A Collaborative Symbolic Music Database for Computational Research on Music

Cory McKay *(Marianopolis College, Canada)*

Emily Hopkins, Gustavo Polins Pedro, Yaolong Ju, Andrew Kam, Julie Cumming and Ichiro Fujinaga *(McGill University, Canada)*

2019 Medieval and Renaissance Music Conference
Basel, Switzerland

Computational musicology: Advantages

- Computational approaches can be very usefully applied to early music:
  - □ Study huge quantities of music very quickly
  - □ Empirically validate (or repudiate) hypotheses
  - □ Do purely exploratory studies of music
    - See music from fresh perspectives

Computational musicology: Challenges

- Require large quantities of music encoded in machine-readable “symbolic” formats
  - □ e.g. Music XML, MEI, MIDI, Sibelius, Finale, etc.
  - □ Transcribed and encoded using consistent and well-document methodologies (Cumming et al. 2018)
- Meaningful, reliable and consistent metadata annotations needed to track, search and contextualize the music
  - □ Structured enough to allow sophisticated exploration, but flexible enough to not compromise usability
- Data ideally open and publicly accessible
  - □ Permits experimental repeatability and inter-scholar refinement

A solution!

- Open, on-line databases of symbolic music designed with the specific needs of musicologists and theorists in mind
- Ideally, such databases should:
  - Permit sophisticated searches of both metadata and musical content
  - Allow access and contributions by any scholar

The need for more repositories

- Unfortunately, there are relatively few large research-grade on-line repositories of symbolic music files
  - Fewer still that that are proper databases
  - Fewer still holding large, broad collections
  - Fewer still that are fully open
- Those few that do exist are used heavily by musicologists and other researchers
  - e.g. the Josquin Research Project
  - Makes it clear how much such resources are needed

