SIMSSA DB: Some details

Cory McKay

Marianopolis College, Canada

CIRMMT, Canada

November 19, 2022. LinkedMusic Project Meeting Montreal, Canada

Review of SIMSSA DB highlights

- Prototype database of symbolic music designed to meet the specific needs of scholars engaging in large-scale computational musicological research
  - Emphasis on tailoring the interface to meet needs of musicologists
- Feature-based search combined with free-text and faceted metadata search
  - Full sets of auto-extracted jSymbolic feature values can also be downloaded
- Emphasis on research-relevant data structuring
  - Modeling of complex abstract musical relationships
    - e.g., relationships between sources and (abstract) works, sections and parts
    - e.g., linking different kinds of musical objects
  - Provenance chains
  - Archiving of specific corpora and features associated with specific studies
  - Authority control and cataloguing standards

Data quality

- Focus on high-quality data
- Quality of individual documents is especially important in early music:
  - Individual details very important to domain experts
    - e.g. a single cadence or even a single note
  - Few extant sources, so limited training/testing data will ever be available and there is limited room 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
  - A quantity vs. quality tension, which will inform ongoing development
    - Both in the amount of data and in the amount of structuring and annotation

Abstract works, sections and parts (1/2)

- The SIMSSA DB maintains a conceptual separation between abstract musical works and particular instantiations of them (as expressed by particular symbolic files, for example)
- Multiple versions of the same abstract work can exist, and these should be both associated with and differentiated from one another
    - e.g. different editions, arrangements, etc. of a work
    - e.g. different digital symbolic encodings of the same manuscript

Abstract works, sections and parts (2/2)

- The SIMSSA DB makes it possible to divide music into abstract Works, Sections and Parts
  - Symbolic files sometimes contain whole pieces, and sometimes only subsets of pieces
- The makes it possible to keep track of complex abstract relationships
  - e.g., a single movement of a mass might be reused in another mass
  - e.g., an orchestral score and a keyboard reduction of it have different parts, but they are also different versions of the same abstract work

