A Database Model for Computational Music Research
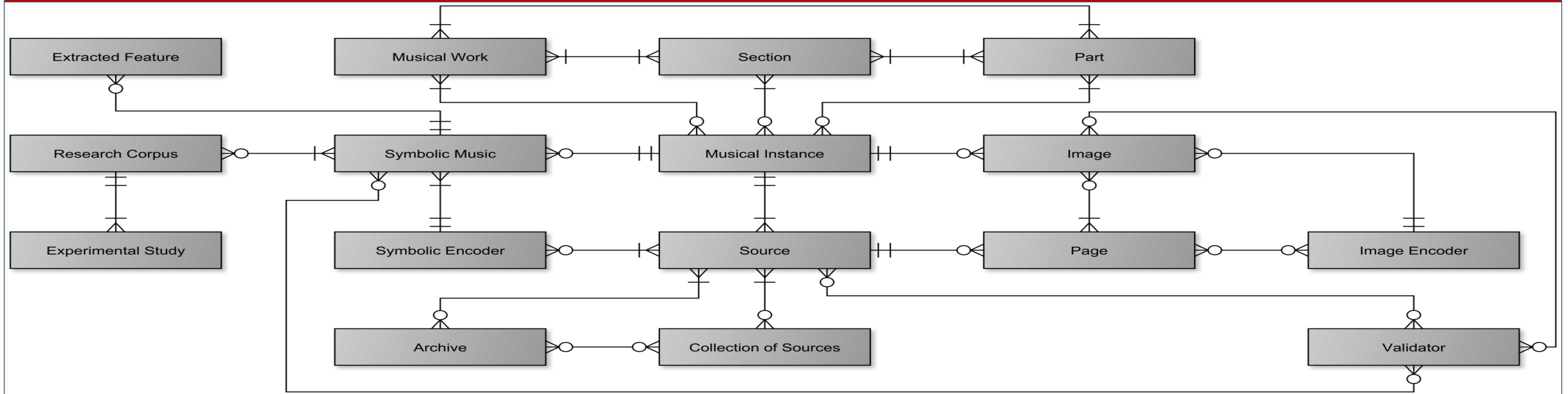
Cory McKay
Marianopolis College

Andrew Hankinson
Bodleian Libraries

Julie Cumming
McGill University

Ichiro Fujinaga
McGill University

Overview Entity Relationship Diagram for the Primary Components of the Data Model

Introduction

Musicologists, theorists, MIR researchers and others interested in empirical music research can benefit greatly from large structured searchable collections of symbolic music in formats like MEI, MusicXML, **kern and MIDI. There is a particular unfilled need for collections that support access to the data used in specific studies, both to ensure reproducibility and to permit direct comparisons of approaches. There is also an unmet need to allow collections of music to be indexed and searched by content, as opposed to just via metadata; one might, for example, wish to retrieve music that has less than a certain amount of chromaticism or parallel motion. Finally, it is important to be able to represent the provenance of objects.

We propose a data model that meets these needs. It further allows digital objects to be conceptualized using a variety of linked layers, such as by work, section, part or source. An emphasis is placed on flexibility and extensibility, while also prioritizing the avoidance of ambiguity and data inconsistencies.

An overview of the core entities of the data model is provided in the red section above. The grey sections to the right describe the meaning of each entity. Space limitations prevent the full data model complete with secondary entities and attributes from being presented here. Information related to provenance and general metadata (e.g. composers, dates, locations, musical notations, text, reliability, workflows, feature metadata, etc.) is thus omitted from this reduced version of the data model.

This is foundational work towards general full-music search and musical content discovery within library collections, and is designed to be integrated with the SIMSSA Musiclibs project, which uses OMR to link symbolic elements to their locations on scans of manuscripts, and to the jSymbolic software, which automatically extracts features from symbolic music. It is ultimately intended to be extended to audio as well.

Structural Observations

- A **Musical Work may** be divided into one or more **Sections** and one or more **Parts**
- A **Section** or **Part** may be reused in multiple **Musical Works**
- **Musical Works**, **Sections** or **Parts** can exist in potentially different forms and combinations in different **Sources**
  - e.g. different arrangements, editions, manuscripts, **Section/Part** subsets, etc.
- Contents of a **Source** can be represented in multiple formats
  - e.g. **Image** or **Symbolic Music** files
- Manifestations like **Sources**, **Images** and **Symbolic Music** must be linked with abstract entities like **Musical Works**, **Sections** and **Parts** in ways that reflect this complexity

