SIMSSA DB: A Database for Computational Musicological Research

Cory McKay

Marianopolis College

2018 International Association of Music Libraries, Archives and Documentation Centres International Congress, SIMSSA Workshop, Leipzig, Germany

Topics

- Currently available musicological research databases and repositories
- Data needs of computational musicology and MIR
- The SIMSSA DB
  - Features and jSymbolic
  - Archiving research
  - Design priorities
  - Data model
  - Prototype interface

Existing music research databases

- There are several excellent on-line databases available that provide researchers with access to:
  - Musical metadata
    - e.g. Bach Digital
  - Images of scores and manuscripts
    - e.g. Musiclibs
  - Audio recordings
    - e.g. Naxos Digital (paid service)

Symbolic music repositories (1/2)

- However, there are relatively few research-grade on-line repositories of symbolic music
  - i.e. Finale, Sibelius, Music XML, MEI, MIDI, etc. files
- Most symbolic music repositories that do exist tend to either:
  - Have unreliable data and metadata (intended for non-specialist use rather than rigorous musicological research)
    - e.g. Classical Archives or Musescore
  - Be limited in scope
    - e.g. the SEILS dataset
  - Have relatively limited metadata structuring and only basic search functionality
    - e.g. Kern Scores

Symbolic music repositories (2/2)

- Those few research-grade symbolic music repositories that do exist are used heavily by musicologists and MIR researchers
  - e.g. the Josquin Research Project
- This makes it clear how much such resources are needed by the research community

Computational musicology and MIR

- Automated data extraction software, statistical analysis techniques and machine learning now allow us to:
    - □ Study huge quantities of music very quickly
        - More than any human could reasonably look at
    - □ Empirically validate (or repudiate) our theoretical predictions
    - □ Do purely exploratory studies of music
    - □ See music from fresh perspectives

We need symbolic data

- But to take full advantage of these techniques, researchers need symbolic music files
  - □ Lots of symbolic music files
  - □ Varied symbolic music files
  - □ High-quality and symbolic music files
  - □ Consistently encoded symbolic music files
- So where can researchers get these?
  - □ *<pause type=“dramatic”>1 sec</pause>* . . .

Introducing the SIMSSA DB

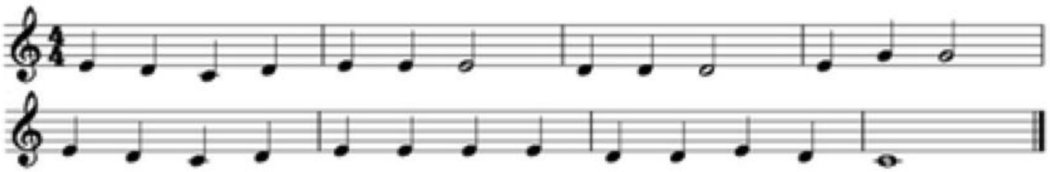
- Emphasizes research-grade symbolic music files
- Permits flexible, high-quality searchable metadata
  - Of the kinds specifically needed by musicologists and MIR researchers
  - Allows modelling of complex relationships
  - Provenance is given particular centrality
- Allows records to be kept of the specific files (and other related information) used in individual research studies
- Permits content-based (as well as metadata-based) search and analysis
  - Let's expand on this for a moment . . .

The notion of a “feature”

- A feature is a piece of information that characterizes something (e.g. a piece of music) in a simple way
- Usually a simple numerical value
  - A feature can be a single value, or it can be a set of related values (e.g. a histogram)
- Can be extracted from pieces in their entirety, or from segments of pieces
- Can use features to compare and look for patterns in different music in a macro sense

Example: A basic feature

- Range (1-D): Difference in semitones between the highest and lowest pitches



- Value of this feature: 7
  - □ G - C = 7 semitones

- In practice, of course, we want many features, not just one . . .

jSymbolic (1/2)