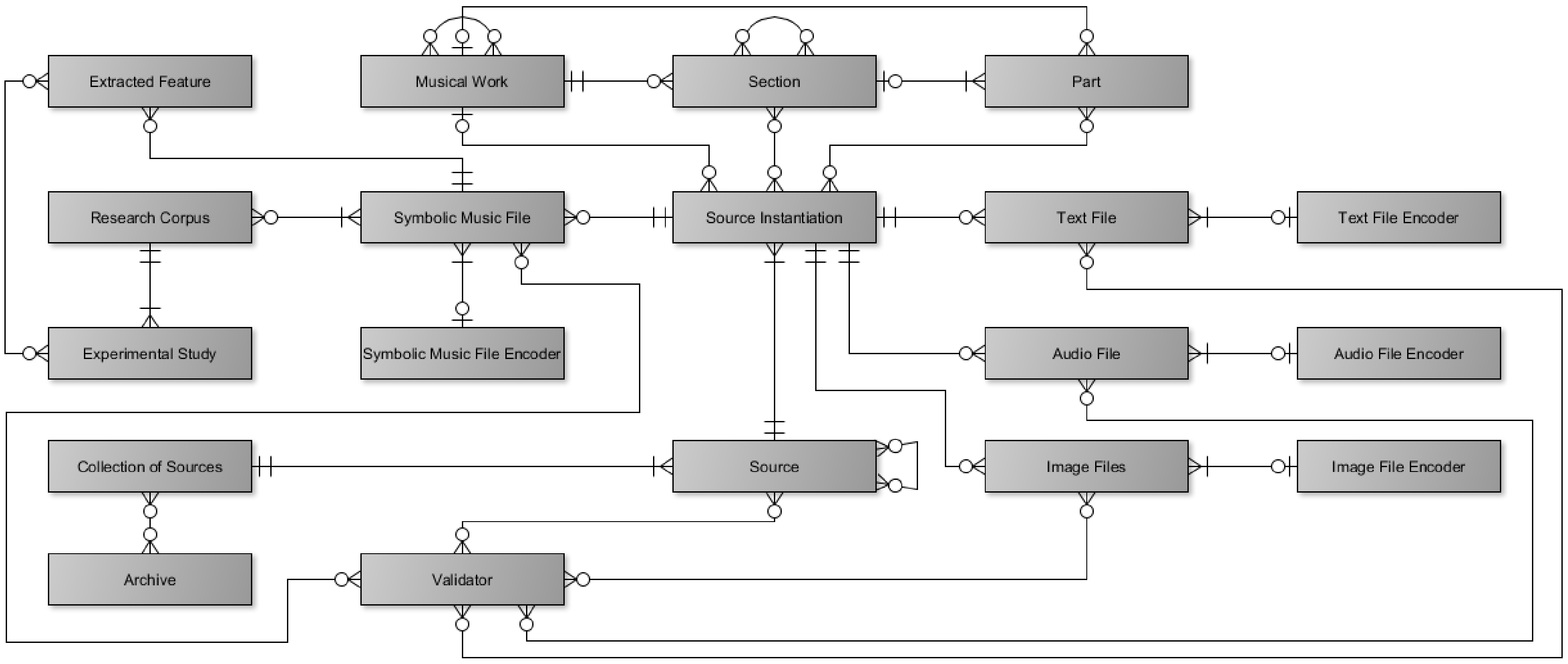
Sources and provenance

- 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 (including, 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., a symbolic 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

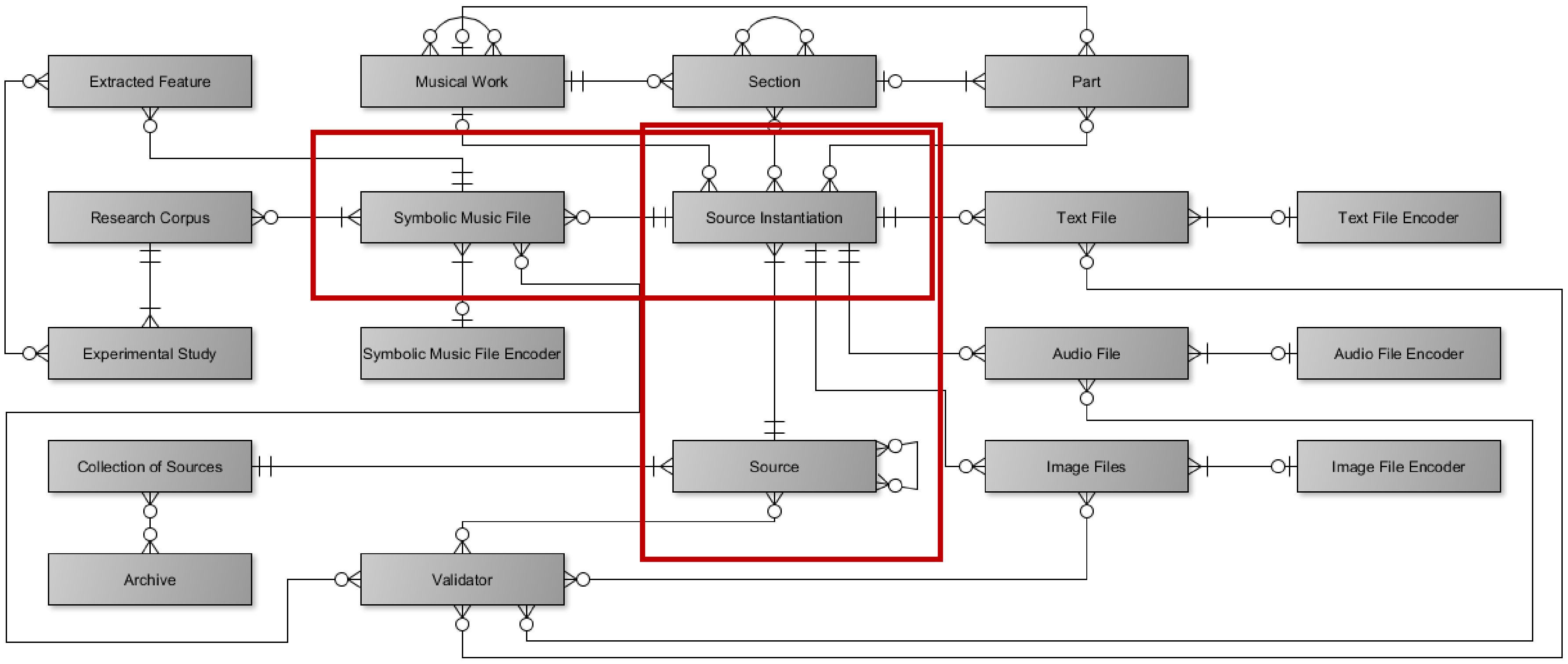
Source Instantiation elements

- Source Instantiation entities link each digital object / Source pair to:
  - Each other *(required)*
  - Abstract Works, Sections and Parts *(optional)*
  - Other digital objects stored in the SIMSSA DB *(optional)*
- A Source Instantiation can encapsulate all of a source or only part of it
  - e.g. an entire score or a single page of a book
- Source Instantiation entities are not exposed to users

