



# Summary Features as the Basis for Content-Based Queries of Symbolic Music Repositories

Cory McKay

*Marianopolis College + CIRMMT, Canada*

Julie Cumming

*McGill University + CIRMMT, Canada*

*IAML Conference*

*2022 July 24 to 29, Prague, Czechia*

# Computational musicology and MIR

- **Automated data extraction** software, **statistical analysis** techniques and **machine learning** can help music researchers to:
  - Study **huge quantities of music** very **quickly**
    - More than any human could reasonably look at
  - Empirically **validate** (or repudiate) theoretical predictions using very large corpora
  - Do purely **exploratory** studies of music
    - Examine music from **fresh perspectives**

# Need for symbolic musical encodings

- But to take full advantage of these techniques, researchers need **machine-readable music files**
  - e.g. symbolic music formats like MEI, Music XML, MIDI, kern, Sibelius, Finale, etc.

# Symbolic music repositories (1/2)

- There are relatively few **large research-grade** on-line repositories of such **machine-readable symbolic music**
- Most symbolic music repositories that do exist tend to either:
  - Have inconsistent or unreliable data and metadata
    - **Intended for non-specialist use** rather than rigorous musicological research
  - Be **limited in scope**
  - Have relatively **limited metadata structuring** and only **basic search functionality**
  - Not provide full **open access**

# Symbolic music repositories (2/2)

- Those few large 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 valued such resources are needed** by the research community

# Introduction to the SIMSSA DB

- **Collaborative** database **prototype** **infrastructure** for holding and accessing **symbolic music files**
- Populated by:
  - **Now**: Datasets we have constructed
  - **Medium-term**: Integration of existing symbolic datasets musicologists, libraries and others have constructed for their own purposes
  - **Long-term**: Auto-population via (verified) OMR
- Focused (for now) on **early music**
- Web **browser** interface

# Current status

- Coming out of a two-year hiatus that followed the loss of our development team during the pandemic lockdowns
- Hoping to soon resume internal user-testing and consultation with domain experts

# An infrastructure, not a dataset

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



# SIMSSA DB prototype contribution form

**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 \*:

Variant Titles:

+ -

Sections:

+ -

**Contributions**

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

**Genre(s)**

What type of piece is this? (e.g., song, symphony, motet)

What style is this piece? (e.g., classical, jazz)

Sacred Or Secular:

Not Applicable ▾

**Medium of Performance**

Please enter the instruments or voices below.

Instruments:

SUBMIT

# Data quality

- Focus on **high-quality** data
- Quality of individual documents especially important in **early music**:
  - Individual **details** very important to domain experts
    - e.g. a single cadence or even a single note
  - **Few extant sources**, so limited training/testing data will ever be available

# Searching

- The SIMSSA DB allows two kinds of searching:
  - Free-text or structured metadata searches
    - e.g. title, composer, date, genre, etc.
  - Searches of musical content via features
    - jSymbolic 2.2 features, specifically

# What is a “feature”?

- Information that measures a **characteristic** of a piece of music in a **simple, consistent and precisely-defined** way
- Represented using **numbers**
  - Can be a **single value**, or can be a **set of related values** (e.g. a vector of histogram bin values)
- Provides a **summary** description of the characteristic being measured
  - Usually **macro**, rather than local
- Usually extracted from pieces or distinct sections (e.g. mass movements) **in their entirety**
  - But can also be extracted from smaller **segments** of music