- jSymbolic is our software platform for automatically extracting features from symbolic music (ISMIR 2018)
- Extracts 246 unique features (version 2.2)
  - □ Some of these are multi-dimensional, including histograms
  - □ Extracts a total of 1497 separate values (version 2.2) per symbolic music file

jSymbolic (2/2)

- Types of information accessed by jSymbolic features:
  - Pitch statistics
  - Melody / horizontal intervals
  - Chords / vertical intervals
  - Texture
  - Rhythm
  - Instrumentation
  - Dynamics

SIMSSA DB and jSymbolic features

- jSymbolic is being integrated into the SIMSSA DB
  - □ Whenever a file is added to the DB, features are automatically extracted and used to index the file
- Users can use these features to search the DB based on musical content as well as metadata
  - □ e.g. retrieve all pieces composed by J. S. Bach in Leipzig that contain vertical tritones or parallel fifths
- Researchers can also download and use features directly as input to statistical analysis and machine learning tools (or use manual analysis) to study things such as:
  - □ Composer attribution (MedRen 2017, ISMIR 2017)
  - □ Genre (MedRen 2018, ISMIR 2010)
  - □ Regional styles (APM 2018)

Archiving research

- Researchers can submit information on particular studies they performed
  - Specifically which symbolic music files were used
  - Specifically which features (if any) were used
  - Workflows, results, analysis, conclusions, publications and other related data
- Essential for repeatability, direct comparison of approaches, iterative refinements, etc.
  - jSymbolic configuration files can be auto-generated for each study in order to facilitate this

Design priorities (1/8)

- Make the repository as accessible as possible to all music researchers, regardless of technological training
  - □ As users
  - □ As data (and metadata) contributors
  - □ As editors / validators
- This requires a front-end that is easy-to-use
  - □ And that hides details of the data model from users that they do not need to be aware of

Design priorities (2/8)

- Use authority control and cataloguing standards to reduce ambiguity and redundancy (and increase consistency) as much as possible
- Initial focus on VIAF authority files, but also looking at:
  - FRBR
  - Wikidata
  - RISM’s Muscat and authority files
  - RDA
  - Library of Congress
- Populate fields with URIs and use linked open data practices when possible
  - But also allow contributors to enter raw text into fields (to meet the realistic needs of and constraints faced by musicologists)

Design priorities (3/8)

- Information relating to quality control and file encoding methodology must be kept
  - □ Who submitted data or metadata
  - □ Who verified or edited data or metadata
  - □ Who (or what software) encoded a symbolic music file, and using what settings
    - Encoding methodologies can significantly influence results if one is not careful (ISMIR 2018)

Design priorities (4/8)

- Keeping a record of provenance is musicologically essential
  - □ Each symbolic music file is linked to a specific source (digital or physical)
  - □ Each source can be linked to its parent source(s) through chains of provenance
  - □ e.g. an MEI file is derived from a printed score J. S. Bach score, which is derived from a hand-written copyist’s manuscript, which is derived from a (potentially lost) original manuscript hand-written by Bach

Design priorities (5/8)

- Maintain a conceptual separation between abstract musical works and particular instantiations of them (as expressed by symbolic files and sources)
  - Multiple versions of the same abstract work can exist, and these should be both associated with and differentiated from one another
  - e.g. different symbolic encodings
  - e.g. different editions, arrangements, etc. of a work

Design priorities (6/8)

- Make it possible to divide abstract musical works into abstract sections and parts
  - □ Symbolic files sometimes contain whole pieces, and sometimes only parts of pieces
- Make it possible to keep track of complex relationships between works, sections and parts
  - □ e.g. a movement of one mass might be reused in another mass
  - □ e.g. an orchestral score and a piano reduction of it have different parts, but they are the same work and have the same sections