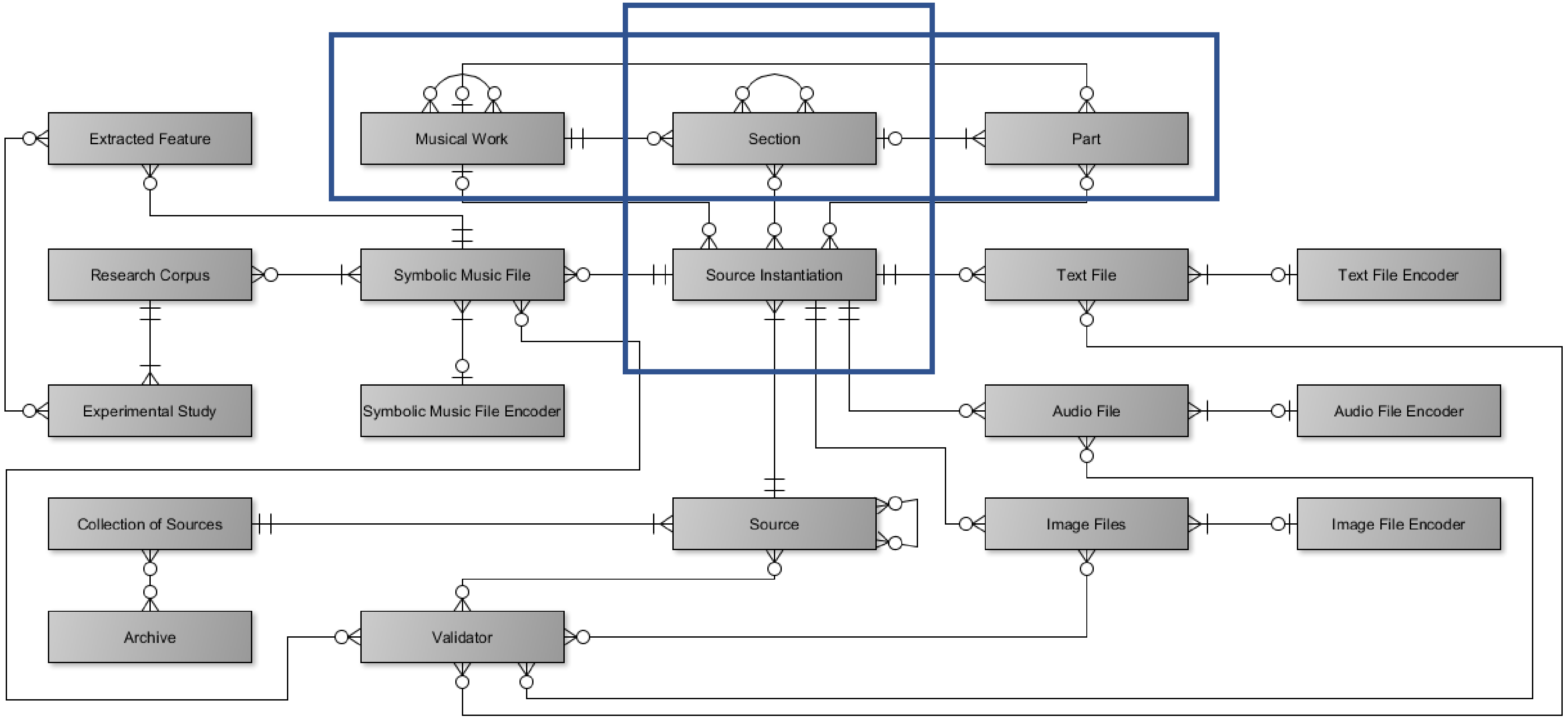
Overview ERD of the SIMSSA DB data model



