

SIMSSA DB and Related Human Factors

Cory McKay

Marianopolis College

Topics

- SIMSSA DB
 - Extracting musical features
 - Musicological research with features
- Issues:
 - Data quality
 - The human element
 - Long-term operational concerns
- The LinkedMusic project

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 Django-based web browser interface
- Populated by:
 - **Now:** Content 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 have seeded it with datasets we have created, 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: [?](#)

Sections: [?](#)

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

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.
- Most existing scholarly music repositories focus on references to physical sources, to images or audio recordings
 - Many repositories do not reference symbolic music files at all
 - Most of those that do reference symbolic music typically:
 - Treat them as an afterthought, rather than as valuable digital objects worthy of careful consideration
 - Neglect essential issues like **provenance** and **documentation of essential editorial and encoding decisions** that are fundamental to conducting proper computational musicological research
 - Limit the **range of symbolic formats available**, contrary to the needs of researchers who in practice will need music available in a range of different formats (and who know that naïve automatic translation can bias or otherwise compromise research results)

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

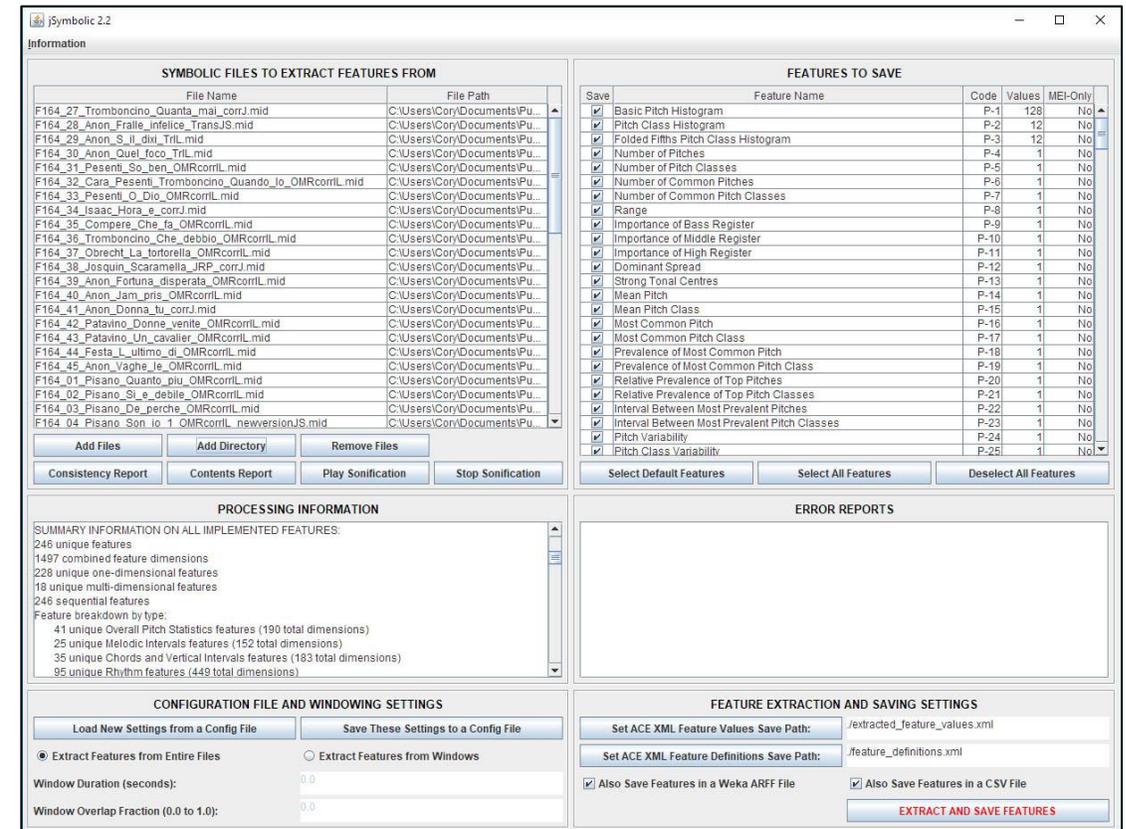
- **Range:** Difference in semitones between the lowest and highest pitches present
 - A 1-dimensional feature