# Example: A simple feature

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



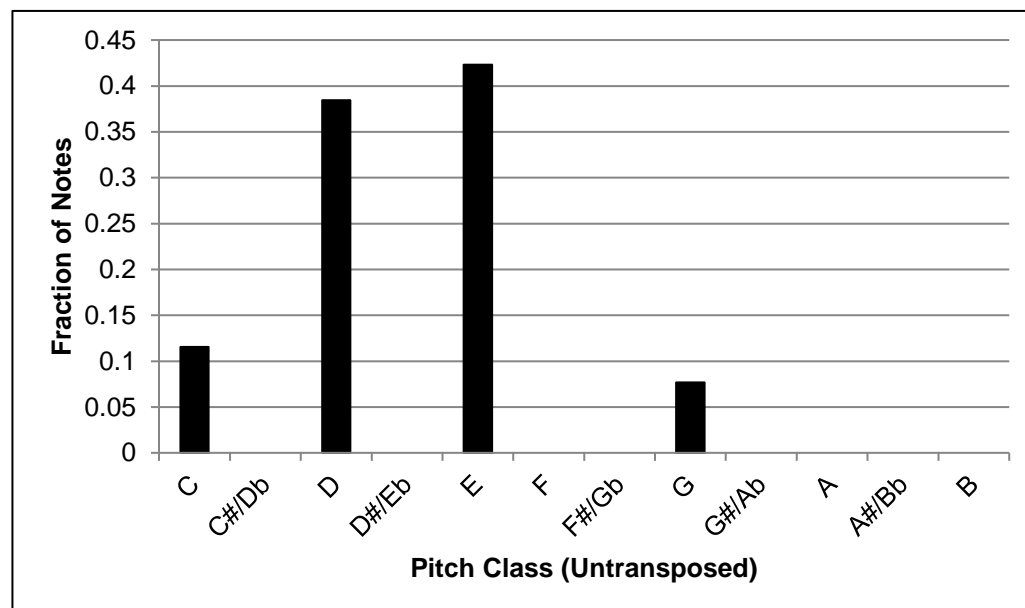
- **Value of this feature:** 7
  - G - C = 7 semitones

# Example: A histogram feature

- **Pitch Class Histogram:** Consists of 12 values, each representing the fraction of all notes belonging to each enharmonic pitch class



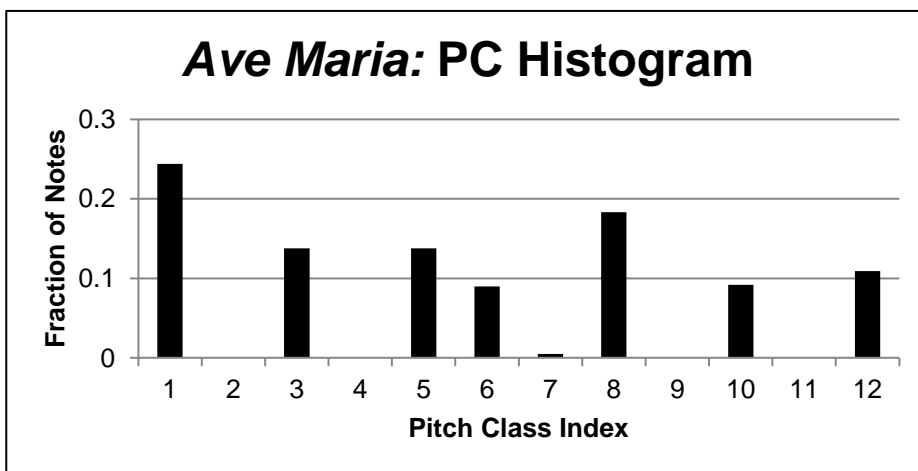
- Histogram graph on right shows feature values
- Pitch class counts:
  - C: 3, D: 10, E: 11, G: 2
- Most common PC is E:
  - 11/26 notes
  - Corresponds to a feature value of 0.423 for E



# Josquin's *Ave Maria . . . virgo serena*

- **Range:** 34 (semitones)
- **Repeated notes:** 0.181 (18.1%)
- **Vertical perfect 4<sup>ths</sup>:** 0.070 (7.0%)
- **Rhythmic variability:** 0.032
- **Parallel motion:** 0.039 (3.9%)

*Ave Maria... Virgo serena*  
Motet  
Josquin Des Prez  
(1440 - 1521)

# jSymbolic 2.2

- The **jSymbolic** research software (McKay et al. 2018) can be used to automatically extract features from digital scores
  - Open source
  - General purpose
- Extracts **246 unique features**
  - Totals **1497 separate feature values**, since many features are multi-dimensional (e.g. histogram vectors)