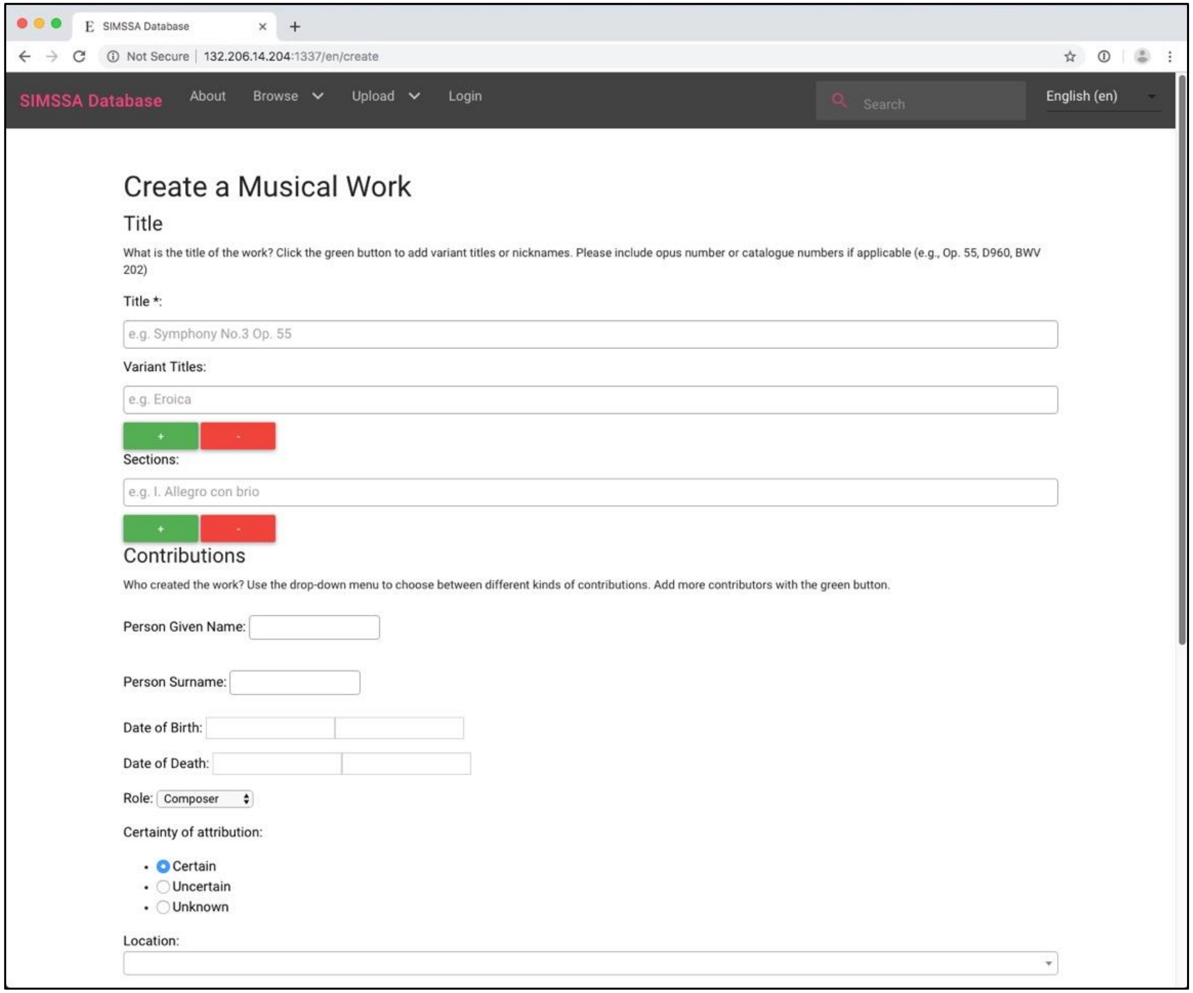
SIMSSA DB

- We are constructing the SIMSSA DB to meet this need
  - Specifically designed for the needs of musicologists and theorists
  - Particular (but not exclusive) focus on early music
- The remainder of this talk will focus on the structure and functionality of the SIMSSA DB

An infrastructure, not a dataset!

- The SIMSSA DB is not simply a repository of music we have transcribed
  - Although it is seeded with our JLSDD (Cumming et al. 2018), Florence 164 (Cumming & McKay 2018), etc. corpus
- Rather, it is a general unified infrastructure to which other scholars can contribute symbolic music files they have used in their own work

SIMSSA DB prototype contribution form



Create a Musical Work

Title

What is the title of the work? Click the green button to add variant titles or nicknames. Please include opus number or catalogue numbers if applicable (e.g., Op. 55, D960, BWV 202)

Title *: e.g. Symphony No.3 Op. 55

Variant Titles: e.g. Eroica

+ -

Sections: e.g. I. Allegro con brio

+ -

Contributions

Who created the work? Use the drop-down menu to choose between different kinds of contributions. Add more contributors with the green button.

Person Given Name:

Person Surname:

Date of Birth:

Date of Death:

Role: Composer

Certainty of attribution:

- ● Certain
- ○ Uncertain
- ○ Unknown

Location:

A long-term goal: OMR

- The SIMSSA DB is also designed to eventually be populated with music auto-transcribed using optical music recognition (OMR) technology
- OMR is not quite accurate enough yet
  - But researchers at SIMSSA and elsewhere are making important progress

Searching

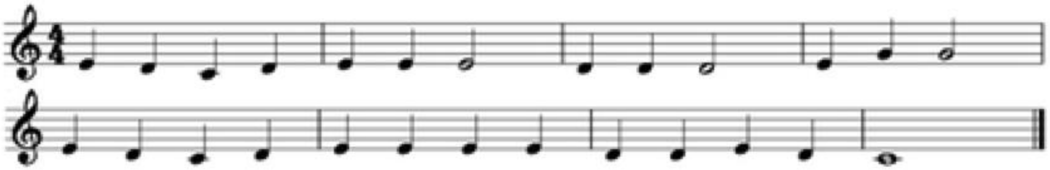
- Aside from contributing music, scholars will of course also wish to access music on the SIMSSA DB
- The SIMSSA DB allows two kinds of searching:
  - □ Free-text or structured metadata searches
    - e.g. title, composer, location, etc.
  - □ Searches of musical content via features
    - Let’s expand on the notion of a “feature”. . .

