:7	Eb:maj7	NA:
7.4	C:7	C:7	G:min6	NA:	NA:	NA:	NA:	C:maj	C:maj	G:min6	NA:	Eb:maj	NA:	NA:	G:maj	C:7	Eb:maj7	NA:
7.5	C:7	C:7	G:min6	NA:	NA:	NA:	NA:	C:maj	F:7	G:min6	NA:	Eb:maj	NA:	NA:	G:maj	C:7	Eb:maj7	NA:
7.6	F:7	F:7	F:7	F:maj	NA:	F:7	F:7	F:7	F:7	F:7	F:maj	Eb:maj	F:7	F:7	G:maj	F:7	Eb:maj7	F:7
7.7	F:7	F:7	F:7	F:maj	NA:	F:7	F:7	F:7	F:7	F:7	F:maj	Eb:maj	F:7	F:7	G:maj	F:7	Eb:maj7	F:7
7.8	F:7	F:7	F:7	F:maj	NA:	F:7	F:7	F:7	F:7	F:7	F:maj	Eb:maj	F:7	F:7	G:maj	F:7	Eb:maj7	F:7
7.9	F:7	F:7	F:7	F:maj	NA:	F:7	F:7	F:7	F:7	F:7	F:maj	Eb:maj	F:7	F:7	G:maj	F:7	Eb:maj7	F:7
8	F:7	F:7	F:7	F:maj	NA:	F:7	F:7	F:7	F:7	F:7	F:maj	Eb:maj	F:7	F:7	G:maj	F:7	Eb:maj7	F:7
8.1	F:7	F:7	F:7	F:maj	NA:	F:7	F:7	F:7	F:7	F:7	F:maj	Eb:maj	F:7	F:7	G:maj	F:7	Eb:maj7	F:7
8.2	F:7	F:7	F:7	F:maj	F:7	F:7	F:7	F:7	F:7	F:7	F:maj	Eb:maj	F:7	F:7	G:maj	F:7	Eb:maj7	F:7
8.3	F:7	F:7	F:7	F:maj	F:7	F:7	F:7	F:7	F:7	F:7	F:maj	Eb:maj	F:7	F:7	G:maj	F:7	Eb:maj7	F:7
8.4	F:7	F:7	F:7	F:maj	F:7	F:7	F:7	F:7	F:7	F:7	F:maj	Eb:maj	F:7	F:7	G:7	F:7	Eb:maj7	F:7
8.5	F:7	F:7	F:7	F:maj	F:7	F:7	F:7	F:7	C:min	F:7	F:maj	Eb:maj	F:7	F:7	G:7	F:7	F:7	F:7
8.6	F:7	F:7	F:7	F:maj	F:7	F:7	F:7	F:7	C:min	F:7	F:maj	Eb:maj	F:7	F:7	G:7	F:7	F:7	F:7
8.7	F:7	F:7	F:7	F:maj	F:7	F:7	F:7	F:7	C:min	F:7	F:maj	Eb:maj	F:7	F:7	G:7	F:7	F:7	F:7
8.8	C:min7	C:min7	F:7	F:maj	C:min	F:7	F:7	C:min	C:min	F:7	F:maj	Eb:maj	C:min	C:min	G:7	F:7	F:7	Eb:maj
8.9	C:min7	C:min7	F:7	F:maj	C:min	F:7	F:7	C:min	C:min7	F:7	Eb:maj	F:maj	C:min	C:min	Bb:maj	G:min	F:7	Eb:maj
9	C:min7	C:min7	F:7	F:maj	C:min	F:7	F:7	C:min	C:min7	F:7	Eb:maj	F:maj	C:min	C:min	Bb:maj	G:min	C:min7	Eb:maj
9.1	C:min7	C:min7	F:7	F:maj	C:min	F:7	F:7	C:min	C:min7	F:7	Eb:maj	F:maj	C:min	C:min	Bb:maj	G:min	C:min7	Eb:maj
9.2	C:min7	C:min7	F:7	F:maj	C:min	F:7	F:7	C:min	C:min7	F:7	Eb:maj	F:maj	C:min	C:min	Bb:maj	G:min	C:min7	Eb:maj
9.3	C:min7	C:min7	C:min7	F:maj	C:min	F:7	F:7	C:min	C:min7	C:min7	Eb:maj	F:maj	C:min	C:min	Bb:maj	G:min	C:min7	Eb:maj
9.4	C:min7	C:min7	C:min7	F:maj	C:min	F:7	F:7	C:min	C:min7	C:min7	Eb:maj	F:maj	C:min	C:min	Bb:maj	G:min	C:min7	Eb:maj
9.5	C:min7	C:min7	C:min7	F:maj	C:min	F:7	F:7	C:min	C:min7	C:min7	Eb:maj	F:maj	C:min	C:min	Bb:maj	G:min	C:min7	F:maj
9.6	C:min7	C:min7	F:maj	F:maj	C:min	F:7	F:7	C:min	C:min7	C:min	Eb:maj	F:maj	C:min	C:min	Bb:maj	F:maj	C:min7	F:maj
9.7	F:7	F:7	F:maj	F:maj	F:maj	F:7	F:7	C:min	C:min	C:min	F:7	F:maj	C:min	C:min	Bb:maj	F:maj	C:min7	F:maj
9.8	F:7	F:7	F:maj	F:maj	F:maj	F:7	F:7	F:7	C:min	C:min	F:7	F:maj	C:min	C:min	Bb:maj	F:maj	C:min7	F:maj
9.9	F:7	F:7	F:maj	F:maj	F:maj	F:7	F:7	F:7	C:min	C:min	F:7	F:maj	C:min	C:min	Bb:maj	F:maj	F:7	F:maj
10	F:7	F:7	F:maj	F:maj	F:maj	F:7	F:7	F:7	F:maj	C:min	F:7	F:maj	F:maj	F:maj	G:7	F:7	F:7	F:maj

Python code and Bohemian Rhapsody example data are available at
https://github.com/jeffkmiller/chord-data-aggregator