Datasets for music research
Audio vs symbolic datasets

• Audio: Continuous data
• Symbolic: note, event data
• Other elements of datasets
  • Metadata
  • Annotations
  • Extractions (chords, melodies, lyrics)
ISMIR data set:

https://www.ismir.net/resources/

• Possibly the largest
• Overwhelmingly contains audio sources but many resources not continuous
• Completely unstandardized

• https://www.ismir.net/resources/datasets/
Music Brainz (London)

- “A dataset of **hierarchical genre definitions**” [metadata for audio features]
- Intended to evaluate recordings of undetermined content
- Collaboration of Barcelona MT group and Music Brainz (London)
- Not updated since July 2022; processing still in progress
- >700,000 million unique titles (annotated database entries)
- Genre designations stated as probabilities
- One example: [https://acousticbrainz.org/f36d9818-019b-4379-ad13-5080feb9ad8a](https://acousticbrainz.org/f36d9818-019b-4379-ad13-5080feb9ad8a)
- Dataset list: [https://acousticbrainz.org/datasets/list](https://acousticbrainz.org/datasets/list)
A word about genre [re: Music Brainz]

- Definitions of genre variable
- Clustering and similarity depend on boundaries that are not fixed
- Similar problems in classical and popular music
- Reasons for changing definitions may be different
- *Billboard Magazine* manipulates boundary definitions
DALI: Singing voice detection

- Resources on github: [https://github.com/gabolsgabs/DALI](https://github.com/gabolsgabs/DALI)
- Hierarchical connections for synchronized audio, lyrics, vocal notes
- Based in Paris (Geoff Pieters, IRCAM et al); > 5,300 songs
ELVIS: duos-Josquin&LaRue-CPDL

• McGill University
• Consistent labeling; symbolic data (77 duos)
• Most files from other sources including the Choral Public Domain Library (CPDL), now ChoralWiki
• https://www.cpdl.org/wiki/index.php/Main_Page
Choral Wiki

- Significant quantities of metadata
- Several symbolic formats
- 4000 composers
- 44,000 scores
- Seven additional languages
- Categories of usage, e.g., church music by season
LabROSA

• Automatic piano transcription (29)
• Disklavier, synthesized audio
• Dan Ellis et al., Columbia U.

• http://labrosa.ee.columbia.edu/sounds/music/
Last.fm

- Listener centric research at U: Lin
- Enormous database of listening events
  http://www.cp.jku.at/datasets/LFM-1b/
- Emphasis on genre; geography; et al:
  - Genre profile
    http://www.cp.jku.at/datasets/LFM-1b/
  - Full dataset 8GB
    http://drive.jku.at/ssf/s/readFile/share/1056/266403063659030189/publicLink/LFM-1b.zip
Giant MIDI piano (bytedance)

- Symbolic data sets for analysis
- Culled names of works and composers from IMSLP
- Sought corresponding audio files
- Used convolutional NNs to detect solos, transcribed them
- 90% live performances
- Sought pitch class, interval, trichords, tetrachords
- Data: [https://github.com/bytedance/GiantMIDI-Piano/tree/master/midis_for_evaluation/giantmidi-piano](https://github.com/bytedance/GiantMIDI-Piano/tree/master/midis_for_evaluation/giantmidi-piano)
MAESTRO dataset

- MIDI and audio for synchronous tracks
- Extracted from International Piano e Composition, 120 GB
- Magenta resource
SUPRA: Piano-roll performances

• 478 performances
• E.g. Paderewski playing his own Minuet in G
• https://github.com/pianoroll/SUPRA/blob/master/metadata/marcxml/d/dw479gk0103.marcxml
Shazam sample data

• https://purl.stanford.edu/fj396zz8014
• Offsets and query dates from Billboard hits (2015)