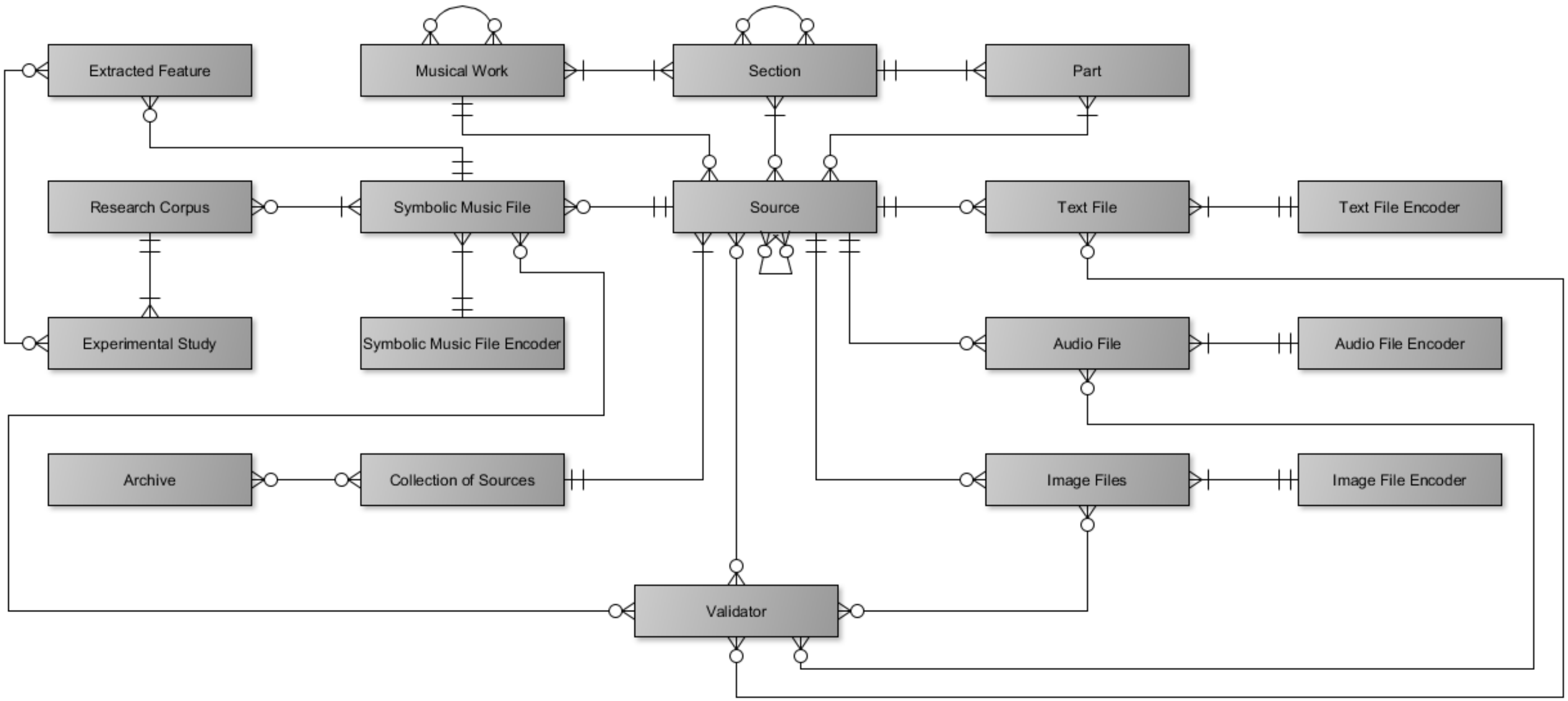
Design priorities (7/8)

- Make it possible to link an abstract musical work (and its sections and parts) to instantiations in multiple formats
  - Symbolic music files
  - Musical texts
  - Images of scores or manuscripts
  - Audio files
- Although our primary focus is on symbolic music, this data is ultimately all related . . .