Overview ERD of the SIMSSA DB data model



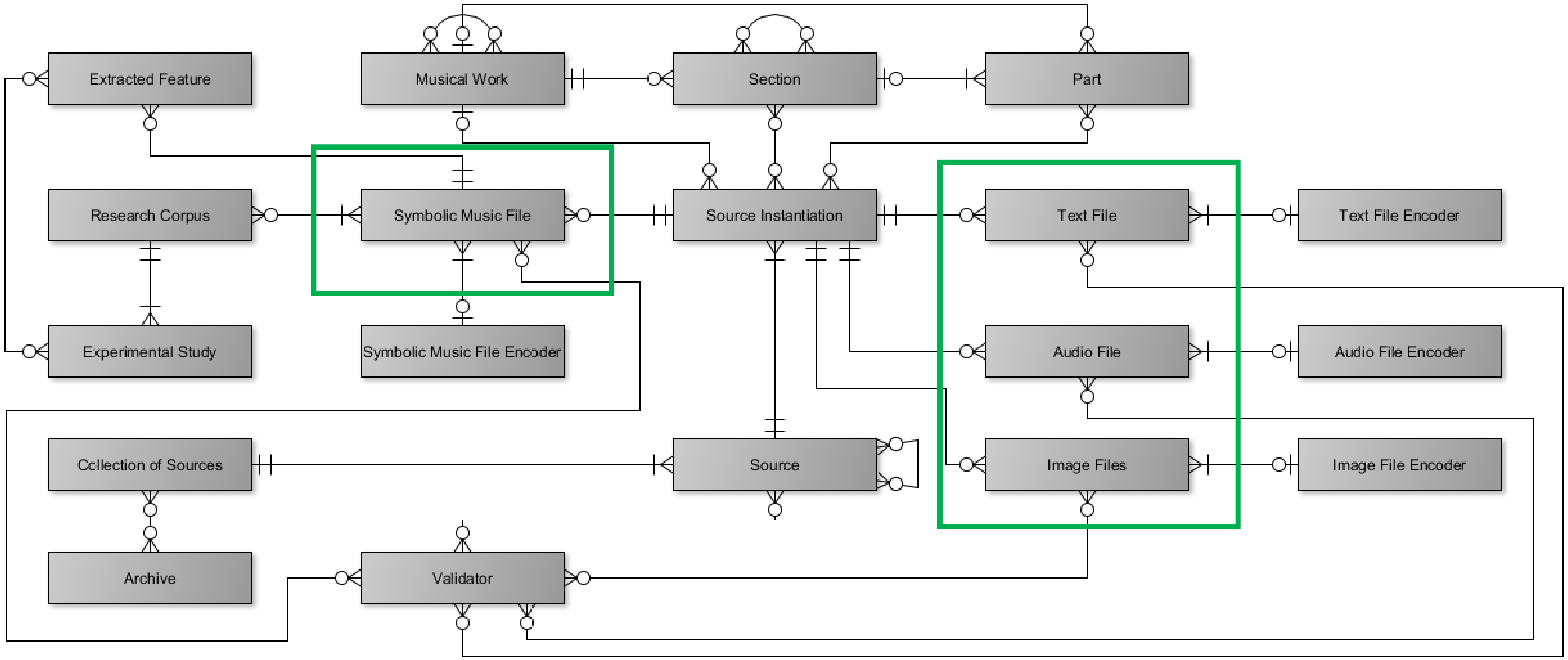
Overview ERD of the SIMSSA DB data model



Other kinds 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
    - Audio
    - Text
- Useful for expanding the scope of the SIMSSA DB
    - Particular focus on facilitating integration with frameworks for generating (validated) symbolic music via OMR
- These are all connected to each other and to sources using Source Instantiation entities