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

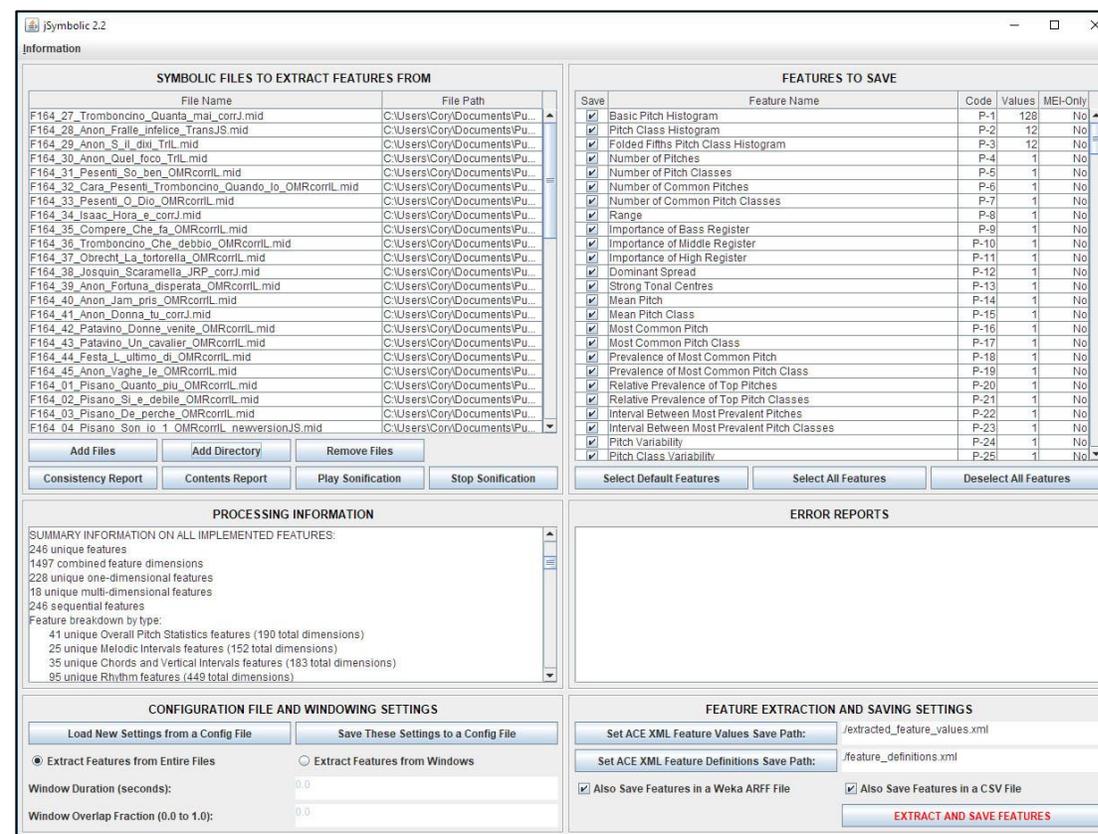
How might one calculate features?

- 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

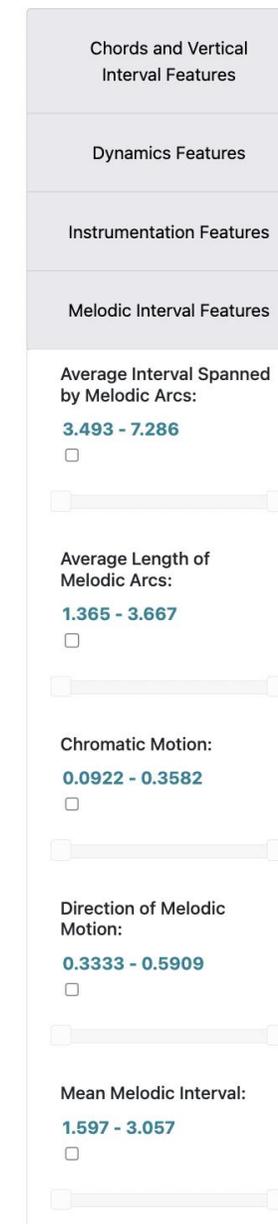
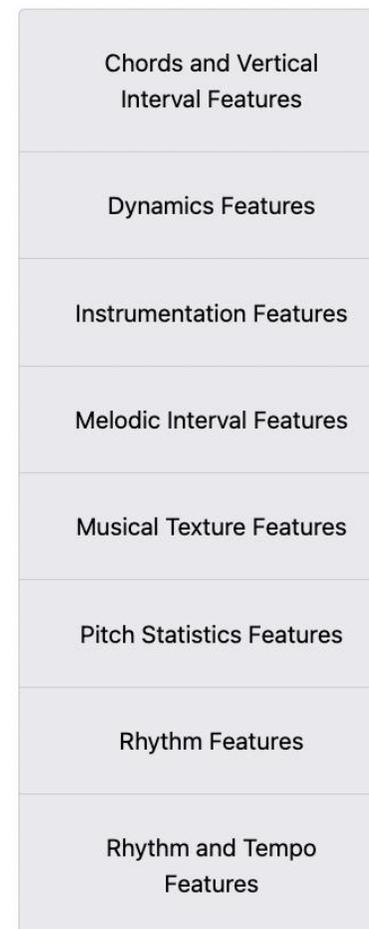