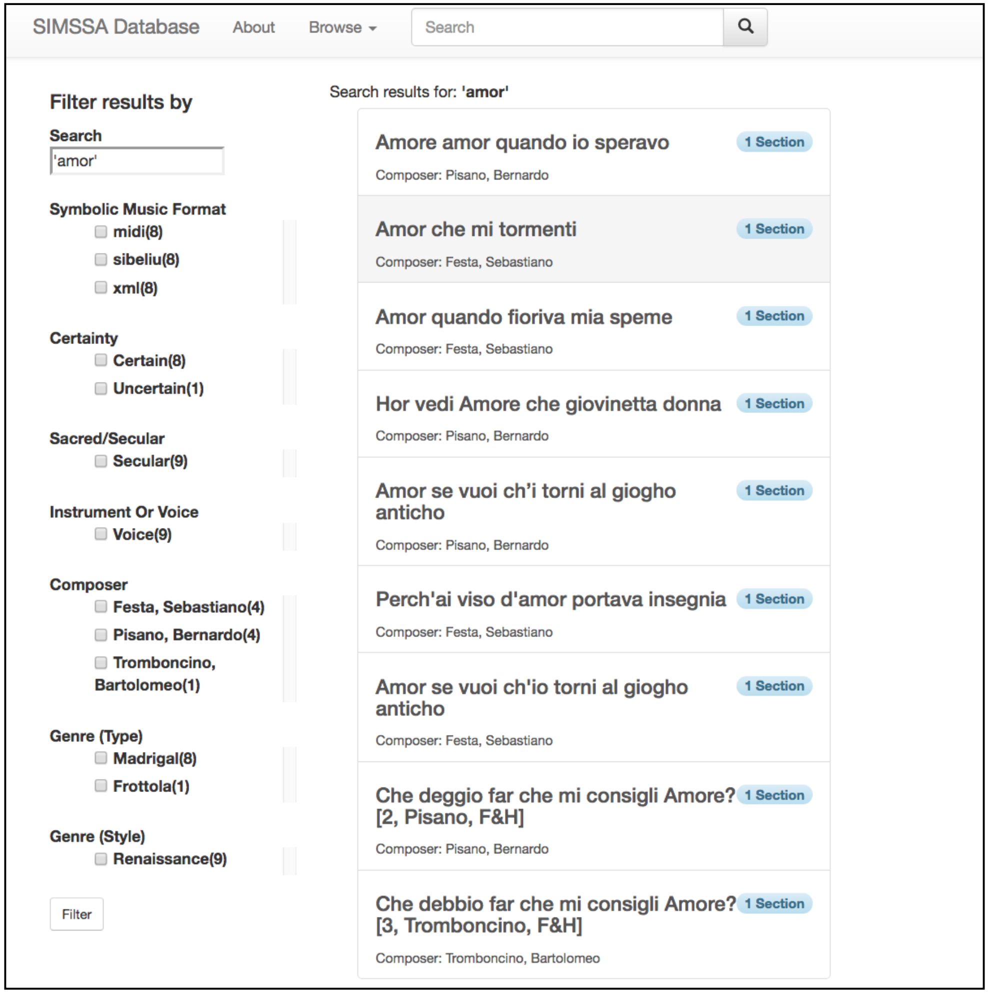
Design priorities (8/8)

- Long-term, we want to:
  - □ Link our data to the contents of other repositories
    - e.g. DOREMUS, Josquin Research Project, etc.
    - We are putting a design emphasis on making it possible to import or export information using linked open data frameworks
    - IIIF-compatibility will certainly help with respect to images
  - □ Take as input symbolic files auto-generated from images using OMR
    - As the technology improves
  - □ Take as input symbolic files auto-generated from audio files using automatic transcription algorithms
    - As the technology improves

Overview ERD of our data model



Prototype interface (1/3)



Filter results by

Search

'amor'

Symbolic Music Format
- □ midi(8)
- □ sibeliu(8)
- □ xml(8)

Certainty
- □ Certain(8)
- □ Uncertain(1)

Sacred/Secular
- □ Secular(9)

Instrument Or Voice
- □ Voice(9)

Composer
- □ Festa, Sebastiano(4)
- □ Pisano, Bernardo(4)
- □ Tromboncino, Bartolomeo(1)

Genre (Type)
- □ Madrigal(8)
- □ Frottola(1)

