

ECOLM Futures Review

Chris Cannam
Particular Programs Ltd
London, UK
chris.cannam@particularprograms.co.uk

David Lewis
Goldsmiths, University of London
London, UK
d.lewis@gold.ac.uk

Tim Crawford
Goldsmiths, University of London
London, UK
t.crawford@gold.ac.uk

- **STATUS Preliminary thought-dump**
- **TODO Eliminate almost all of the bulleted lists, turning into prose or something else such as tables**

The ECOLM database as available online contains about 2,000 tablature encodings, manually curated, of relatively high quality with accompanying metadata.

1 BACKGROUND AND MOTIVATION

1.1 What is ECOLM?

ECOLM¹ or “Electronic Corpus of Lute Music” was a series of research projects aiming to develop a queryable online database of lute tablature encodings, of quality suitable for scholarly use.

Two critical, and still relevant, goals were:

- (1) To store and deliver encodings of music, not only metadata;
- (2) To be trustworthy for scholarly use: for example, sources are identified, reliability of attribution is noted, editorial changes are pointed out, and the schema distinguishes between performance and diplomatic transcriptions.

Here we use the name ECOLM broadly to refer to this design of database application, as well as to the past research projects of that name and to the existing system² that they produced.

1.2 Aim and Structure of this Review

We aim to review the premise and outcomes of ECOLM and to consider whether a “lightweight path to sustainability” can be found that can be incrementally extended to other tablature resources.

In section 2 we first set out the history of ECOLM and enumerate other resources of interest. Section 3 identifies some typical users of such resources and describes our findings from user interviews. In section 4 we outline the technical makeup of each of these resources and some related sites of interest. Section 5 summarises the desirable qualities we seek in a solution, and section 6 suggests three possible future courses of action.

2 ECOLM AND OTHER RESOURCES

2.1 The ECOLM Projects

- **ECOLM** (1999-2002) was a project run by Tim Crawford, initially at King’s College London, which produced a queryable database of lute encodings with metadata with a web interface. The resulting service is still accessible today through a public-facing server hosted at Goldsmiths.
- **ECOLM II** (2002-2006) was a successor project which expanded the ECOLM database and used it for some computational musicological investigations.
- **ECOLM III** (2012) was a short project with the goal of adding further high-quality encodings by crowd-sourcing corrections of OMR (optical music recognition) scans.

2.2 Other Lute Tablature Resources

Several other collections of lute music have been collected and placed online by various curators. Of particular interest are:

- **Mss.slweiss.de**³ curated by Peter Steur and the late Markus Lutz. A metadata catalogue of around 68,000 listings of which the majority have incipits (opening ideas) encoded.
- **Lutemusic.org**⁴ curated by Sarge Gerbode. Around 20,000 encodings in playing editions with semi-structured metadata, informally curated with limited version tracking or editorial notes.
- **Lute Society publications** curated by John Robinson. Scans from printed periodicals intended for players, containing around 7,000 encodings consisting of printed music, prose commentary, and semi-structured metadata.
- **Phalèse** curated by Jan Burgers. Around 1,000 encodings transcribed from editions of 16th-century publisher Pierre Phalèse with publication metadata.

There are concerns about the ongoing sustainability of all of these, similar to those about ECOLM: curation and maintenance by individuals or small groups of enthusiasts, in some cases of retirement age; maintenance in limited periods of spare time, perhaps following initial short-term funding; data management using ad-hoc methods or private systems that are not accessible to third-party reproduction; lack of data export facilities or support for common interchange formats.

Therefore, we would prefer to find a solution with the potential to incorporate and maintain data from these resources as well.

The lutemusic.org transcriptions are explicitly Creative Commons NC-SA licensed, and the maintainers of the other listed resources have indicated willingness to contribute to a potential combined dataset.

3 UNDERSTANDING USER CONTEXT

The resources we are considering serve a spectrum of audiences. At one end, lutemusic.org aims at performers and includes edited transcriptions with relatively little scholarly metadata or editorial comment. At the other, ECOLM was aimed at computational musicologists and prioritises diplomatic facsimiles and transcriptions that preserve original scribal idiosyncracies.

