

LinkedMusic, SIMSSA DB and Feature-Based Musicology

Cory McKay, Marianopolis College

(with some slides based on a deck by Ichiro Fujinaga)

Topics

- The LinkedMusic project
- SIMSSA DB
 - Extracting musical features
 - Musicological research with features

LinkedMusic: Scope

- Funded for 7 years (2022–2029): \$3.2M
 - SSHRC Partnership Grant
 - FRQSC Research Team Support Grant
 - Based at McGill
- Broad international involvement
 - 7 co-investigators
 - 18 collaborators
 - 9 partners
 - 4 advisory board members

LinkedMusic: Co-Investigators

- *PI: Ichiro Fujinaga (McGill University)*
- Jennifer Bain (Dalhousie University)
- Housman Behzadi (McGill University)
- *Julie Cumming (McGill University)*
- Debra Lacoste (University of Waterloo)
- *Audrey Laplante (Université de Montréal)*
- *Cory McKay (Marianopolis College)*
- Laurent Pugin (RISM-Digital)

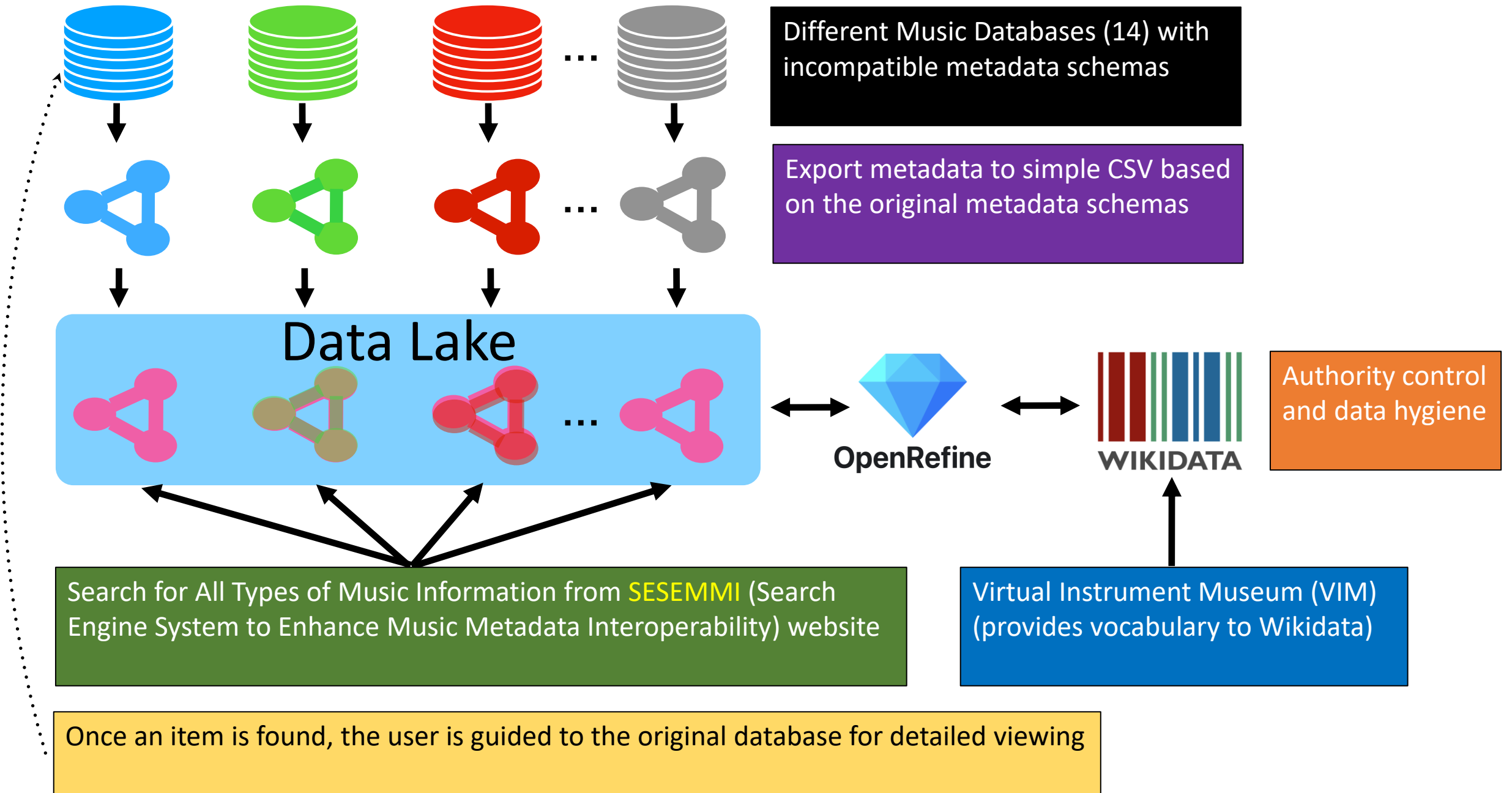
- Many CIRMMT students have also been involved