Genre (Style)
- □ Renaissance(9)

Filter

Search results for: **'amor'**

Amore amor quando io speravo — 1 Section
Composer: Pisano, Bernardo

Amor che mi tormenti — 1 Section
Composer: Festa, Sebastiano

Amor quando fioriva mia speme — 1 Section
Composer: Festa, Sebastiano

Hor vedi Amore che giovinetta donna — 1 Section
Composer: Pisano, Bernardo

Amor se vuoi ch'i torni al giogho anticho — 1 Section
Composer: Pisano, Bernardo

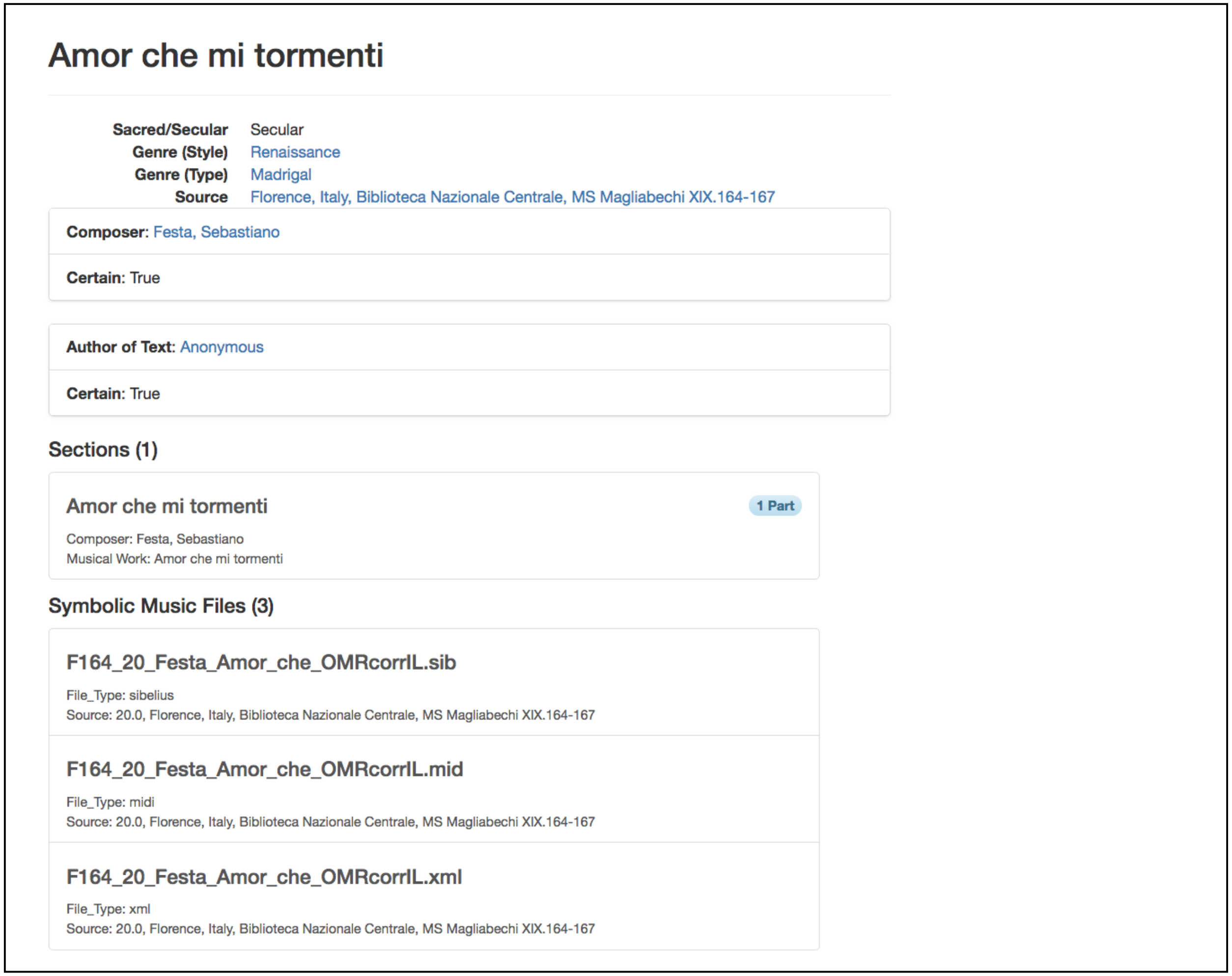
Perch'ai viso d'amor portava insegnia — 1 Section
Composer: Festa, Sebastiano