In this review we are particularly concerned with sustainability for musicology and other academic purposes. To this end, we conducted informal interviews with three exemplary users of online

¹<http://igor.gold.ac.uk/isms/ecolm/>

²<http://doc.gold.ac.uk/isms/ecolm/database/>

³<https://mss.slweiss.de/>

⁴<https://lutemusic.org/>

Table 1: Status of data and metadata in online lute tablature resources

	Data				Metadata				
	Encoded tablature	OTR scanned pages	Facsimile images	Published PDFs	Works linked to encodings	Ordered work-lists	Textual commentary	Textual references to models	Structured metadata
ECOLM I/II	Yes	Partial	Yes		Yes	Yes		Partial	Yes
Mss.slweiss.de	Yes					Yes			Partial
Lutemusic.org	Yes		Partial	Yes	Partial				Partial
Lute Society	Yes			Yes		Yes	Yes	Yes	
Phalèse	Yes			Yes	Partial	Partial	Yes	Yes	

early-music resources, in order to understand scholarly expectations. These were a “traditional” musicologist, a computational musicologist, and a lute performer and teacher.

3.1 Musicologist

Depending on the material they are looking for, the musicologist we spoke to may begin by searching the RISM⁵ or Cantus⁶ databases. They routinely start with a search by composer or source, because titles tend to have too many historical variants.

They find diplomatic transcriptions (i.e. closely following the source without editorial intervention) the most useful, but are grateful to find any transcription. However, they always refer to the facsimile as well, regardless of the status of any transcriptions, so can often do without editorial notes.

This musicologist was particularly interested in dual tablature and staff renderings, because they are not a specialist in tablature. They would also appreciate the opportunity to annotate or correct unreliable transcriptions for their own use.

3.2 Computational Musicologist

The computational musicologist we spoke to would typically also begin by searching RISM. They trust that metadata in RISM is more authoritative than elsewhere.

They consider trust very important, and appreciate annotations about the original source, transcriber, and editorial interventions. They can work with unreliable transcriptions, if their quality is known and original sources are properly described.

When considering the user interface to a dataset, they appreciate a simple presentation and single search function as their first entry point. They described the ability to refine results via facets as more useful than the ability to construct complex queries from the outset.

The computational musicologist expects the ability to download results (up to possibly the whole dataset) or to query data via API, for use with computational tools such as music21 or Humdrum locally.

3.3 Lute Performer and Teacher

The lute performer and teacher we spoke to would often begin by using the most informal performance resources, simply because they have the most material available. This causes problems cross-referencing with more authoritative material. They find information

about the original source, transcriber, and editorial interventions extremely important, but these are often lacking in performance resources.

As a teacher, they expect students to know the history of the editions they use when performing. They would greatly appreciate something containing modern performing editions, as at lutemusic.org, but with more reliable editorial commentary.

In the absence of trustworthy editorial information, they effectively need to compare every note with the facsimile before making serious use of a transcription.

3.4 Common Threads

Trust and provenance are common themes in discussion with all three of our exemplary users. They have different requirements for content, format, detail of editorial notes and so on, but share a desire to know the quality of transcription and level of editorial intervention they are dealing with.

There was also some consensus about the value of simple search with subsequent refinement, of a cleanly-designed results layout including inline incipits, and of API and data provision.

The musicological specialists were comfortable with RISM and would prefer some level of compatibility, perhaps as far as having the works indexed from RISM and metadata managed there.

None of the three indicated they would hope to *contribute* material to a dataset like this, although they might appreciate the ability to make corrections.

The users mentioned some other sites which they regarded as particularly useful or as good examples to learn from. These are listed in section 4.2 below.

4 TECHNICAL REVIEW

We studied the technical makeup of each of these resources and of other sites of interest, including retrieving data and schema dumps and mapping the schema where applicable. Schema diagrams and accompanying notes are included separately at the end of this document.