LinkedMusic: Goals

- Make more musical information accessible to more people in the world
 - With a particular focus on making queries available in languages **other than English**
- Use linked data and semantic web technologies to create a data lake infrastructure allowing one to **search across multiple music databases from one website**
 - Wikidata for authority control
 - OpenRefine to improve data hygiene
 - SPARQL and other search engines (e.g., Solr, ElasticSearch) for queries
- Create a **Virtual Instrument Museum**
 - A crowd-sourced website
 - Images and recordings of musical instruments
 - Name of each instrument in the local language, with translations

Initial 14 databases to import into data lake

1. **SIMSSA DB**
2. Cantus Ultimus
3. Cantus Database
4. DIAMM
5. RISM
6. Cantus Index
7. Canadian Chant Database
8. Global Jukebox
9. DTL1000 (Dig That Lick)
10. MusicBrainz
11. AcousticBrainz
12. CritiqueBrainz
13. ListenBrainz
14. MOTET Database
(Jennifer Thomas)



What is the SIMSSA DB?

- **Collaborative** database **prototype infrastructure** for holding and accessing **symbolic music files**, associated auto-extracted content-based **feature values**, and **musicologically-focused metadata**
 - With a web Django-based browser interface
- Populated by:
 - **Now:** Samples from research datasets we have constructed
 - **Medium-term:** Import existing open symbolic datasets that musicologists, libraries and others have already constructed
 - We can import such datasets, or users can **contribute them directly**
 - **Long-term:** Auto-population via (verified) OMR
- Focused (for now) on **early music**

An infrastructure, not a corpus

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

SIMSSA DB prototype contribution form

Create a Musical Work

Title

Check if the work is already in the database. If so, then select it. If not, then check the "Musical Work not in database" checkbox below and enter the title in the field that appears. Please include opus number or catalogue numbers if applicable (e.g., Op. 55, D960, BWV 202).

Musical Work **not** in database

Title*: [?](#)

Variant Titles: [?](#)

e.g. Eroica

Sections: [?](#)

1. Kyrie

Genre(s)

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

Type not in database

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

Style not in database

Sacred Or Secular:

Medium of Performance

Please enter the instruments or voices below.

Instruments:

Instrument not in database

Contributors [?](#)

Please complete one contributor before adding another. Who created the work? Use the drop-down menu to choose between different kinds of contributions. Add more contributors with the green button.

Contributor's Name:

Person is not in database

Role:

Certainty of attribution:

Certain

Uncertain

Unknown

Location:

Location not in database

e.g. Court of Marie V

Date of Contribution (range):

Core focus: Symbolic music files

- **Research-grade symbolic music files** are surprisingly difficult to access
 - e.g., MEI, MusicXML, MIDI, etc.