Defining a “feature”

- A feature is a piece of statistical information that characterizes some aspect of a piece of music using a simple, consistent measurement
  - □ Each feature is represented as one or more simple numerical values
- Can use features to find patterns and compare music and in a macro sense

A basic sample feature: Range

- Range: Difference in semitones between the highest and lowest pitches



- Value of this feature for this music: 7
  - G - C = 7 semitones
- In practice, of course, we want many features, not just one

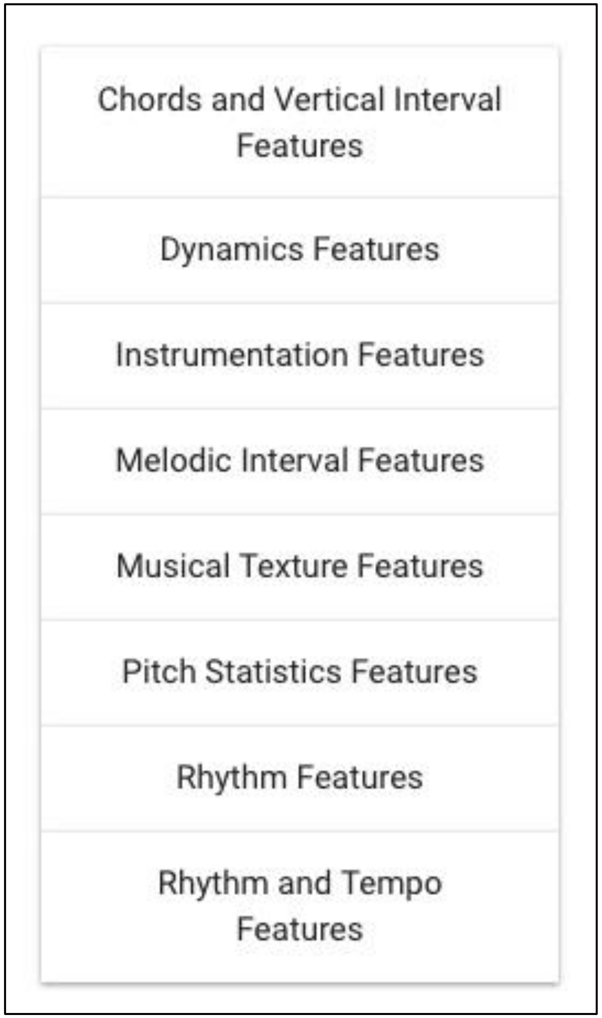
jSymbolic

- jSymbolic is our software platform for automatically extracting features from music (McKay et al. 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

SIMSSA DB and features (1/3)

- jSymbolic has been integrated into the SIMSSA DB
  - Whenever a file is uploaded to the DB, features are automatically extracted and used to index the file
- Users can use these features to search the database based on musical content
  - Can also be combined with metadata searches
  - e.g. retrieve all sacred pieces composed by Josquin that contain parallel fifths

SIMSSA DB and features (2/3)



Chords and Vertical Interval Features

Dynamics Features

Instrumentation Features

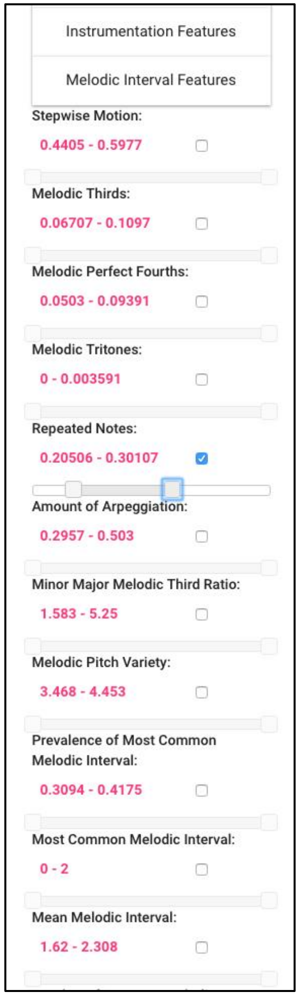
Melodic Interval Features

Musical Texture Features

Pitch Statistics Features

Rhythm Features

Rhythm and Tempo Features



Instrumentation Features

Melodic Interval Features

Stepwise Motion:
0.4405 - 0.5977 ☐

Melodic Thirds:
0.06707 - 0.1097 ☐

Melodic Perfect Fourths:
0.0503 - 0.09391 ☐

Melodic Tritones:
0 - 0.003591 ☐

Repeated Notes:
0.20506 - 0.30107 ☑

Amount of Arpeggiation:
0.2957 - 0.503 ☐

Minor Major Melodic Third Ratio:
1.583 - 5.25 ☐

Melodic Pitch Variety:
3.468 - 4.453 ☐

Prevalence of Most Common Melodic Interval:
0.3094 - 0.4175 ☐

Most Common Melodic Interval:
0 - 2 ☐

Mean Melodic Interval:
1.62 - 2.308 ☐

- Users can specify feature-range searches via a slider for each feature they are interested in