Sample musicological feature-based research

- 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 Maria Elena Cuenca and Esperanza Rodríguez-García*):
 - 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 a symbolic music 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 a given composer that contain tritones

Sample query combining metadata and features

The screenshot displays a search interface for musical works. On the left, there are several filter sections: **Search** (input: amor), **Sort By** (Best Match), **Composition Year From** and **To** (empty), **Genre (Type of Work)** (Madrigal(8), Frottola(1)), **Genre (Style)** (Renaissance(9)), **Composer** (Festa, Sebastiano(4), Pisano, Bernardo(4), Tromboncino, Bartolomeo(1)), **Instrument/Voice** (Voice(9)), **Sacred or Secular** (Secular(9)), and **File Format**. 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 (Renaissance). It lists available file formats: xml, midi, pdf, and sibelius. 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, a sidebar contains a note about feature applicability and a list of feature categories: Chords and Vertical Interval Features, Dynamics Features, Instrumentation Features, Melodic Interval Features, and Musical Texture Features. The "Musical Texture Features" section is expanded, showing "Average Number of Independent Voices" (1 - 3.938) and "Contrary Motion" (0.079 - 0.2071), both with range sliders.

Sources and provenance

- Keeping a record of **provenance** is musicologically essential
- Each digital object in the SIMSSA DB (e.g., a symbolic music file) is therefore linked to a **Source**
 - A “source” is a **reference** (ideally a URI) to a **physical or digital** document from which a digital object in the SIMSSA DB (e.g., a Music XML file) was derived
- Each source can in turn be linked to its parent source(s) through (eventually) **chains of provenance**
 - e.g., an 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
 - Currently uses **VIAF** authority files for genre (type of work) and location
 - Will expand in the future to other fields
 - Populates fields with **URIs** and uses **linked open data** practices when possible
- The goal is to have metadata tags **auto-suggested** as users type based on these authority files when they submit contributions
 - e.g. composer name, genre name, etc.

Other types of digital objects

- The data model is designed to ultimately permit structured access not just to symbolic music files and features extracted from them, but also to related files containing:
 - **Images** (e.g, of the score)
 - **Audio** (e.g., an audio recording of the score)
 - **Text** (e.g., critical edition text annotations)
- Useful for expanding the scope of the SIMSSA DB
 - Particular focus on facilitating integration with frameworks for generating (validated) symbolic music via **OMR**

Other aspects of the SIMSSA DB

- Conceptual separation between abstract musical works, sections and parts and particular instantiations of them
- Grouping digital objects into corpora
- Forming associations with specific experimental studies

Issues: Data quality

- The current focus is on **high-quality** symbolic data
- Quality of individual documents is especially important in **early music**:
 - Individual **details** can be very important to domain experts
 - e.g. a single cadence or even a single note
 - There are often **few extant sources**, so limited training/testing data will ever be available, and there thus limited tolerance for statistical noise
- **Problem**: Ensuring high-quality structured data requires expertise and effort on the part of contributors and validators
 - One of the reasons the SIMSSA DB is designed primarily for use by musicologists and, to a lesser extent, MIR researchers
- This tension between **quantity vs. quality** is not yet fully resolved; we may choose to find a different balance between them in the future
 - In terms of both the **amount of data** and in the amount of **structuring** and **annotation**