Metadata and feature searches

- SIMSSA DB may be searched using traditional metadata queries:
 - **Free-text** search
 - **Faceted** metadata filters, such as:
 - Contributor
 - Composer, arranger, author of text, transcriber, etc.
 - Instruments / voices
 - Sacred / secular
 - Genre / type of work
 - e.g. madrigal, motet, etc.
 - Etc.
- SIMSSA DB also permits **content-based searches** based on **features**

Wait, what is a “feature?”

- Information that **measures a characteristic** of a segment 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 provides a **macro** rather than local view
- 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

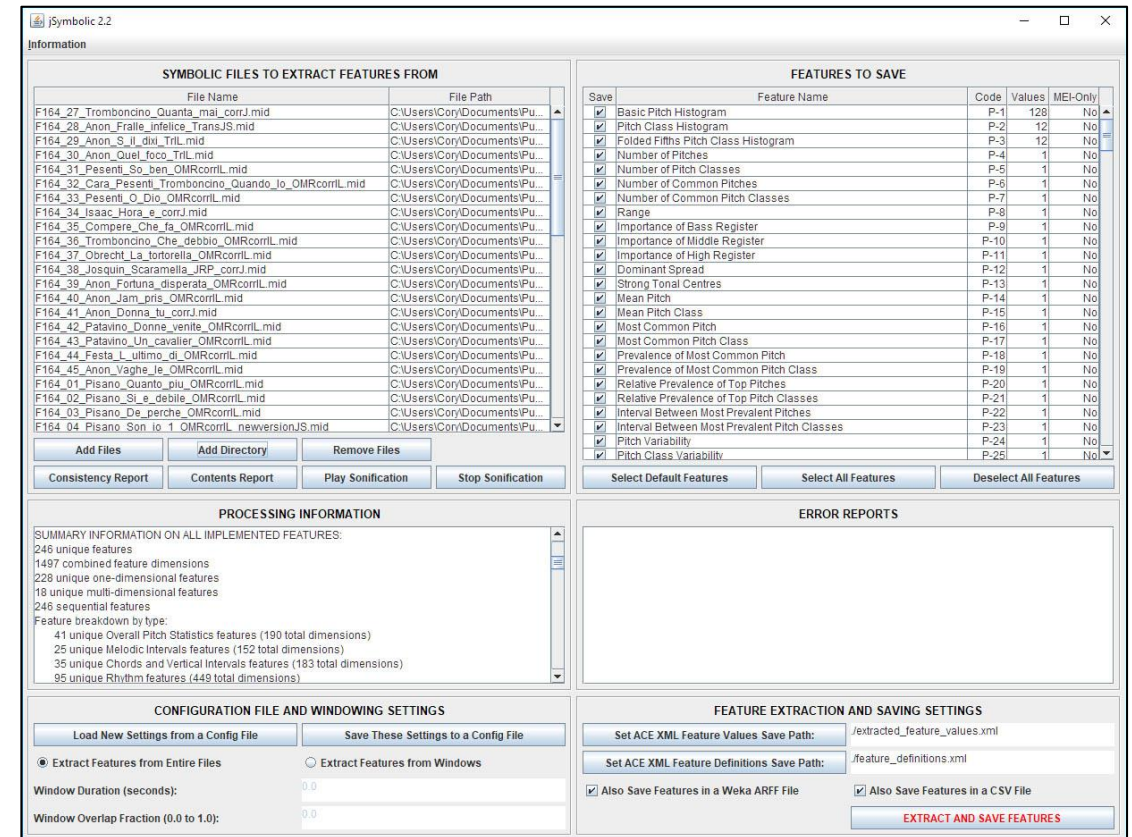
- **Range:** Difference in semitones between the lowest and highest pitches present
 - A 1-dimensional feature



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

How might one calculate features?