Summary of Resulting Insights

- It is best to avoid hierarchically organizing or assigning primacy to any of these entities, as doing so will impose inflexible and potentially contradictory structures
- Abstract **Musical Instances** provide a convenient intermediary that facilitates flexibility while avoiding ambiguity and inconsistency
- Specifying **Encoders** and **Validators** is important for the sake of provenance and reliability
- Grouping **Symbolic Music** files into reusable **Research Corpora** that can be used in **Empirical Studies** facilitates research repeatability and extensibility
- Storing pre-**Extracted Features** enables both content-based searching and re-use of features in **Experimental Studies**

Musical Work

- A complete work of music
  - e.g. a mass, opera, song, symphony, sonata, etc.
- A purely abstract entity that can manifest in differing versions
- Can be divided into one or more **Sections** and **Parts**

Section

- A component or division of a **Musical Work**
  - e.g. an aria of an opera or a movement of a symphony
  - Could alternatively be a **Musical Work** in its entirety
- A purely abstract entity that can manifest in differing versions
- Can potentially exist in more than one **Musical Work**
- Can be divided into one or more **Parts**

Part

- A single voice or instrument in a **Section** of a **Musical Work**
  - e.g. the piano part of a lied
- A purely abstract entity that can manifest in differing versions
- Can potentially exist in more than one **Section** and more than one **Musical Work**

Source

- A reference to all or part of a physical or electronic document that contains the music defining a specific **Musical Instance**, which in turn corresponds to all or part of a **Musical Work**
  - e.g. a printed score, a written manuscript, an audio recording, etc.
- Can potentially be divided into **Pages**
- Can be a component of a **Collection of Sources**
- Can be stored in an **Archive**

Collection of Sources

- A reference to a (physical or electronic) collection of one or more **Sources** grouped together
  - e.g. a book of masses or an album of songs

Archive

- A reference to a (physical or electronic) location where one or more **Sources** or **Collection of Sources** are stored
  - e.g. an on-line database or a monastery library

Provenance

- Indicating provenance is of essential musicological value
  - Including **Sources**, **Collections of Sources**, **Symbolic Music** and **Images**
- Can represent provenance using chains of arbitrary length connecting objects with transformation mappings
  - These objects can be any musical representation
    - e.g. physical scores, images of scores, symbolic music files, audio recordings of performances, etc.
  - An object can be an abstract reference, or can refer to a specific available resource
  - Each transformation mapping should indicate how the object was changed and by what agent(s)
- Each root object (i.e. the earliest known object in a provenance chain) should contain metadata indicating how it was created, by who and with what

Musical Instance

- An abstract entity defined by the music specified by a particular **Source**, which corresponds to a particular instantiation of all or part of a **Musical Work**
  - e.g. the music indicated on a specific score
- Can manifest as **Image** or **Symbolic Music** files

Symbolic Music

- A manifestation of a **Musical Instance** as a digital music file
  - In a format such as MEI, MIDI, Music XML, **kern, etc.
- Generated from a **Source** by a **Symbolic Encoder** via manual entry with a score editor or performance capture system, or via automatic transcription or optical music recognition

Image

- The manifestation of a **Musical Instance** as a digital image of one or more **Pages** of a **Source**
  - In a format such as TIFF, PSD, PNG, JPEG, etc.
  - e.g. a JPEG of a manuscript **Page**
- Generated from a **Source** by an **Image Encoder**

Encoder *or* Validator

- A person or piece of software that encoded a **Symbolic Music** or **Image** file from a **Source**, or that verified their quality
- Includes details on workflow and reliability

Research Corpus

- A collection of **Symbolic Music** files that can be used in specific empirical studies

Experimental Study

- A particular empirical study based on the **Symbolic Music** files in a **Research Corpus**

Extracted Feature

- Content-based data extracted from a **Symbolic Music** file, either manually or automatically with software like jSymbolic
  - e.g. amount of chromaticism
- Includes details on what the feature is and the workflow used to extract it

Implementation Priorities

- Include flexible import/export functionality to facilitate integration with other data repositories
- Make it easy for users to contribute and access data
  - Implement both manual web interfaces and APIs
  - Allow incomplete data, as in practice very few data providers will cooperate if complete data is required
- Include steps to ensure contributions are of a high quality and do not violate copyright
- Adopt some of the principles of Linked Data whenever possible
  - Populate fields with URIs
  - Adhere to the principles of RESTful design

Contact

Cory McKay
Marianopolis College, CIRMMT
cory.mckay@mail.mcgill.ca
https://simssa.ca
http://jmir.sourceforge.net

Acknowledgements