SIMSSA DB: Symbolic Music Discovery and Search

Emily Hopkins (emily.hopkins@mcgill.ca),
Yaolong Ju, Gustavo Polins Pedro, Cory McKay,
Julie Cumming, Ichiro Fujinaga

INTRO

SIMSSA DB is designed to integrate with the optical music recognition workflow developed as part of the SIMSSA (Single Interface for Music Score Searching and Analysis) Project (simssa.ca). However, the popularity of specialized projects such as the Josquin Research Project, KernScores, and the SEILS dataset implied the need for a general-purpose database of symbolic files for researchers to discover, access, and contribute files in a range of formats for any genre or era of music. In development since summer 2017, we aim for a full public release in summer of 2020.

DATA MODEL

- Inspired by DIAMM, RISM, and the IFLA-LRM
- Describing provenance: users can enter a source (e.g., IMSLP) as well as a chain of "parent" sources (e.g. Breitkopf & Härtel, 1862)
- Capture information about software and digitization workflows
- Describe relationships: arrangements, parts, prints, subsections
- Support for symbolic files with plans to support audio and scanned images for OMR

SEARCH & DOWNLOAD

- Content search based on Cory McKay's jSymbolic, which extracts musical features (e.g. range) on upload (MIDI, MusicXML,s MEI)
- Users will be able to batch download search results and use these filters to develop research corpora

METADATA

- Harvesting metadata whenever possible; e.g. from VIAF (Virtual International Authority File) and Library of Congress (Medium of Performance Thesaurus) reduces errors and makes things easier
- Users can also add new values (as in RISM's Muscat)
- Developing workflows with RISM to contribute these new values to RISM (just as they contribute data back to VIAF)
- User expertise can be shared and preserved beyond this database

LINKED DATA

- Building in linked-data compatibility using external resources that assign URIs (e.g. VIAF, Library of Congress Linked Data Service)
- Unite variant spellings, extra names, and different dates with URIs
- Enables semantic queries
- Future work: linked data quads to track metadata provenance
- Triple: Composer—composed—symphony
- Quad: Composer—composed—symphony—according to historian

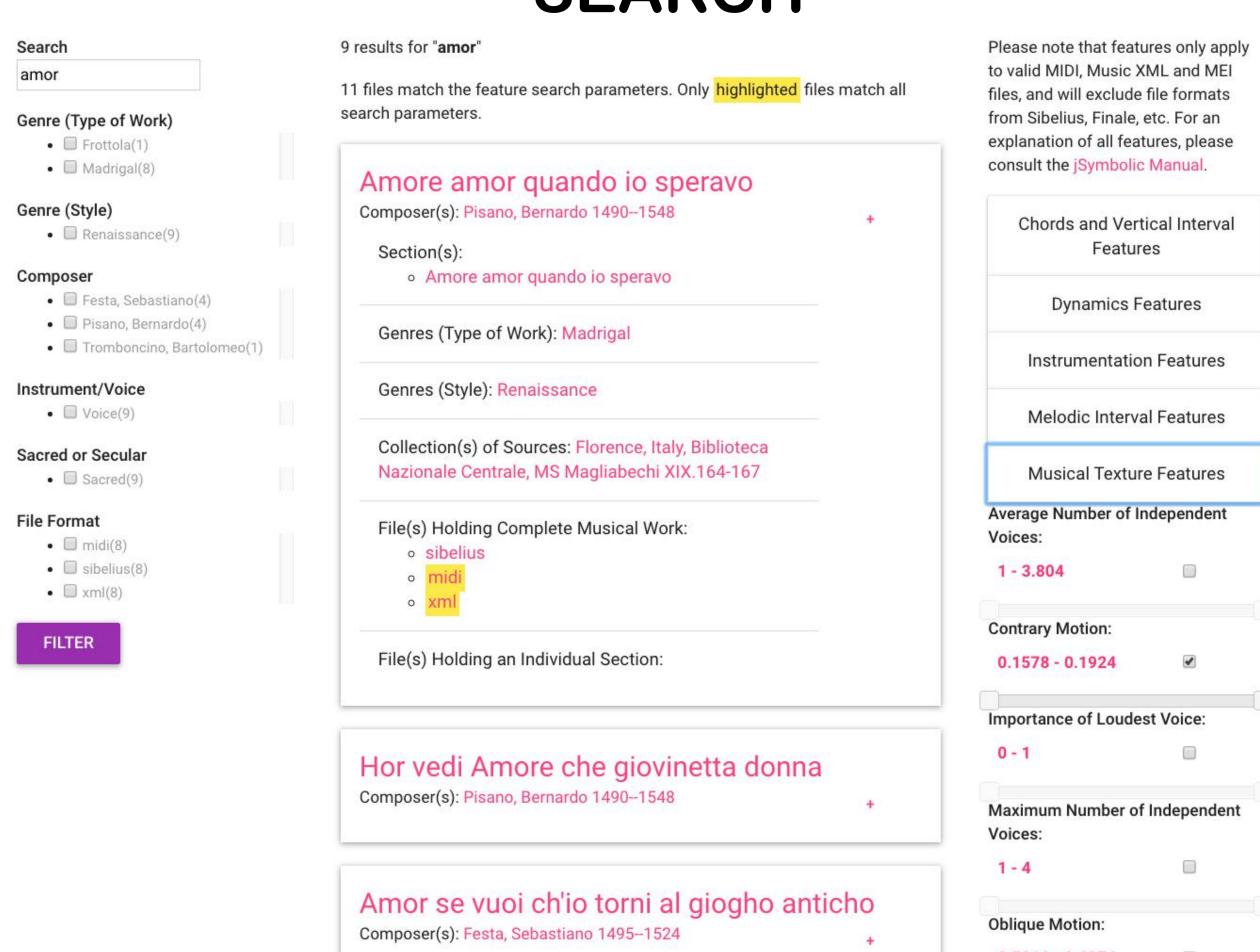
RESEARCH ARCHIVING & REDISTRIBUTION

- Stable data sets required for machine learning and repeatability
- Problem: Users can edit their own contributions, so files may not remain identical over time
- Integrate with external repositories (e.g. Zenodo, Dataverse) to store complete, static version of dataset and link back to DB
- Include licensing metadata for proper redistribution and attribution

