Enabling Music Search and Analysis: A Database for Symbolic Music Files

As music researchers embrace the possibilities of computational musicology, the need for music in symbolic, machine-readable formats is growing. Symbolic music data is an invaluable resource for computational music analysis tasks such as key finding (Albrecht and Shanahan 2013), cadence detection (Bigo et al. 2018), melodic similarity measurement (Urbano et al. 2011), and harmonic analysis (Condit-Schultz et al. 2018). With the help of automated feature extraction software (e.g., jSymbolic[1]), statistical analysis, and machine learning, symbolic music data allows us to study large quantities of music, empirically test theoretical predictions, and conduct exploratory studies. In order to facilitate this research, we need numerous, high-quality symbolic music files made accessible to researchers via a single, searchable interface.

Although there are several online repositories that provide researchers with access to musical metadata (Bach Digital[2]), audio recordings (Naxos Digital[3]), images of scores/manuscripts (Musiclibs[4]) and mixed formats (IMSLP[5]), few research-grade online repositories of symbolic music exist. Those that do exist have limitations: user-contributed metadata may be inconsistent or unreliable (Classical Archives,[6] Musescore,[7] and ChoralWiki[8]), limited in scope (the SEILS dataset[9]), or designed to support primarily one format (Kern Scores[10]). One particularly high-quality symbolic music repository, the Josquin Research Project,[11] has been used extensively by musicologists and music information retrieval researchers (Brinkman et al. 2016; McKay et al. 2017b), which makes it clear how much such resources are needed by the research community.

In this paper, we present the SIMSSA DB, a large-scale database for research-grade symbolic music files. This database is the next iteration of our earlier ELVIS DB,[12] which featured a user-friendly interface, but was limited by a data model that did not allow enough flexibility to fully accommodate the needs of music researchers.

We began our work on the new database by developing a more flexible and robust data model (McKay et al. 2017a). The relationships among musical works,

---

[1] http://jmir.sourceforge.net/index_jSymbolic.html
[2] https://www.bach-digital.de/content/index.xed
[3] https://www.naxos.com/aboutus.asp?page=naxos-digital-services
[4] https://musiclibs.net/
[5] https://imslp.org/
[6] https://www.classicalarchives.com/
[7] https://musescore.com/
[8] https://www.cpdl.org/wiki/
[9] https://github.com/SEILSdataset/SEILSdataset
[10] http://kern.ccarh.org/
[11] http://josquin.stanford.edu/
[12] https://database.elvisproject.ca/

sections, and parts, along with the associated metadata, can be quite complex. In our new model, musical works can be related to one another (e.g., different arrangements of the same work), and sections of a work can relate to other works or sections (e.g., a cantus firmus used in multiple masses). We drew inspiration from existing music databases such as DIAMM[13] (the Digital Image Archive of Medieval Music) to provide us with a strong foundation. We also made use of existing library cataloging models, such as IFLA-LRM[14] (International Federation of Library Associations-Library Reference Model), in order to gain insights into how best to handle the relationships between conceptual works (e.g., a symphony) and their various expressions (e.g., an image of a printed score or manuscript, a set of parts, and a symbolic file).

The SIMSSA DB also features high-quality metadata. We modeled our structure on RISM's Muscat,[15] and for authority control we connect to VIAF[16] (Virtual International Authority File). We also populate the database with linked data URIs (Uniform Resource Identifiers) when possible, and use controlled vocabularies for genres and instrumentation.[17] The result is that metadata fields in the SIMSSA DB can be auto-completed based on information drawn from authority resources. This helps guard against typographical errors, manages variant spellings of fields, and increases interoperability.

We also record provenance information for all our materials: we describe the sources for each item in the database, both digital and physical, and link sources to the larger collections or archives where they can be found. It is also possible to specify recursive provenance chains for sources themselves (e.g., this symbolic file was digitized from this printed work, which was based on this original manuscript).