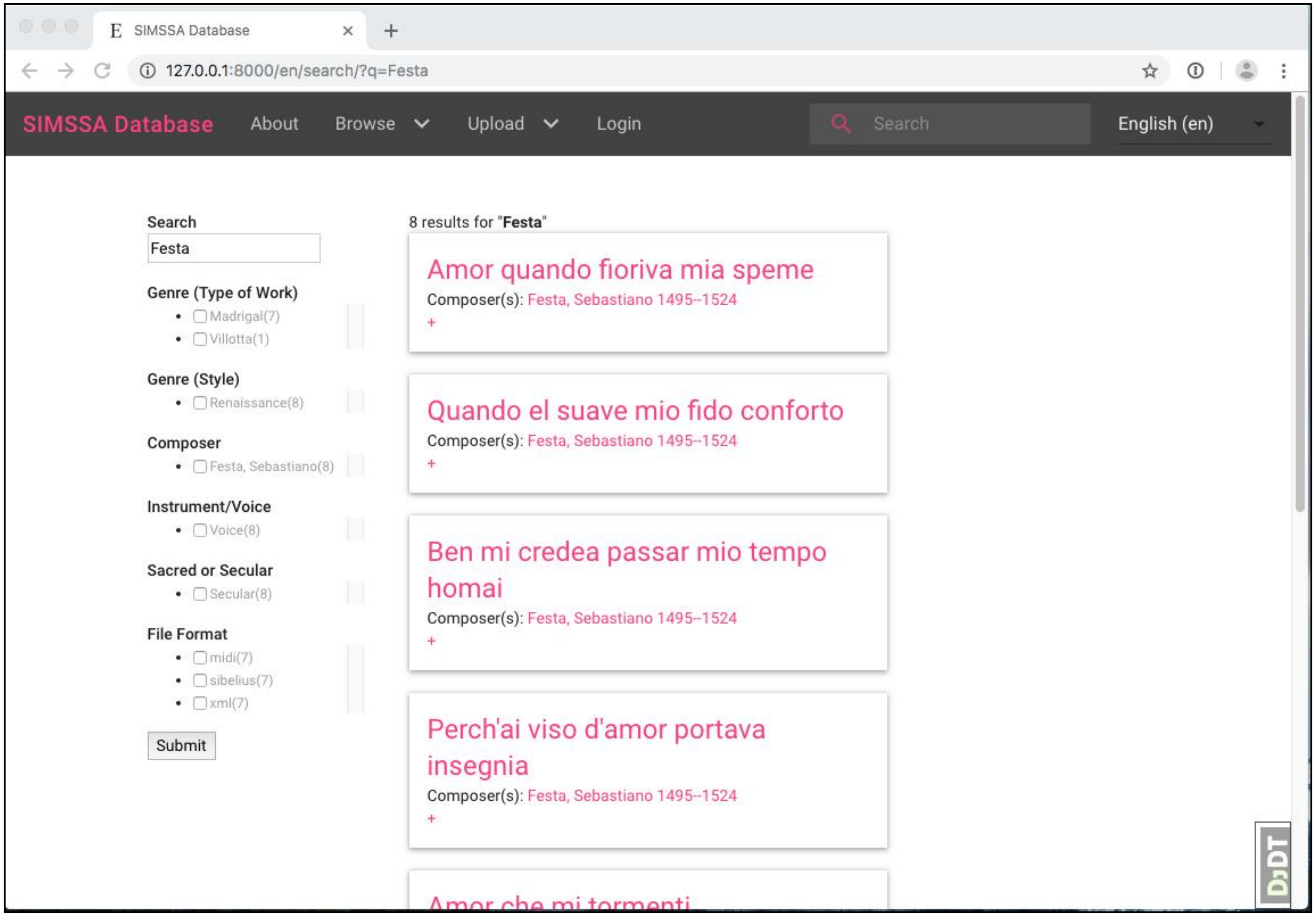
SIMSSA DB and features (3/3)

- Scholars can also download complete feature sets directly and use them as input to statistical analysis and machine learning tools (or use manual analysis) to study things such as:
  - Composer attribution (McKay et al. 2017)
  - Origins of the madrigal (Cumming & McKay 2018)
  - Regional styles (Cuenca & McKay 2019)

Metadata and “faceted” search

- The DB may also be searched using more traditional metadata queries:
  - Free-text search
  - “Faceted” metadata filters, such as:
    - Contributor
      - Composer, arranger, author of text, transcriber, etc.
    - Sacred, secular, etc.
    - Instruments / voices
    - Genre / type of work
      - e.g. madrigal, motet, etc.