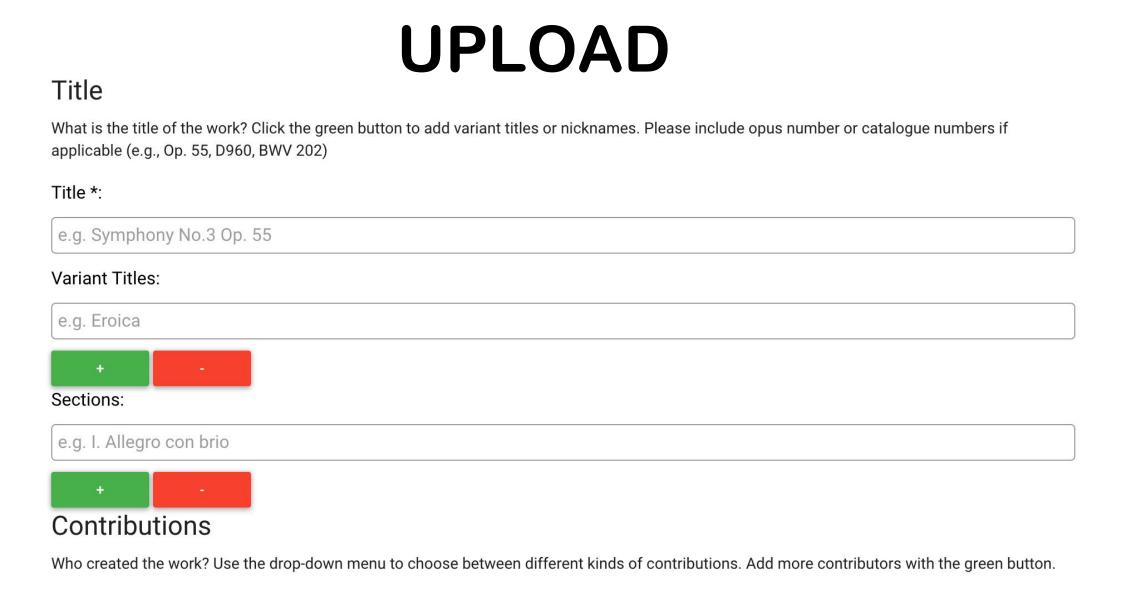
SIMSSA DB improves discovery and access for symbolic music files by combining metadata and musical content search.



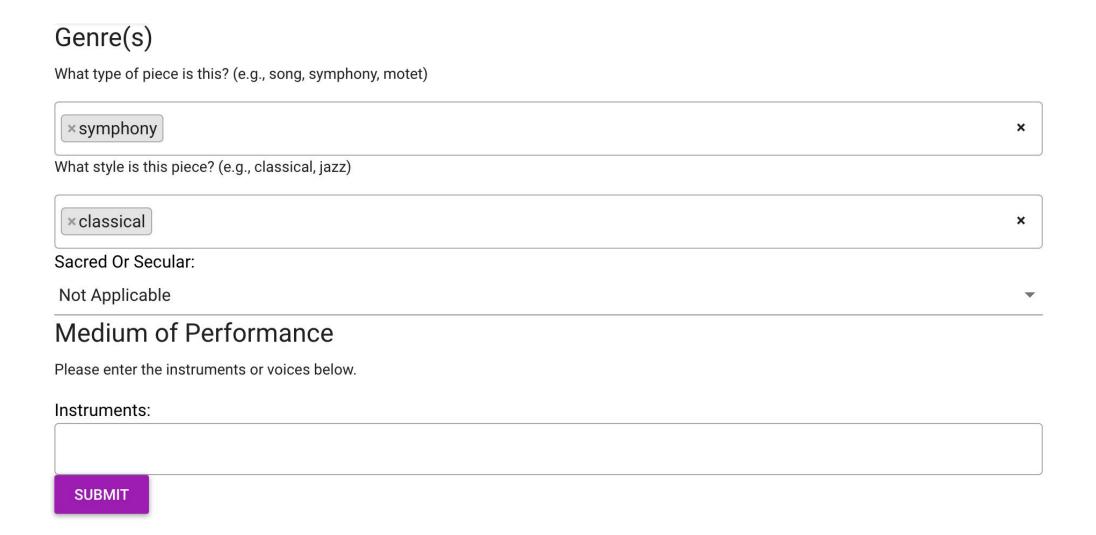
SEARCH



A keyword search for "amor" combined with a search for contrary motion in the features menu. Here, metadata facets are shown on the left, results in the middle, and feature sliders on the right. Feature value slider limits are calculated dynamically according to search results.



Users can enter variant titles or nicknames, and can choose from different roles for contributors, such as composer, arranger, author of text, transcriber, improvisor, and performer.



Genre as in "Symphony" or genre as in "Classical"? We have separated the most common understandings of the term into two fields. Medium of Performance auto-completes from the LC Medium of Performance Thesaurus.

ACKNOWLEDGEMENTS

This research is part of the SIMSSA Project, which is funded by a Partnership Grant from the Social Sciences and Humanities Research Council of Canada (SSHRC 895-2013-1012) and also funded by the Fonds de recherche du Québec-Société et culture (FRQSC).