- The **jSymbolic** research software (McKay et al. 2018) can be used to automatically extract features from **symbolic digital scores**
 - Open source
 - Applicable to diverse musics
- Version 2.2 extracts **246 unique features**
 - 1497 separate feature values, since many features a multi-dimensional (e.g. histogram vectors)
- The upcoming Version 3 extracts 533 unique features
 - 2040 feature values, including **n-gram features**



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

The screenshot displays the jSymbolic 2.2 application window, which is divided into several functional panels:

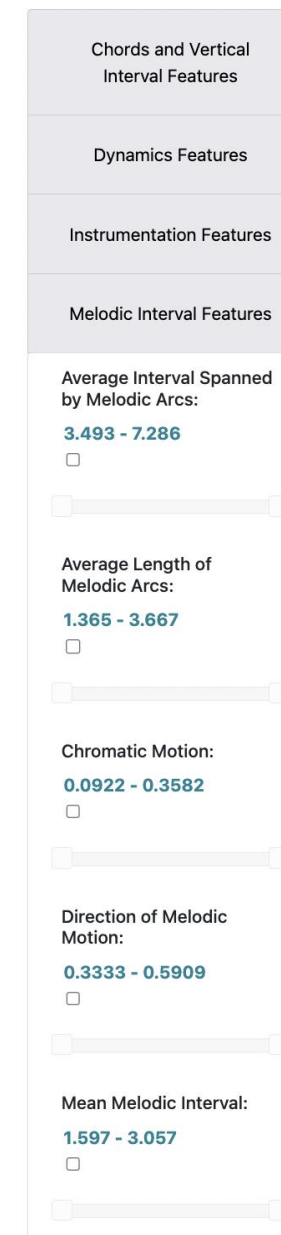
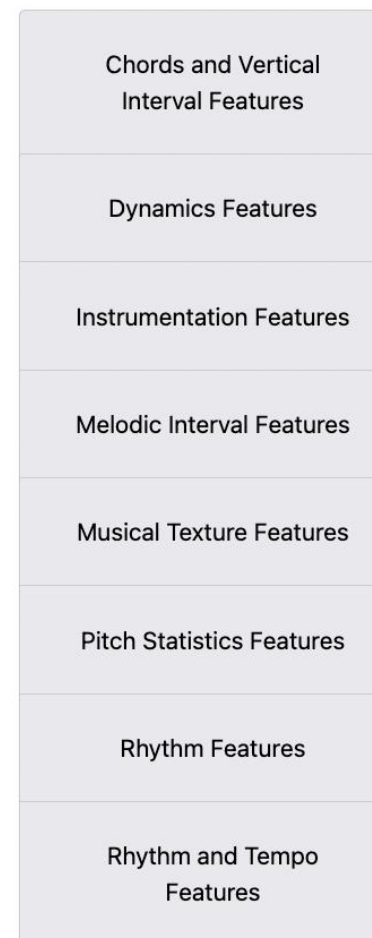
- Information:** A table titled "SYMBOLIC FILES TO EXTRACT FEATURES FROM" listing various MIDI files with their names and file paths.
- FEATURES TO SAVE:** A table listing 25 different features, each with a checkbox, a code (e.g., P-1), a value (e.g., 128), and a "MEI-Only" checkbox.
- PROCESSING INFORMATION:** A text area providing a summary of implemented features, including counts for unique features, dimensions, and a breakdown by type.
- CONFIGURATION FILE AND WINDOWING SETTINGS:** A section with buttons for loading and saving settings, and radio buttons for selecting the extraction scope (entire files or windows).
- FEATURE EXTRACTION AND SAVING SETTINGS:** A section for setting save paths for ACE XML files and checkboxes for saving features in Weka ARFF and CSV formats.
- ERROR REPORTS:** An empty text area for displaying any errors during the process.