Amor se vuoi ch'io torni al giogho anticho — 1 Section
Composer: Festa, Sebastiano

Che deggio far che mi consigli Amore? [2, Pisano, F&H] — 1 Section
Composer: Pisano, Bernardo

Che debbio far che mi consigli Amore? [3, Tromboncino, F&H] — 1 Section
Composer: Tromboncino, Bartolomeo

Prototype interface (2/3)



Amor che mi tormenti

Sacred/Secular	Secular
Genre (Style)	Renaissance
Genre (Type)	Madrigal
Source	Florence, Italy, Biblioteca Nazionale Centrale, MS Magliabechi XIX.164-167

Composer: Festa, Sebastiano

Certain: True

Author of Text: Anonymous

Certain: True

Sections (1)

Amor che mi tormenti

Composer: Festa, Sebastiano
Musical Work: Amor che mi tormenti

Symbolic Music Files (3)

F164_20_Festa_Amor_che_OMRcorrlL.sib

File_Type: sibelius
Source: 20.0, Florence, Italy, Biblioteca Nazionale Centrale, MS Magliabechi XIX.164-167

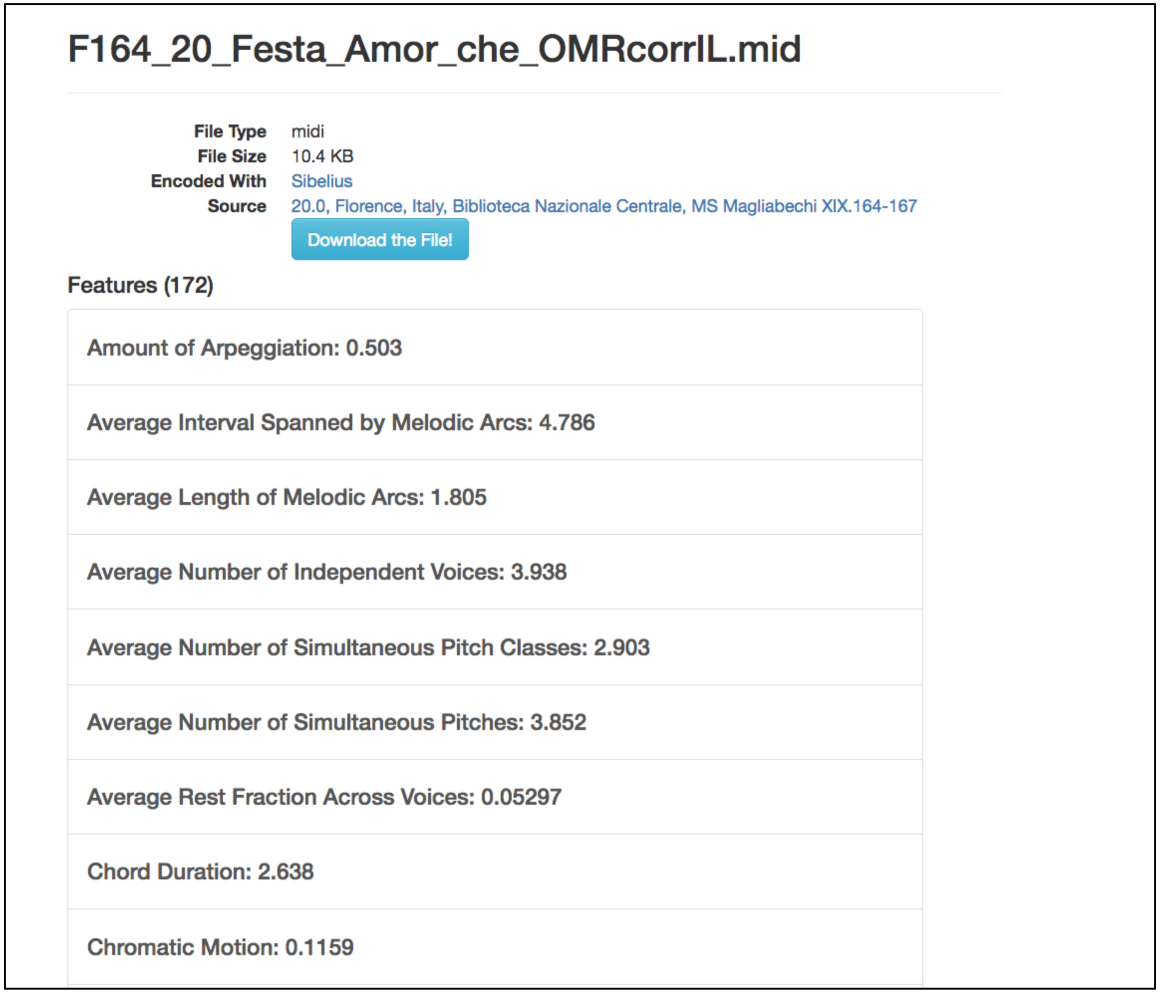
F164_20_Festa_Amor_che_OMRcorrlL.mid

File_Type: midi
Source: 20.0, Florence, Italy, Biblioteca Nazionale Centrale, MS Magliabechi XIX.164-167

F164_20_Festa_Amor_che_OMRcorrlL.xml

File_Type: xml
Source: 20.0, Florence, Italy, Biblioteca Nazionale Centrale, MS Magliabechi XIX.164-167

Prototype interface (3/3)



F164_20_Festa_Amor_che_OMRcorrIL.mid

File Type	midi
File Size	10.4 KB
Encoded With	Sibelius
Source	20.0, Florence, Italy, Biblioteca Nazionale Centrale, MS Magliabechi XIX.164-167

Features (172)

- Amount of Arpeggiation: 0.503
- Average Interval Spanned by Melodic Arcs: 4.786
- Average Length of Melodic Arcs: 1.805
- Average Number of Independent Voices: 3.938