This database also introduces an important innovation: content-based music search using features extracted with the jSymbolic software (McKay et al. 2018). A feature is a piece of information that characterizes something in a simple quantitative way. For example, the range of a piece of music, is defined as the difference between its highest and lowest pitches in semitones. The latest version (v. 2.2) of jSymbolic can extract 246 unique features associated with information such as pitch statistics, melody, vertical intervals, texture, rhythm, instrumentation, and dynamics. Researchers can submit queries to the database combining both metadata and features (e.g., retrieve all keyboard pieces composed by J. S. Bach that contain vertical tritones or parallel fifths), download the retrieved symbolic music and use it as the input for statistical analysis and machine learning tools (or use manual analysis) to study various research topics. For example, we have already used feature data in studies on composer attribution (McKay

[13] https://www.diamm.ac.uk/
[14] https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017.pdf
[15] http://www.rism.info/community/muscat/
[16] https://viaf.org/
[17] Library of Congress. 2013. "Medium of Performance for Music Working List of Terminology". Retrieved from: https://www.loc.gov/catdir/cpso/medprf.pdf.

et al. 2017b), genre (Cumming and McKay 2018), and regional styles (McKay 2018), and there remain many areas for further study.

This database will also encourage the archiving of music datasets developed for specific studies. The data model is designed with this in mind, allowing files to be grouped into corpora associated with specific research projects. A given corpus can be linked to a Zenodo[18] repository that includes a static collection of the music files as studied, including associated extracted features, workflows, results, analyses, publications, and other related data. Zenodo assigns a DOI (Digital Object Identifier) to datasets, allowing us to cite them and ensure that resources used in studies remain accessible even as the database expands. As computational musicology continues to grow, it will become increasingly important to facilitate repeatability and refinement of experiments, as well as reusability of datasets.

Overall, the SIMSSA DB permits the storage and distribution of numerous and varied symbolic music files; provides high-quality, meaningfully structured metadata; emphasizes the provenance of resources; facilitates the archiving of research experiments; and offers content-based search. Since it is publicly accessible online,[19] this database will serve as an invaluable resource for the fields of musicology, music theory, and music information retrieval.

[18] https://zenodo.org/
[19] http://db.simssa.ca

References:

Albrecht, Joshua, and Daniel Shanahan. 2013. “The Use of Large Corpora to Train a New Type of Key-finding Algorithm: An Improved Treatment of the Minor Mode.” *Music Perception* 31: (1): 59–67.

Bigo, Louis, Laurent Feisthauer, Mathieu Giraud, and Florence Levé. 2018. “Relevance of Musical Features for Cadence Detection.” In *Proceedings of the 19th International Society for Music Information Retrieval Conference,* 355–61*.*

Brinkman, Andrew, Daniel Shanahan, and Craig Sapp. 2016. “Musical Stylometry, Machine Learning, and Attribution Studies: A Semi-Supervised Approach to the Works of Josquin.” In *Proceedings of the Biennial International Conference on Music Perception and Cognition*, 91–7.

Condit-Schultz, Nat, Yaolong Ju, and Ichiro Fujinaga. 2018. “A Flexible Approach to Automated Harmonic Analysis: Multiple Annotations of Chorales by Bach and Prætorius.” In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 66–73.

Cumming, Julie E., and Cory McKay. 2018. “Revisiting the Origins of the Italian Madrigal.” Presented at the Medieval and Renaissance Music Conference, Maynooth University, Maynooth, Ireland.

McKay, Cory. 2018. “Performing Statistical Musicological Research using jSymbolic and Machine Learning”. Presented at the Anatomy of Polyphonic Music around 1500 International Conference, 34–5.

McKay, Cory, Andrew Hankinson, Julie Cumming, and Ichiro Fujinaga. 2017a. “A Database Model for Computational Music Research”. Presented at the International Workshop on Digital Libraries for Musicology.

McKay, Cory, Tristano Tenaglia, Julie Cumming, and Ichiro Fujinaga. 2017b. “Using Statistical Feature Extraction to Distinguish the Styles of Different Composers.” Presented at *the Medieval and Renaissance Music Conference*, Prague, Czech Republic.

McKay, Cory, Julie Cumming, and Ichiro Fujinaga. 2018. “JSYMBOLIC 2.2: Extracting Features from Symbolic Music for Use in Musicological and MIR Research.” In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 348–54.

Urbano, Julián, Juan Lloréns, Jorge Morato, and Sonia Sánchez-Cuadrado. 2010. “Melodic Similarity through Shape Similarity.” In *Proceedings of the International Symposium on Computer Music Modeling and Retrieval*, 338–55.