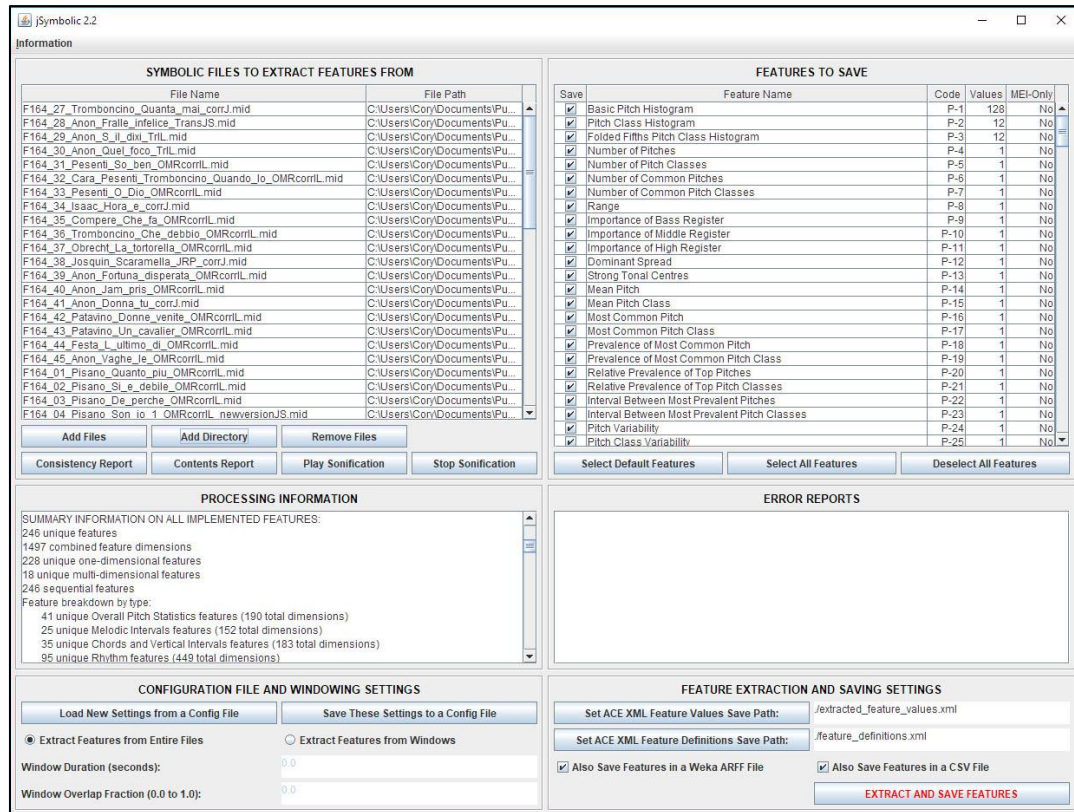


# jSymbolic 2.2's feature types

- Pitch statistics
  - e.g. Range
- Melody / horizontal intervals
  - e.g. Most Common Melodic Interval
- Chords / vertical intervals
  - e.g. Vertical Minor Third Prevalence
- Texture
  - e.g. Parallel Motion
- Rhythm
  - e.g. Note Density per Quarter Note
- Instrumentation
  - e.g. Note Prevalence of Unpitched Instruments
- Dynamics
  - e.g. Variation of Dynamics

# jSymbolic: User interfaces

- Graphical user interface
- Command line interface
- Java API
- Rodan workflow for distributed processing



Information

**SYMBOLIC FILES TO EXTRACT FEATURES FROM**

File Name	File Path
F164_27_Tromboncino_Quanta_mai_corrJ.mid	C:\Users\Cory\Documents\IPu...
F164_28_Anon_Fralle_infelice_TransJS.mid	C:\Users\Cory\Documents\IPu...
F164_29_Anon_S_ii_divi_TritL.mid	C:\Users\Cory\Documents\IPu...
F164_30_Anon_Quel_foco_TritL.mid	C:\Users\Cory\Documents\IPu...
F164_31_Pesenti_So_ben_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_32_Cara_Pesenti_Tromboncino_Quando_lo_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_33_Pesenti_O_Dio_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_34_Isaac_Hora_e_corrJ.mid	C:\Users\Cory\Documents\IPu...
F164_35_Compere_Che_fa_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_36_Tromboncino_Che_debbio_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_37_Obrecht_La_tortorella_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_38_Josquin_Scaramella_JRP_corrJ.mid	C:\Users\Cory\Documents\IPu...
F164_39_Anon_Fortuna_disperata_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_40_Anon_Jam_pris_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_41_Anon_Donna_tu_corrJ.mid	C:\Users\Cory\Documents\IPu...
F164_42_Patavino_Donne_venite_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_43_Patavino_Un_cavaliere_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_44_Festa_L_ultimo_di_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_45_Anon_Vaghe_Le_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_01_Pisano_Quanto_piu_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_02_Pisano_Si_e_debito_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_03_Pisano_De_perche_OMRcorrlL.mid	C:\Users\Cory\Documents\IPu...
F164_04_Pisano_Son_to_1_OMRcorrlL_newversionJS.mid	C:\Users\Cory\Documents\IPu...