- Average Number of Simultaneous Pitch Classes: 2.903
- Average Number of Simultaneous Pitches: 3.852
- Average Rest Fraction Across Voices: 0.05297
- Chord Duration: 2.638
- Chromatic Motion: 0.1159

Highlights of the SIMSSA DB

- Designed to meet the specific needs of researchers wishing to engage in large-scale computational musicological and MIR research
- Focus on symbolic music files
  - □ But also permits links with images, audio files and texts
- Emphasis on accessibility to researchers
- Emphasis on quality and consistency of both metadata and data
  - □ Authority control and cataloguing standards
- Modeling of complex musical relationships
  - □ Relationships between (abstract) works, sections and parts
  - □ Mapping musical instantiations (e.g. files) to abstract musical entities
  - □ Emphasis on provenance
- Archiving of experiments
- Content-based search and analysis based on features
  - □ As well as metadata-based searches, of course

Thanks for your attention

- **E-mail:** cory.mckay@mail.mcgill.ca
- **The SIMSSA DB team:**
  - Julie E. Cumming, Ichiro Fujinaga, Andrew Hankinson, Emily Hopkins, Yaolong Ju, Andrew Kam, Gustavo Polins Pedro

McGill

Schulich School of Music
École de musique Schulich

CIRMMT
Centre for Interdisciplinary Research in Music Media and Technology

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada

Canada

Marianopolis College

Fonds de recherche sur la société et la culture
Québec

DDMAL
Distributed Digital Music Archives & Libraries Lab

SIMSSA
Single Interface for Music Score Searching and Analysis

jMIR