4.1 Tablature Resources

4.1.1 ECOLM. ECOLM is a web application written in PHP backed by an SQL database.

Entity relationships are modelled directly in the schema rather than as literal relations in the RDF or triple-store sense. Relations

⁵<https://rism.info/>

⁶<https://cantus.uwaterloo.ca/>

are hardcoded and cannot be changed once the database has been loaded.

A concept known as “clusters” is used to give some support for more general relations within the schema. The most common use is for grouping a number of “pieces” (representations within published sources) into a “work” (a single musical composition).

The schema supports modelling of confidence levels for relations, and dates are modelled with a custom type that supports partially-bounded queries and queries of varying precision.

The same database is used for administrative work (user logins and editorial control) as for content records; there is an expectation that data entry and management are carried out within the same ECOLM application as query and retrieval.

Transcriptions are stored in TabCode, a format devised for the purpose and not widely used elsewhere. Converters between TabCode and other formats are readily available.

The degree of rigour in organisation means that, while it may be tricky to convert or adapt to another format or system, such an effort will probably succeed without too many loose ends.

See section 7.1 for more technical details and a schema diagram.

4.1.2 *mss.slweiss.de*. This is a web application written in PHP and driven entirely from CSV-like files with a semicolon-separated tabular format.

There is a flat directory containing one CSV-like file per source. Separate index files in the same CSV-like format contain manuscript metadata and concordances.

Incipits are embedded in the CSV files, in ABC format, and rendered to SVG from the PHP scripts to be served to the browser. Query capabilities are limited and full transcriptions are not included.

The dataset has been version-controlled since 2013, and is well-organised and looks relatively easy to deal with.

4.1.3 *lutemusic.org*. This is implemented by directly exposing a static file hierarchy through a web server.

The site uses a hierarchical organisation, with separate filesystem trees by composer, source, and facsimile. Composer and source trees contain Fronimo tab transcriptions with derived MIDI and PDF renderings.

The facsimile tree contains images (typically PNG) closely cropped with thresholding, apparently intended for clear reading from screen rather than as historical page facsimiles.

A separate, apparently older, tab hierarchy also present with tab-format files.

The files are indexed using a hand-maintained spreadsheet which contains metadata and the Fronimo file index. There does not appear to be versioning for files or the index.

This design is attractively simple, but its irregular organisation may make adaptation relatively high risk.

4.1.4 *Lute Society*. This consists of facsimiles and transcriptions from the Lute News paper publication, organised by issue number.

The organisation is on filesystem, with PDFs and transcriptions of both text and tablature.

There is a simple front-end available, provided by Tim Crawford⁷ which provides a web index via Javascript requests on the client side.

The content seems well organised. The difficulty is the wide variety of types of material present (including a lot of prose commentary) and the original linear organisation intended for readers and players.

4.1.5 *Phalèse*. We have relatively little technical information about the Phalèse dataset. It is understood to contain transcriptions in Tab format, along with EPS and JPEG facsimiles and documentation mapping the pieces to the original volume and location.

4.2 Other Sites of Interest

4.2.1 *RISM*. RISM⁸ (Répertoire International des Sources Musicales) is a catalogue of musical sources. It “documents what exists and where it is kept”⁹. That is, it is not a library but an index of libraries and the sources they contain. It has a historical focus on physical sources rather than abstract works (although this may be changing) and although some records have incipits attached, it does not otherwise serve musical content.

As we saw in section 3, musicologists routinely expect sources to be indexed in RISM and may expect it to offer the most authoritative source metadata.

The RISM project also publishes a web application for entry and management of musical source catalogue data, called Muscat. See section 7.2 for details about the schema used by Muscat.

RISM records are stored in MARC¹⁰ (Machine Readable Cataloging) format and are available as MARC or RDF data via API.

4.2.2 *DIAMM*. DIAMM¹¹ (Digital Image Archive of Medieval Music) is an archive of scanned manuscripts, mainly from before 1600. It also indexes sources whose images are stored elsewhere. Images are typically of high quality with accompanying metadata describing the physical artifact in some detail.