Sample query: Free-text



Search

Festa

Genre (Type of Work)

- □ Madrigal(7)
- □ Villotta(1)

Genre (Style)

- □ Renaissance(8)

Composer

- □ Festa, Sebastiano(8)

Instrument/Voice

- □ Voice(8)

Sacred or Secular

- □ Secular(8)

File Format

- □ midi(7)
- □ sibelius(7)
- □ xml(7)

Submit

8 results for "**Festa**"

Amor quando fioriva mia speme

Composer(s): Festa, Sebastiano 1495--1524

+

Quando el suave mio fido conforto

Composer(s): Festa, Sebastiano 1495--1524

+

Ben mi credea passar mio tempo homai

Composer(s): Festa, Sebastiano 1495--1524

+

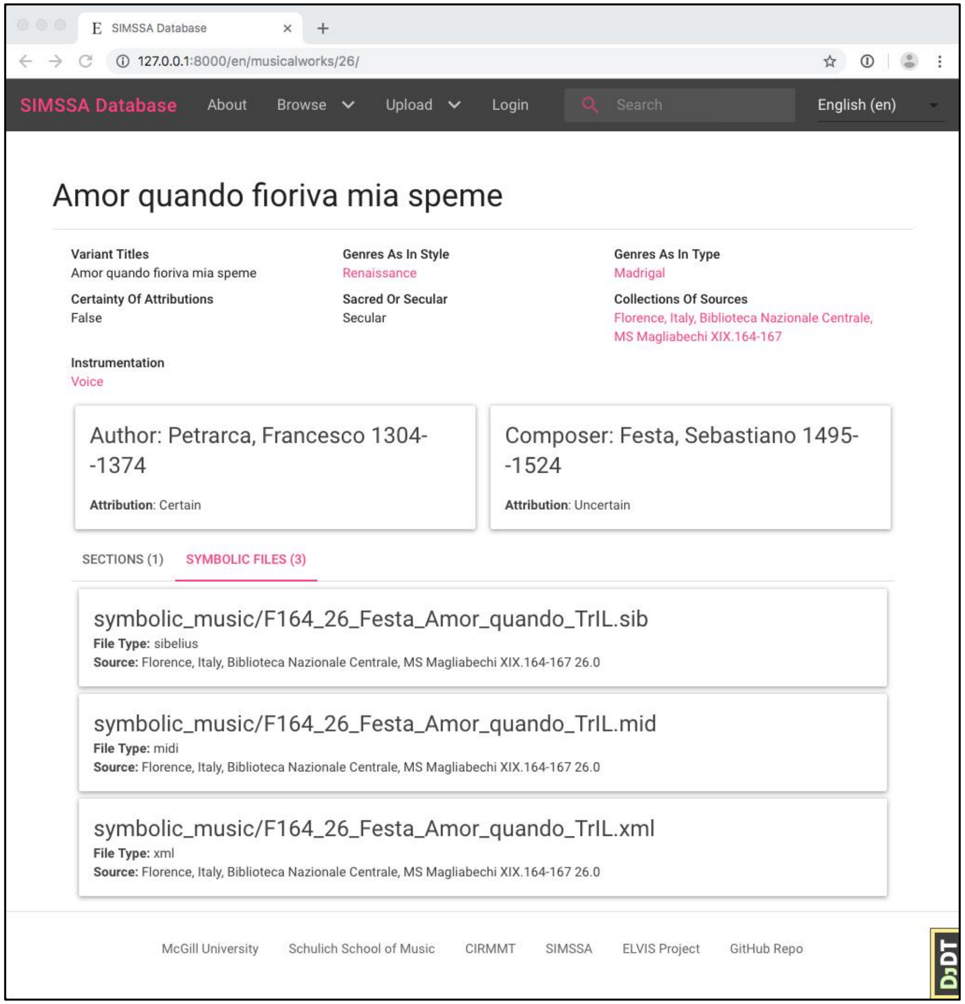
Perch'ai viso d'amor portava insegnia

Composer(s): Festa, Sebastiano 1495--1524

+

Amor che mi tormenti

Sample query: Expanding a work



Amor quando fioriva mia speme

Variant Titles
Amor quando fioriva mia speme

Genres As In Style
Renaissance

Genres As In Type
Madrigal

Certainty Of Attributions
False

Sacred Or Secular
Secular

Collections Of Sources
Florence, Italy, Biblioteca Nazionale Centrale, MS Magliabechi XIX.164-167