**FEATURES TO SAVE**

Save	Feature Name	Code	Values	MEI-Only
<input checked="" type="checkbox"/>	Basic Pitch Histogram	P-1	128	No
<input checked="" type="checkbox"/>	Pitch Class Histogram	P-2	12	No
<input checked="" type="checkbox"/>	Folded Fifths Pitch Class Histogram	P-3	12	No
<input checked="" type="checkbox"/>	Number of Pitches	P-4	1	No
<input checked="" type="checkbox"/>	Number of Pitch Classes	P-5	1	No
<input checked="" type="checkbox"/>	Number of Common Pitches	P-6	1	No
<input checked="" type="checkbox"/>	Number of Common Pitch Classes	P-7	1	No
<input checked="" type="checkbox"/>	Range	P-8	1	No
<input checked="" type="checkbox"/>	Importance of Bass Register	P-9	1	No
<input checked="" type="checkbox"/>	Importance of Middle Register	P-10	1	No
<input checked="" type="checkbox"/>	Importance of High Register	P-11	1	No
<input checked="" type="checkbox"/>	Dominant Spread	P-12	1	No
<input checked="" type="checkbox"/>	Strong Tonal Centres	P-13	1	No
<input checked="" type="checkbox"/>	Mean Pitch	P-14	1	No
<input checked="" type="checkbox"/>	Mean Pitch Class	P-15	1	No
<input checked="" type="checkbox"/>	Most Common Pitch	P-16	1	No
<input checked="" type="checkbox"/>	Most Common Pitch Class	P-17	1	No
<input checked="" type="checkbox"/>	Prevalence of Most Common Pitch	P-18	1	No
<input checked="" type="checkbox"/>	Prevalence of Most Common Pitch Class	P-19	1	No
<input checked="" type="checkbox"/>	Relative Prevalence of Top Pitches	P-20	1	No
<input checked="" type="checkbox"/>	Relative Prevalence of Top Pitch Classes	P-21	1	No
<input checked="" type="checkbox"/>	Interval Between Most Prevalent Pitches	P-22	1	No
<input checked="" type="checkbox"/>	Interval Between Most Prevalent Pitch Classes	P-23	1	No
<input checked="" type="checkbox"/>	Pitch Variability	P-24	1	No
<input checked="" type="checkbox"/>	Pitch Class Variability	P-25	1	No

**PROCESSING INFORMATION**

SUMMARY INFORMATION ON ALL IMPLEMENTED FEATURES:

- 246 unique features
- 1497 combined feature dimensions
- 228 unique one-dimensional features
- 18 unique multi-dimensional features
- 246 sequential features

Feature breakdown by type:

- 41 unique Overall Pitch Statistics features (190 total dimensions)
- 25 unique Melodic Intervals features (162 total dimensions)
- 35 unique Chords and Vertical Intervals features (183 total dimensions)
- 95 unique Rhythm features (449 total dimensions)