DIAMM has an API providing metadata in a JSON encoding, though apparently not RDF or JSON-LD.

4.2.3 *Vihuela Database*. The Vihuela Database¹² of John Griffiths is a research-focused index of vihuela music and information about the vihuela. It includes a browseable and searchable list of pieces with incipits in image form and some text commentary. Significant fantasia themes are indexed separately by melody. The site does not appear to offer an API or data linking.

Our consulted lute performer/teacher praised this site for its clear presentation of search results with inline incipits (figure 1) and use of editorial commentary.

4.2.4 *Josquin Research Project*. The Josquin Research Project¹³ from Jesse Rodin and Craig Sapp at Stanford is an index of early polyphonic music with full digitised scores. It is driven from a transparent catalogue of Humdrum-format scores, maintained under

⁷https://doc.gold.ac.uk/mas01tc/jhr_web/

⁸<https://rism.info/>

⁹<https://opac.rism.info/main-menu-/kachelmenu/about>

¹⁰<https://www.loc.gov/marc/>






¹¹<https://www.diamm.ac.uk/>

¹²<https://vihuelagriffiths.com/>

¹³<https://josquin.stanford.edu/>

Figure 1: Vihuela Database search results example

5 records

	fu166	Miguel de Fuenllana	Cobarde caballero [Vásquez]	Orphenica Lyra (1554)	fol. 162v
	mi063	Luis Milán	Aquel caballero madre	El Maestro (1536)	fol. Q1v
	mi063a	Luis Milán	Aquel caballero madre [another version]	El Maestro (1536)	fol. Q2
	mu072	Alonso Mudarra	Gentil caballero	Tres libros de música en cifra (1546)	fol. III/50
	pi004	Diego Pisador	Dezilde al caballero	Libro de música para vihuela (1552)	fol. 4

version control in a Git repository with a submodule per composer and available from the website in-page or via an API.

Our consulted computational musicologist praised this site for its straightforward search interface, presentation of results including useful details such as vocal range plots, and publication of raw data for computational use.

4.2.5 IMSLP / Petrucci. IMSLP (the International Music Score Library Project or Petrucci Music Library) is a very widely used crowd-sourced library of public-domain and Creative Commons licensed sheet music. Managed using MediaWiki, it makes a priority of encouraging community contributions over authoritative editorial review. Works are often available in multiple versions including scans and transcriptions, typically rendered as PDFs rather than in machine-readable form. Depending on the composition, works may feature in full score, parts, and arrangements, and tablature often appears where applicable. There is some linkage to other informal resources such as Wikipedia as well as to formal authorities such as VIAF.

5 DESIRABLE QUALITIES OF A SOLUTION

5.1 Social

- talk about types of user and their expectations
- talk about crowdsourcing and ECOLM III

5.2 Technical

5.2.1 Required.

- ability to continue to absorb upstream changes, for adaptations of datasets that are also being maintained elsewhere
- standard formats where they exist! e.g. always MEI unless there's some very good reason
- automated testing for format conversions (e.g. from external sources)
- data served through API
- stable identifiers for works *and* for transcriptions, so that the latter can be used in principle by other services e.g. similarity
- disambiguation of sources among multiple datasets
- clear relation to identifiers in other sources (online or offline) where available
 - particularly, cross-reference to RISM identifiers for composers and sources (where they exist)

- ability to handle substantial textual and other unstructured data including diagrams and multimedia

5.2.2 Desired.

- RDF
- possibly “immutable pipeline” - rebuildable from source format that is friendly for humans to work with
 - but note the formats that other maintainers actually choose to use! CSV (probably exported from a spreadsheet) and XLS... much as I like e.g. RDF/Turtle, few people want to edit that or JSON-LD directly
- ability to provide more than one front-end
- multiple types of facsimile as well as potentially of transcription—for example if a source has both Gerbode-edited PNG and a detailed scan with limited editing, it would be useful to retain both, with suitable metadata
- RISM indexing compatibility
- natively version-controlled