Issues: The human element

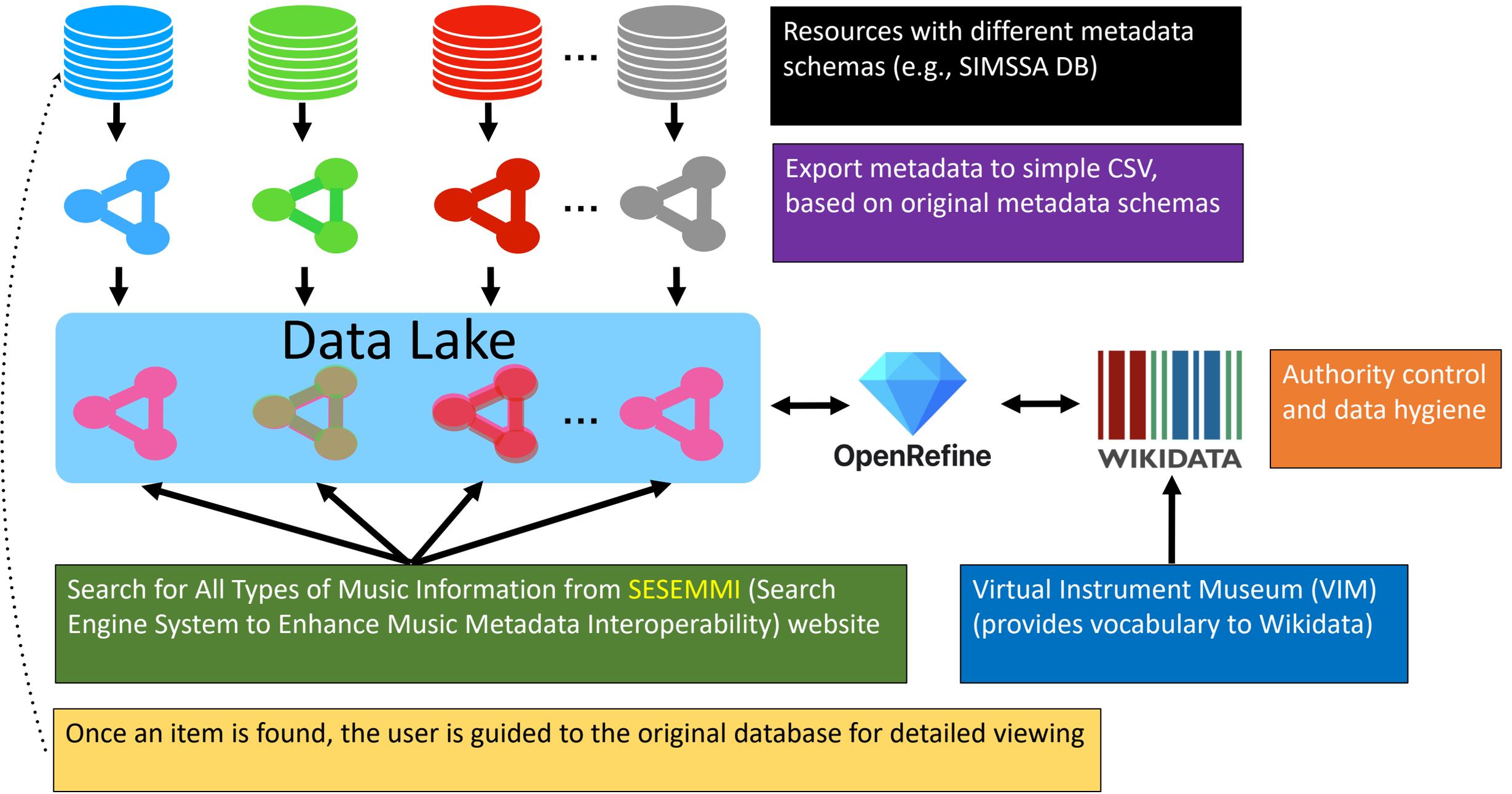
- How can one motivate those not directly involved with the project to:
 - Contribute (high quality) music with (high quality) metadata?
 - Validate data and metadata submitted by others?
 - Use the resource in their own work?
- How can one create a resource that:
 - Meets the needs of those inclined to use it?
 - Needs that might be different from what the creators expect
 - Alerts potential users to new ways of using it they might not have considered?
- **Consultation** with domain experts and potential users of all kinds is essential
 - Throughout, from initial planning to implementation to operation

Issues: Long-term operational concerns

- How does one:
 - Manage administration and maintenance succession once those who created the resource move on?
 - Maintain operational funding after development is complete?
 - Integrate one's own resource with other resources to facilitate access for users who want the benefits of a broad range of resources, but do not want to have to learn and query each resource individually?

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 resources 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 resources 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)

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
 - Including Ichiro Fujinaga (PI), Julie Cumming, Cory McKay
 - 18 collaborators
 - 9 partners
 - 4 advisory board members

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 and implemented substantial improvements, especially in connection to user uploads
- **Hong Van Pham** has worked on deployment and towards LinkedMusic integration
- **Ichiro Fujinaga** generously hosted SIMSSA DB development in his lab

Please try the SIMSSA DB prototype yourself

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

Thanks for your attention!

cory.mckay@mail.mcgill.ca

<https://db.simssa.ca>