**CONFIGURATION FILE AND WINDOWING SETTINGS**

Load New Settings from a Config File | Save These Settings to a Config File

Extract Features from Entire Files |  Extract Features from Windows

Window Duration (seconds): 0.0

Window Overlap Fraction (0.0 to 1.0): 0.0

**FEATURE EXTRACTION AND SAVING SETTINGS**

Set ACE XML Feature Values Save Path: ./extracted\_feature\_values.xml

Set ACE XML Feature Definitions Save Path: ./feature\_definitions.xml

Also Save Features in a Weka ARFF File

**EXTRACT AND SAVE FEATURES**

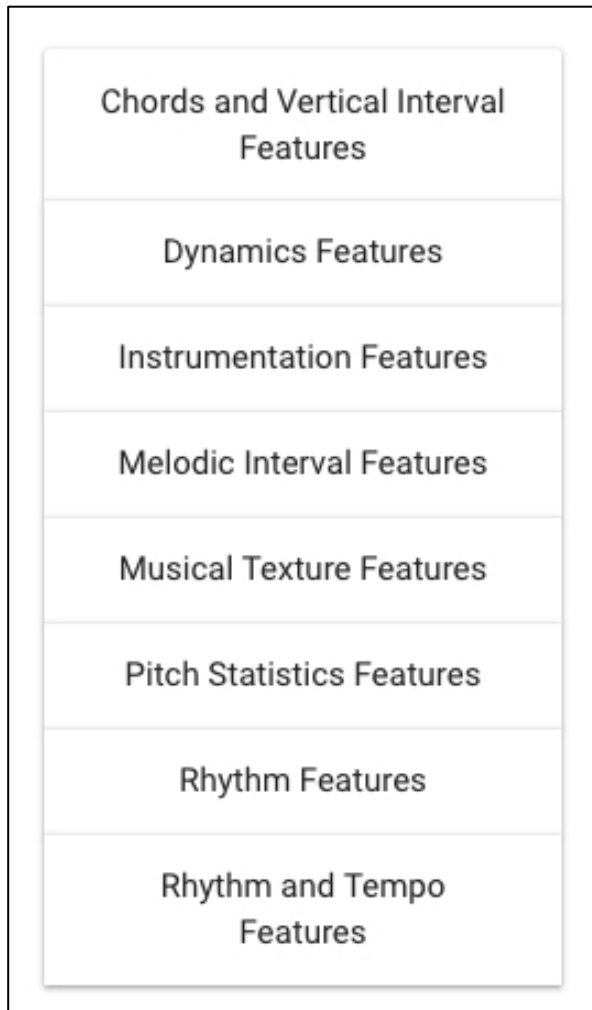
# Sample previous musicological jSymbolic research

- Composer attribution
  - McKay et al. 2017
- N-gram features
  - McKay et al. 2020
- Coimbra manuscripts
  - Cuenca & McKay 2019; Cuenca & McKay 2021
- Ave verum corpus and O decus virgineum
  - Rodriguez-Garcia & McKay 2021
- Ave festiva ferculis
  - Rodriguez-Garcia & McKay 2021
- Morales and Guerrero
  - McKay & Cuenca 2021
- Origins of the madrigal
  - Cumming & McKay 2018; Cumming & McKay 2021
- Buch
  - Cuenca & McKay 2022

# SIMSSA DB and features (1/3)

- jSymbolic has been **integrated into the SIMSSA DB**
  - Whenever a file is uploaded to the SIMSSA DB, features are automatically pre-extracted and stored
- 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 attributed to Josquin that contain parallel fifths

# SIMSSA DB and features (2/3)



A screenshot of the SIMSSA interface showing various feature sliders and checkboxes. The features listed are:

- 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 queries** via a **slider** for each feature they are interested in

# SIMSSA DB and features (3/3)

- Can also download complete feature sets directly and use them as **input to statistical analysis** and **machine learning tools** (or analyze them **manually**)
  - As in the jSybmolic studies referred to earlier

# Metadata 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.
    - Etc.

# Sample query

**Search**

**Genre (Type of Work)**

- Frottola(1)
- Madrigal(8)

**Genre (Style)**

- Renaissance(9)

**Composer**

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

**Instrument/Voice**

- Voice(9)

**Sacred or Secular**

- Sacred(9)

**File Format**

- midi(8)
- sibelius(8)
- xml(8)

**FILTER**

9 results for "amor"

11 files match the feature search parameters. Only **highlighted** files match all search parameters.

Please note that features only apply to valid MIDI, Music XML and MEI files, and will exclude file formats from Sibelius, Finale, etc. For an explanation of all features, please consult the [jSymbolic Manual](#).

