

Appendix A: Candidate Music IR Test Collections

Donald Byrd
School of Music
Indiana University
+1-(812)-856-0129
donbyrd@indiana.edu

Last revised 28 March 2003

By themselves, these collections might well be useful for a number of research projects. But with suitable queries and relevance judgments (if at all possible, human-produced), they would be much more useful, especially for cross-project comparisons via EMIR (tentative name of the proposed TREC-like "Evaluation of Music IR"). NB: appearance in this list means only that I think the collection exists in machine-readable form somewhere and *might* be available; there are serious copyright as well as other availability issues for most of these! Exception: collections whose name is prefixed with "*" don't even exist yet in electronic form, as far as I know.

The RWC database is the one most clearly available (or soon to be available) for research purposes, but it's very, very small. The Uitenbogerd & Zobel collection has the only existing set of human relevance judgments I know of, but the judgments are not at all extensive.

This list is a continuing work-in-progress: I have very little time for it, but attempt to maintain it as a public service. In particular, I'm aware that the listings for CMN in image format-scanned scores and/or sheet music-are not very up-to-date. Comments are very welcome!

Entries are in alphabetical order by name.

| <i>Name</i> | <i>Representation</i> | <i>Encoding</i> | <i>Complexity</i> | <i>Approx size</i> | <i>Comments</i> |
|------------------------|-----------------------|------------------|-------------------|--------------------|--|
| Bach Chorales | Event | MIDI (SMF) | Polyphonic | 400 | Short 4-part contrapuntal pieces: 185 (from BG v.39) + c.200 (from elsewhere) |
| Barlow and Morgenstern | CMN | ?? | Monophonic | 10,000 | Themes of classical pieces plus "index" |
| CCARH MuseData | CMN | MuseData | Polyphonic | 4000 | Complete movements of 881 Classical pieces from c. 1680 to 1815, at least 1/3 Bach. Includes 185 Bach chorales. See www.musedata.org |
| CCARH MuseData | CMN | kern | Polyphonic | 3000 | Complete movements of c. 700 Classical pieces from c. 1680 to 1815, about half Bach. Includes 185 Bach chorales. (Based on the above.) |
| CD Sheet Music | "CMN" | Image | Polyphonic | 7000? | Scanned scores. 48 CDs, each of c. 750-3800 pages of PDFs; total c. 86,400 pages |
| Classical Archives | Event | MIDI (SMF) | Polyphonic | 17,000 | Contributed MIDI files of public-domain works; quality uneven. See www.classicalarchives.com |
| Classical Archives | Audio | Lossy (MP3, WMA) | Polyphonic | 5200 | Recordings of public-domain works. |
| *ECOLM | CMN | LTN | Polyphonic | 1000 | Electronic Corpus of Lute Music: complete lute pieces |
| Harvard/MIT | "CMN" | Image | Polyphonic?? | ?? | Scanned scores |

continued on next page ...

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

| <i>Name</i> | <i>Representation</i> | <i>Encoding</i> | <i>Complexity</i> | <i>Approx size</i> | <i>Comments</i> |
|------------------------------|-----------------------|--------------------------------------|-------------------|--------------------|---|
| Huron | CMN | Humdrum kern | ?? | 5000 | |
| HymnQuest | CMN | ?? | ?? | 13,000 | Published by Stainer & Bell on CDROM (with monophonic content searching) |
| IRCAM | Audio | Most lossy (MP2), some lossless (CD) | Polyphonic | 5000 | contemporary music works, over 50% non-commercially available; all with metadata |
| JHU/Levy Sheet Music | "CMN", CMN | Image, some Guido | Polyphonic | 29,000 | scanned sheet music (c.100,000 pages, c. 80% public-domain) |
| L of C/Copyright Sheet Music | "CMN" | Image | Polyphonic | 22,000 | scanned sheet music ("American Sheet Music, 1870-1885": pieces copyrighted in those years) |
| L of C/Duke Sheet Music | "CMN" | Image | Polyphonic | 3000 | scanned sheet music ("Historic American Sheet Music, 1850-1920": pieces from the Duke collection) |
| Meldex (NZDML) Folksongs | CMN | Meldex | Monophonic | 10,000 | 9400 German, Chinese, and Anglo-American folksongs from Schaffrath's collection and Digital Tradition. (Better would be *"MELDEX Plus": add back in the c. 200 containing tuplets NZDML removed.) |
| Mutopia | CMN | LilyPond(?) | Polyphonic | 300 | Contributed classical works: www.mutopiaproject.org/index.html |
| Nightingale | CMN | Nightingale | Polyphonic | 600 | movements and excerpts, mostly classical (same music as below) |
| Nightingale | Event | MEF | Polyphonic | 600 | movements and excerpts, mostly classical (same music as above) |
| NZDML Fake book | CMN | Meldex(?) | Monophonic?? | 1200 | popular tunes |
| NZDML MidiMax | Event | MIDI (SMF) | Polyphonic?? | 100,000 | Standard MIDI files (collected from the Web?) |
| Parsons | Pitch Contour | Pitch Contour | Monophonic | 11,000 | up/down/repeat encoding of pitches of themes of classical pieces |
| Pickens | Event | MIDI (SMF) | Polyphonic | 1200 | Piano-only MIDI files, collected by hand from the web, in a variety of styles: Ragtime, classical, popular, original |
| RISM | CMN | Plaine &Easie | ?? | 250,000 | Incipits from 188,000 works by end of 1995 |
| RWC | Event | MIDI (SMF) | Polyphonic | 200 | Pieces in four genres: 100 pop, 15 "royalty-free" (sic), 50 classical, 50 jazz; most original, all freely available for research (same as below) |
| RWC | Audio | Lossless (CD) | Polyphonic | 200 | Pieces in four genres: 100 pop, 15 "royalty-free" (sic), 50 classical, 50 jazz; most original, all freely available for research (same as above) |
| Sunhawk | CMN | Proprietary | Polyphonic | 25,000 | Many genres. See www.sunhawk.com |

continued on next page ...

| <i>Name</i> | <i>Representation</i> | <i>Encoding</i> | <i>Complexity</i> | <i>Approx size</i> | <i>Comments</i> |
|-------------------------|-----------------------|------------------|-------------------|--------------------|--|
| Templeton | "CMN" | Image | Polyphonic | 22,000 | scanned sheet music; at Mississippi State University |
| Themefinder | CMN | MuseData, kern? | Monophonic | 37,000 | incipits of classical pieces and folksongs. See www.themefinder.org |
| UCLA Pop. American | "CMN" | Image | Polyphonic?? | 600 | Pieces of popular American music |
| Uitdenbogerd & Zobel | Event | MIDI (SMF) | Polyphonic | 10,500 | MIDI files (collected from the Web). Some "real" relevance judgments exist. |
| Variations, Variations2 | "CMN" | Image | Polyphonic | ?? | Scanned scores |
| Variations, Variations2 | Audio | Lossy (MP2, MP3) | Polyphonic | ?? | c. 9,000 hours (original Variations) |

Other possibilities:

| | | | | | |
|---------------|-------|---------------|----|----|---|
| *?? | Audio | WAV, WMA, MP3 | ?? | ?? | Preferably encoded in a lossless format |
| *Web crawling | Event | MIDI (SMF) | ?? | ?? | ?? |