5.3 User Experience

6 POSSIBLE PATHS

6.1 “Enhanced ECOLM”

Relational database derived from current ECOLM

Advantages

- preserves existing code; first set of data already imported
- somewhat structured, encouraging consistency
- some good domain-specific decisions made in the schema design
- importing other data is work in the mature field of ETL with corresponding tooling etc
- can focus on user interfaces and data conversion rather than rebuilding representation from scratch

Disadvantages

- versioning is difficult
- does not solve the stable identifiers problem
- has proven somewhat overspecified for its current use (e.g. data maturity fields bodged)
- little in common with other solutions people have used, so all import/export is custom
- cannot easily drop in other “views” apart from any we have written ourselves
- providing APIs is manual work

6.2 Graph-based

Fundamental representation is a graph of triples in the RDF mould; all metadata converted to that for import and from it for query; data such as transcriptions, multimedia etc referred to by identifiers in the same space (e.g. URIs)

Advantages

- widely understood model, not least because other systems often make RDF available via API. Although no longer trendy, RDF has not died out and probably won't do soon
- lowest-common-denominator representation as target for data conversions

- not too awful for versioning (even just using git, with Turtle or JSON-LD)
- good target for “idempotent” one-way conversion flows, supporting automated tests for reliability, offering possibility to integrate ongoing upstream changes
- makes serving data via API almost free
- in principle can use existing tools for review, query, inferencing, format conversion

Disadvantages

- in practice using existing tools for inferencing means entering the world of galaxy-brained semantic web savants
- normal people do not choose to hand-maintain data in triples of URIs, so none of our candidate datasets use this format — it’s no CSV
- commonplace ontology mismatch giving “silently missing data” problems on query — good testing strongly advised
- similarly, typeless or with awkward typing
- problem of which ontology to target does not appear to have a single best solution (and we don’t really want to be in the business of developing one)
- no built-in solution to problem of identifying and retrieving bulkier data such as images

6.3 “RISM-aligned”

Fundamental metadata representation is MARC, and we use Muscat to maintain it?