Instrumentation
Voice

Author: Petrarca, Francesco 1304--1374

Attribution: Certain

Composer: Festa, Sebastiano 1495--1524

Attribution: Uncertain

SECTIONS (1) SYMBOLIC FILES (3)

symbolic_music/F164_26_Festa_Amor_quando_TrIL.sib
File Type: sibelius
Source: Florence, Italy, Biblioteca Nazionale Centrale, MS Magliabechi XIX.164-167 26.0

symbolic_music/F164_26_Festa_Amor_quando_TrIL.mid
File Type: midi
Source: Florence, Italy, Biblioteca Nazionale Centrale, MS Magliabechi XIX.164-167 26.0

symbolic_music/F164_26_Festa_Amor_quando_TrIL.xml
File Type: xml
Source: Florence, Italy, Biblioteca Nazionale Centrale, MS Magliabechi XIX.164-167 26.0

Provenance

- Keeping a record of provenance is musicologically essential
- Each symbolic music file in the DB is therefore linked to specific source(s) (digital or physical)
- Each source can be linked to its parent source(s) through (eventually) chains of provenance
  - □ e.g. a symbolic MEI file transcribed from a printed score, derived from a hand-written copyist's manuscript, derived from a hand-written original manuscript in the composer's hand

Authority control

- Important for the DB to be able to automatically match differing but equivalent metadata annotations and queries
  - □ e.g. “Stravinsky” and “Stravinski”
  - □ e.g. “Le Sacre du printemps” and “The Rite of Spring”
- The SIMSSA DB uses authority control and cataloguing standards to reduce ambiguity and redundancy (and increase consistency) as much as possible
  - □ The DB is currently using VIAF authority files
  - □ Populates fields with URIs and uses linked open data practices when possible
- Metadata tags are auto-suggested as users type based on these authority files
  - □ e.g. composer name, genre name, etc.

Abstract works, sections and parts (1/2)

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

Abstract works, sections and parts (2/2)

- The SIMSSA DB makes it possible to divide music into abstract works, sections and parts
  - Symbolic files sometimes contain whole pieces, and sometimes only parts of pieces
- This makes it possible to keep track of complex abstract relationships
  - e.g. a movement of one mass might be reused in another mass
  - e.g. an orchestral score and a keyboard reduction of it have different parts, but they are also different versions of the same abstract work

Archiving research dataset

- Facilitating repeatability of research and iterative refinements across research groups are key aspects of scientific music research
- Specific datasets used in specific studies can thus be archived on the well-established Zenodo open research repository
  - These can then be linked to directly from the SIMSSA DB
- Other scholars can then access the precise symbolic music files used in any given study
  - And perform their own research on them

Long-term goals

- Optical music recognition (OMR) integration
- Allow melodic and harmonic queries
  - i.e. local queries, in addition to the global feature-based queries we already have
  - David Garfinkle and Yaolong Ju have started work on this
- Store linked multimodal data (not just symbolic music files)
  - Images of scores or manuscripts
  - Musical texts
  - Audio files

Highlights of the SIMSSA DB

- Designed to meet the specific needs of scholars wishing to engage in large-scale computational musicological research
  - Emphasis on access and usability
  - Web browser interface
- Content-based search centered on features
  - Can also download full sets of pre-extracted feature values
- Free-text and faceted metadata search
- Emphasis on musicologically relevant metadata and data structuring
  - Modeling of complex abstract musical relationships
    - e.g. relationships between (abstract) works, sections and parts
  - Emphasis on provenance
  - Authority control and cataloguing standards
  - Open linked data when possible
- Encourages archiving of specific corpora and studies

Upcoming public release

- The SIMSSA DB is currently undergoing internal user testing
  - □ We want it to be as user-friendly as possible, to meet the specific interface needs of musicologists
- Once this is complete, we will release a beta version to the research community:
  - □ http://db.simssa.ca
- In the meantime, we would be very grateful for any ideas, wants or needs you may have:
  - □ Is there anything you would especially like the SIMSSA DB to be able to do?
  - □ Do you have any music you would like us to host?

Thanks for your attention!

- **E-mail:** cory.mckay@mail.mcgill.ca
- **E-mail:** julie.cumming@mcgill.ca
- **SIMSSA DB:** http://db.simssa.ca

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