Sample musicological feature-based research

- Musical genre
 - Origins of the madrigal (*with Julie Cumming and others*)
 - Delineating popular music genres (*with Ichiro Fujinaga and others*)
- Compositional style (*with Julie Cumming and others*)
 - Empirically differentiating the styles of similar composers
 - Confirming historical evidence for Josquin attribution certainty
- Attribution of anonymous and doubtfully attributed works (*with Esperanza Rodríguez-García and Maria Elena Cuenca*):
 - Masses transcribed by Siro Cisilino
 - Coimbra manuscripts
 - *Ave verum corpus* and *O decus virgineum*
 - *Ave festiva ferculis*
 - Gaffurius Codices
- Regional style in Iberian Renaissance music (*with Maria Elena Cuenca*):
 - Musical influences of Pedro Fernández Buch
 - Musical Influences of Cristóbal de Morales and Francisco Guerrero

SIMSSA DB and features (1/2)

- jSymbolic 2.2 has been integrated into the SIMSSA DB
 - Whenever an extractable file is uploaded to the SIMSSA DB, **features are automatically pre-extracted**, stored and indexed
- Users can specify **feature-range queries** via a **slider** for each feature they are interested in



SIMSSA DB and features (2/2)

- Can **download complete feature sets** directly and use them as input to statistical analysis and machine learning tools (or analyze them manually)
- Feature searches can also be **combined with metadata searches**
 - e.g. retrieve all sacred pieces attributed to Josquin that contain parallel fifths

Sample query combining metadata and features

The screenshot displays a search interface for musical works. On the left, there are several filter sections: 'Search' with the input 'amor', 'Sort By' set to 'Best Match', 'Composition Year From' and 'To' fields, 'Genre (Type of Work)' with checkboxes for Madrigal(8) and Frottola(1), 'Genre (Style)' with Renaissance(9), 'Composer' with Festa, Sebastiano(4), Pisano, Bernardo(4), and Tromboncino, Bartolomeo(1), 'Instrument/Voice' with Voice(9), and 'Sacred or Secular' with Secular(9). A 'File Format' section is partially visible at the bottom.

The main results area shows 9 musical works for the query "amor" and selected facets. The first result is "Amore amor quando io speravo" by Pisano, Bernardo (1490--1548), categorized as Madrigal (Type of Work) and Renaissance (Style). Below the title, there are four file format options: xml, midi, pdf, and sibelius, each with a green plus icon. The second result is "Che deggio far che mi consigli Amore? [2, Pisano, F&H]" by Pisano, Bernardo (1490--1548). The third result is "Hor vedi Amore che giovinetta donna" by Pisano, Bernardo (1490--1548).

On the right side, there is a note: "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](#)." Below this note is a vertical list of feature categories: Chords and Vertical Interval Features, Dynamics Features, Instrumentation Features, Melodic Interval Features, and Musical Texture Features (which is highlighted with a blue border). Under the highlighted category, there are three feature analysis sections: 'Average Number of Independent Voices' with a range of 1 - 3.938, 'Contrary Motion' with a range of 0.079 - 0.2071, and 'Maximum Number of Independent Voices' with a range of 1 - 4. Each section includes a small square icon and a horizontal slider.

Other aspects of the SIMSSA DB

- Chains of provenance
- Conceptual separation between abstract musical works, sections and parts and particular instantiations of them
- Authority control
- Grouping into corpora
- Associations with specific experimental studies
- Links to other types of data (text, audio, images, etc.)

SIMSSA DB: Credit to the deserving

- I designed the original data model and provided high-level guidance to the project, along with **Julie Cumming** and **Emily Hopkins**
- **Gustavo Polins Pedro** and **Yaolong Ju** implemented the first version
- **Rebecca Mizrahi** recently resurrected the DB implemented substantial improvements
- **Hong Van Pham** has worked on deployment and towards LinkedMusic integration
- **Ichiro Fujinaga** generously hosted SIMSSA DB development in his lab

Please try the prototype yourself

- <https://db.simssa.ca>

Thanks for your attention!

cory.mckay@mail.mcgill.ca