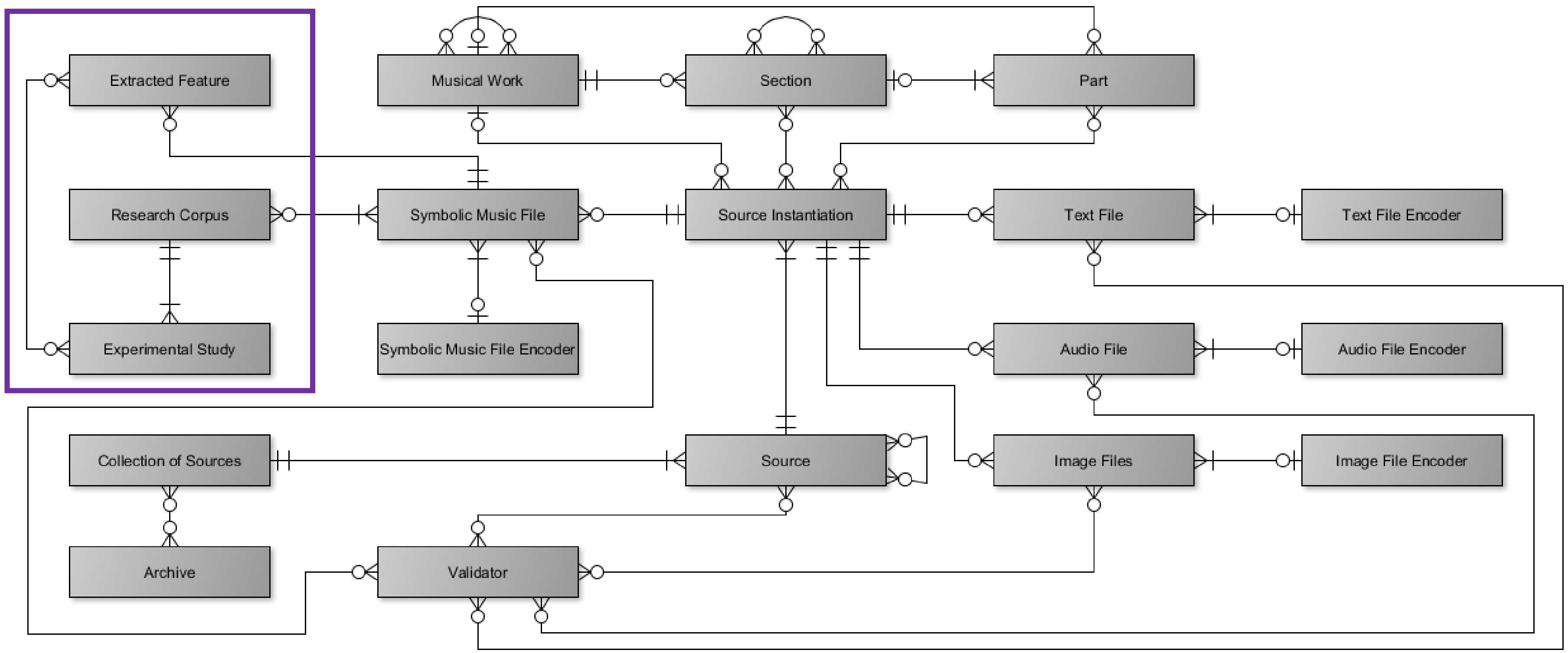
Overview ERD of the SIMSSA DB data model



Archiving specific research datasets

- In scientific music research, facilitating repeatability of research and iterative refinements is essential
- Specific datasets used in specific studies can be archived on open research repositories, such as Zenodo
  - These can then be linked to directly from the SIMSSA DB
  - The SIMSSA DB can also internally represent a specific Research Corpus of collected symbolic music files and features that were used in a specific Experimental Study
- Other scholars can then access the precise symbolic music files and feature values used in a given study
  - Access to such snapshots are important because the both encoding details and feature implementations matter and can change