**Amore amor quando io speravo**

Composer(s): **Pisano, Bernardo 1490–1548** +

Section(s):

- o Amore amor quando io speravo

---

Genre (Type of Work): **Madrigal**

---

Genres (Style): **Renaissance**

---

Collection(s) of Sources: **Florence, Italy, Biblioteca Nazionale Centrale, MS Magliabechi XIX.164-167**

---

File(s) Holding Complete Musical Work:

- o **sibelius**
- o **midi**
- o **xml**

---

File(s) Holding an Individual Section:

**Hor vedi Amore che giovinetta donna**

Composer(s): **Pisano, Bernardo 1490–1548** +

**Amor se vuoi ch'io torni al giogho anticho**

Composer(s): **Festa, Sebastiano 1495–1524** +

**Chords and Vertical Interval Features**

---

**Dynamics Features**

---

**Instrumentation Features**

---

**Melodic Interval Features**

---

**Musical Texture Features**

---

**Average Number of Independent Voices:**

**1 - 3.804**

---

**Contrary Motion:**

**0.1578 - 0.1924**

---

**Importance of Loudest Voice:**

**0 - 1**

---

**Maximum Number of Independent Voices:**

**1 - 4**

---

**Oblique Motion:**

**0.5066 - 0.6374**



# Provenance

- Keeping a record of **provenance** is musicologically essential
- Each symbolic music file in the SIMSSA DB can therefore be 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

- Should be able to automatically match **differing but equivalent metadata**
  - 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 SIMSSA DB currently uses **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 subsets of pieces
- The **flexible data model** makes it possible to **keep track of complex abstract relationships**
  - e.g. a single movement of a 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 specific research datasets

- In scientific music research, facilitating **repeatability of research** and **iterative refinements** is very important
- Specific datasets used in specific studies can be archived on open research repositories, such as **Zenodo**
  - These can then be linked to directly from the SIMSSA DB
  - The SIMSSA DB can also internally represent specific corpora of collected symbolic music files that were used in specific studies
- Other scholars can access the precise **symbolic music files** used in any given study
  - And perform their own research on them

# Highlights of the SIMSSA DB

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

# Long-term goals

- (Verified) optical music recognition (**OMR**)
- Store **multimodal data** linked to symbolic music files
  - Images of scores or manuscripts
  - Musical texts
  - Audio files
- Formalize **editorial** and **encoding** practices
  - e.g. music ficta, rhythmic note values, etc. in early music
  - (Cumming, McKay, Stuchbery and Fujinaga 2018)
- Allow **local melodic** and **harmonic** queries
  - In addition to the global feature-based queries that the SIMSSA DB already has

# Medium term goals

- Expand the feature set to include the upcoming **jSymbolic 3** features
  - Includes n-gram features
- Use features in more sophisticated ways, such as:
  - Metadata **auto-tagging** using AI-based predictions (with manual verification)
    - e.g. the key of a piece
    - These could then be used in queries
  - **Similarity** measurements
    - e.g. tracking musical influences of composers or individual pieces
    - e.g. investigating contested composer attributions
    - e.g. search by similarity (like Google image reverse searches)
  - Exploratory research using **unsupervised clustering**



# Immediate next steps

- We are hoping to rebuild our development team soon to finalize our first formal release
- We already have a live development (**not** release) version of the SIMSSA DB
  - <http://db.simssa.ca>
  - For internal testing

# Acknowledgements

- Thanks to our previous development team, who did most of the work:
  - Emily Hopkins
  - Gustavo Polins Pedro
  - Yaolong Ju
- Thanks to our many collaborators on the SIMSSA, LinkedMusic, MIRAI, and Construction de la bibliothèque musicale mondiale du 21e siècle projects in general
  - Especially Ichiro Fujinaga
- Thanks for the generous funding of granting agencies:
  - Fonds de recherche du Québec - Société et culture (FRQSC)
  - Social Sciences and Humanities Research Council of Canada (SSHRC)

# Feedback please

- 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?
  - How might you integrate feature-based data or queries into your own work?
  - Do you have any music you would like us to host?

# Thanks for your attention

cory.mckay@mail.mcgill.ca

julie.cumming@mcgill.ca



Social Sciences and Humanities  
Research Council of Canada

Conseil de recherches en  
sciences humaines du Canada

Canada



Fonds de recherche  
sur la société  
et la culture

Québec