7 DATA REPRESENTATION IN EXISTING SYSTEMS

7.1 ECOLM II

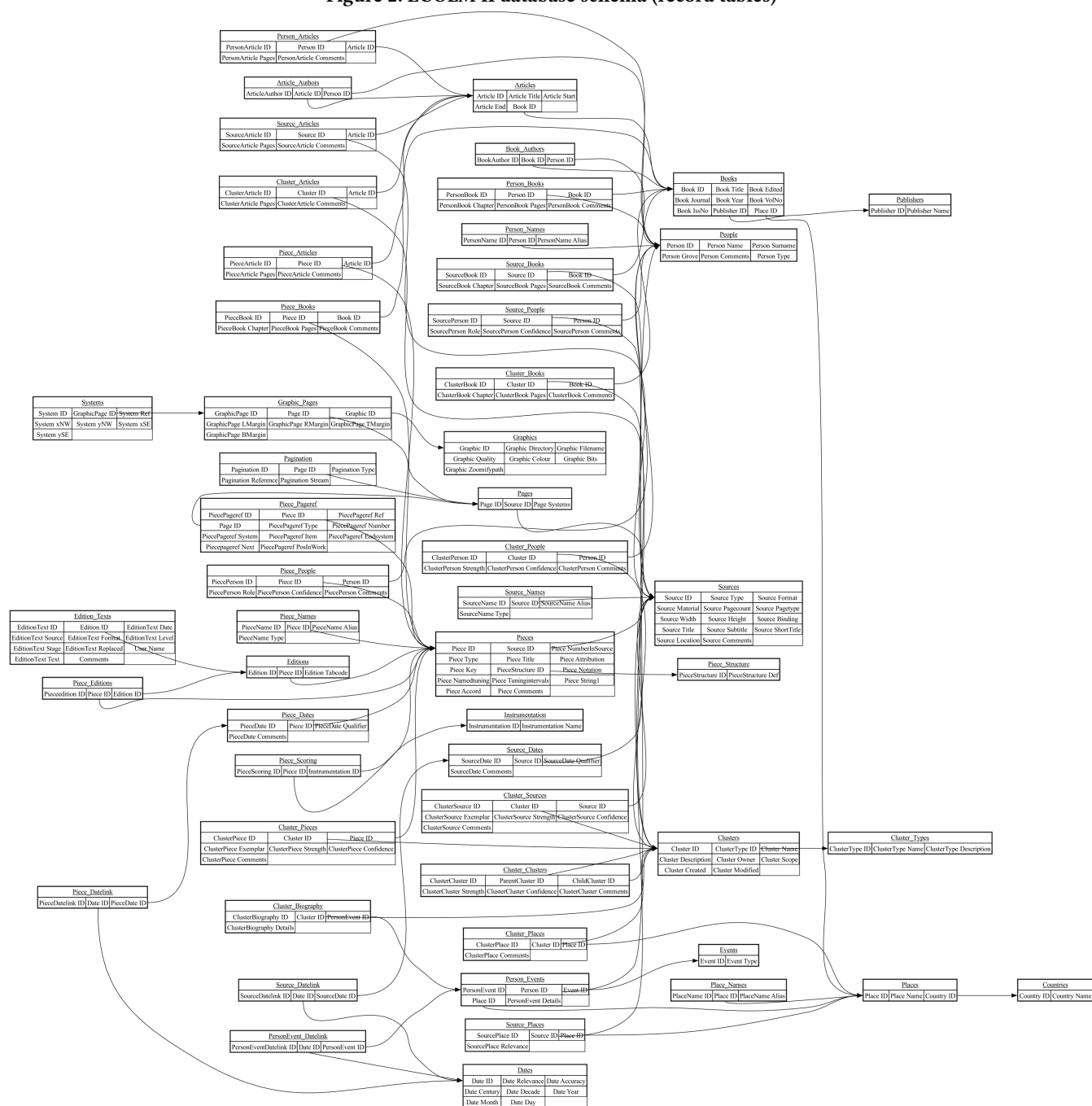
ECOLM I and II store all data directly in a relational database. The database contains both editorial tables, about users of ECOLM and their contributions, and record tables, about the works in the dataset. Figure 2 shows the record tables.

The schema models join relationships using either foreign keys (e.g. `Source_ID` in `Pieces`) or join tables (e.g. `Person_Events`) depending on the presence of metadata about the relationship.

ECOLM has specific definitions of “piece” and “work”. A piece is “a single musical entity within a specified source”, while a work is “a cluster of pieces in different sources that all represent the same musical work”. The database also uses “cluster” to record groups other than works. For example, the ECOLM II database contains only 9 “pieces” directly linked to John Dowland as composer or scribe, but 109 “pieces” linked to him through clusters: 105 as members of the *Lachrimae* “group” cluster, and 4 others through “work” clusters.

The ECOLM schema is unusual today in using mixed-case naming with spaces in the column names.

Figure 2: ECOLM II database schema (record tables)



7.2 RISM Muscat web application

Muscat¹⁴ is a web application published by RISM for cataloguing musical sources, written using Ruby on Rails. Figure 3 shows the main tables.

Muscat uses a hybrid schema, in that each table has a single `marc_source` column containing an authoritative record in MARC21¹⁵ concise text format. Most of the other columns are apparently used to “cache” data from the MARC record that may be needed quickly for display or search. At core, everything is represented using MARC.

Muscat also adds metadata, such as MARC tags, to joins (the arrows in figure 3) through the use of separate join tables.

As an example, the core `sources` table contains columns for numerical RISM source ID, standardised title, manuscript title, composer, shelf mark, language, and date, along with downcased simplified versions of the composer and titles for search purposes. But the authoritative data is found in the `marc_source` column. There are no foreign key relations, as joins are managed through join tables such as `sources_to_people`, `sources_to_sources` etc.

¹⁴<https://rism.info/community/muscat.html>

¹⁵<https://www.loc.gov/marc/>

Figure 3: RISM Muscat schema summary