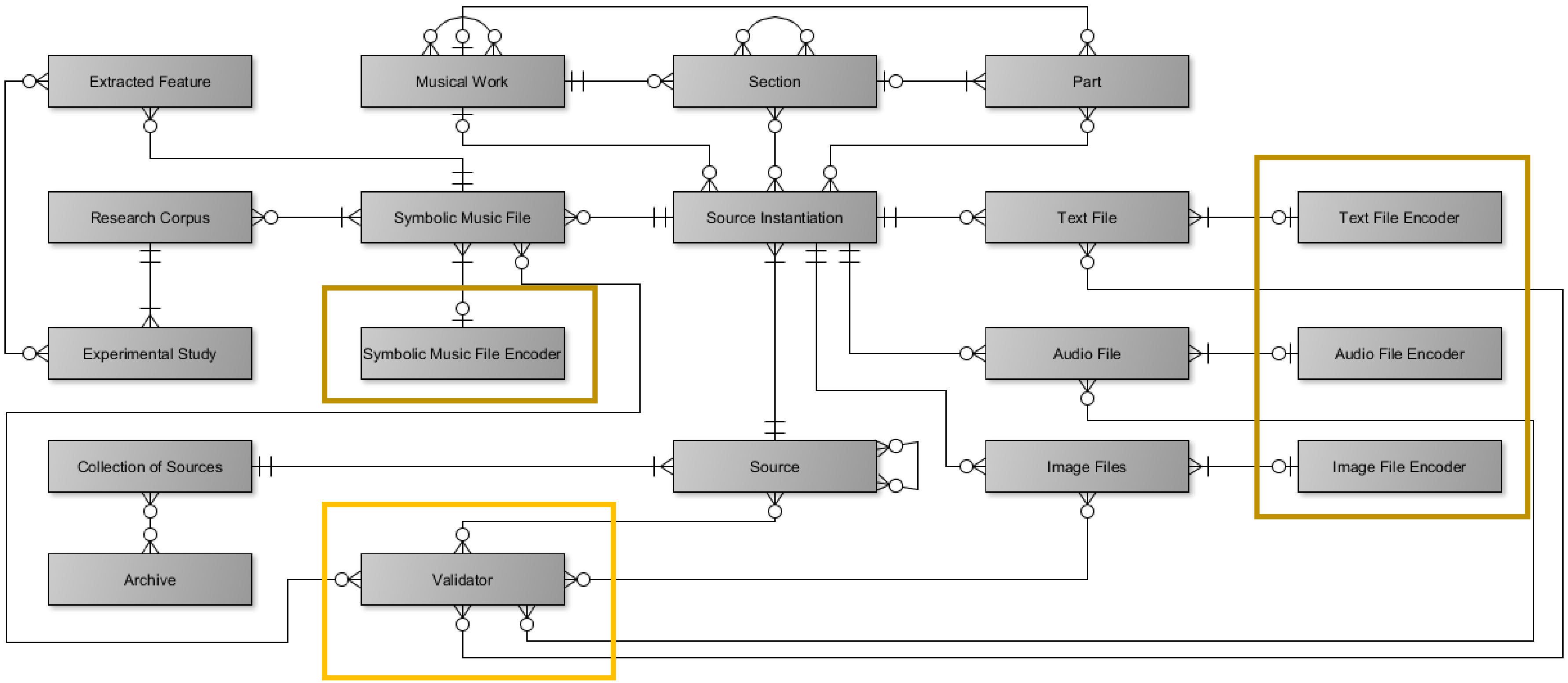
Overview ERD of the SIMSSA DB data model



Quality control and quality control provenance

- Encoder objects keep track of who (or what) generated digital objects stored on SIMSSA DB
- Validator objects keep track of who validated / verified digital objects and metadata about them
    - And sources!

Overview ERD of the SIMSSA DB data model



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
  - Populates fields with URIs and uses linked open data practices when possible
- Metadata tags are auto-suggested as users type based on these authority files when they submit contributions
  - e.g. composer name, genre name, etc.

Medium term goals

- User studies with musicologists to improve the web interface
- Expand the feature set to include the upcoming jSymbolic 3 features
  - Including n-gram features
- Use features in more sophisticated ways, such as:
  - Metadata auto-tagging using AI-based predictions (with manual verification)
    - e.g., modes found in a piece
    - These could then be used in queries
  - Feature-based similarity measurements
    - e.g., tracking musical influences of composers or individual pieces
    - e.g., search by similarity (like Google image reverse searches)
  - Exploratory research using unsupervised clustering

Long-term goals

- Store the product of (verified) optical music recognition (OMR)
  - And associated multimodal data linked to symbolic music files, like images of manuscripts, text extracted from them, etc.
- Formalize editorial and encoding practices
  - e.g. music ficta, rhythmic note values, etc. in early music
  - We have already done some initial work in this direction (Cumming, McKay, Stuchbery and Fujinaga 2018)
- Allow local melodic and harmonic queries
  - In addition to the global feature-based queries SIMSSA DB already has

Live demo

- Not all functionality is enabled in the test version that is currently live
    - e.g., upload is disabled in this version
- https://db.simssa.ca

Feedback please

- We would be very grateful for any ideas, wants or needs you may have:
  - How can SIMSSA DB in general be integrated with your own systems and research?
  - More generally, how might feature-based data or queries be integrated into your own systems or research?
  - Is there anything you would especially like the SIMSSA DB to be able to do?

Thanks for your attention!

cory.mckay@mail.mcgill.ca

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Fonds de recherche sur la société et la culture Québec

MARIANOPOLIS COLLEGE

jMIR

CIRMMT Centre for Interdisciplinary Research in Music Media and Technology

McGill

Schulich School of Music
École de musique Schulich