ECOLM Futures Review

Status: Preliminary thought-dump

1 Background and Motivation

1.1 What is ECOLM?

ECOLM[1], or "Electronic Corpus of Lute Music", is the name of a series of research projects aiming to develop a queryable online database of lute tablature encodings, of quality suitable for scholarly use.

Two critical design goals are:

1. It gives encodings of music, not only metadata;
2. It is trustworthy for scholarly use: for example, sources are identified, reliability of attribution is noted, editorial changes are pointed out, and it 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 with that name and to the existing system[2] that they produced.

1.2 Aim 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.

The following paragraphs set out the history of ECOLM and enumerate other resources of interest.

1.3 ECOLM History

- **ECOLM** (1999-2002) was a project run by Tim Crawford which produced a queryable database of lute encodings with metadata with a web interface. Presented through a public-facing server at Goldsmiths, the service is still accessible today.
- **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.

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

1.4 Other Lute Tablature Resources

Several other collections of lute music have been placed online by various curators. Some of these are licensed appropriately for inclusion in any future dataset, or have been offered for this use. In some cases there are concerns about their ongoing sustainability that make it particularly desirable to include them in future work.

- **Lutemusic.org**[3] curated by Sarge Gerbode. Around 20,000 encodings in playing editions with semi-structured metadata, informally curated with limited version tracking or editorial notes.
- **Mss.slweiss.de**[4] 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.
- **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.

2 User Context

We conducted informal interviews with three exemplary users of online early-music resources, in order to understand scholarly expectations. These were a "traditional" musicologist, a computational musicologist, and a lute performer and teacher.

2.1 Musicologist

The musicologist gave the following indications about their use of online resources of this type:

[1] http://igor.gold.ac.uk/isms/ecolm/
[2] http://doc.gold.ac.uk/isms/ecolm/database/
[3] https://lutemusic.org/
[4] https://mss.slweiss.de/

- Depending on the material, may begin by searching RISM[5] or Cantus[6] databases;
- Routinely starts with a search by composer or source, since titles tend to have too many historical variants;
- Finds diplomatic transcriptions (i.e. closely following the source without editorial intervention) the most useful, but grateful for any transcription;
- Always refers to the facsimile as well, regardless of status of transcriptions, so can often do without editorial notes;
- Is particularly interested in dual tablature and staff renderings;
- Would appreciate opportunity to annotate or correct unreliable transcriptions.

2.2 Computational Musicologist

The computational musicologist gave the following indications about their use of such resources:

- Will typically begin by searching RISM, and trusts that metadata in RISM is more authoritative than elsewhere;
- Finds trust very important, appreciating annotations about the original source, transcriber, and editorial interventions;
- Can work with unreliable transcriptions if their quality is known and original sources are properly described;
- Appreciates a simple presentation and single search function as their first entry point;
- Finds the ability to refine results via facets more useful than the ability to construct complex queries from the outset;
- Wants the ability to download results (up to the whole dataset) or query via API, to use with computational tools such as music21 or Humdrum locally.

2.3 Lute Performer and Teacher

The lute performer and teacher gave the following indications about their use:

- Will often begin using the most informal performance resources because they have the most material, but this causes problems cross-referencing with more authoritative material;
- Finds information about the original source, transcriber, editorial interventions extremely important;
- Expects students to know those things about the editions they use when performing;
- Would greatly appreciate something containing modern performing editions as at lutemusic.org but with more reliable editorial commentary;
- In the absence of trustworthy information about the transcription, needs to compare every note with facsimile before using.

2.4 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.

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.

3 Technical Review

3.1 Tablature Resources

3.1.1 ECOLM

- SQL database
- Entity relationships modelled in schema—pre-RDF, not triples, relations hardcoded
- Structured using “clusters” (give examples)—attempt to support more general relations within schema
- Confidence levels modelled
- Same database used for user logins / editorial control as for content records—expectation that data managed “within ECOLM”
- 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

3.1.2 lutemusic.org

- Hierarchical organisation with separate trees by composer, source, and facsimile

[5] https://rism.info/
[6] https://cantus.uwaterloo.ca/

- Composer and source trees contain Fronimo tab transcriptions with derived MIDI and PDF renderings
- Facsimile tree contains images (typically PNG) closely cropped with thresholding, apparently intended for reading from screen rather than as historical page facsimiles
- Separate tab hierarchy also present with tab-format files, probably older
- Hand-maintained spreadsheet contains metadata and index for Fronimo files
- Website presents the filesystem hierarchy directly (entirely static, uses web server index pages, nothing generated)
- Irregular organisation may make adaptation relatively high risk

3.1.3 mss.slweiss.de

- PHP web application driven entirely from CSV files (actually semicolon-separated rather than comma-separated)
- Flat directory containing one CSV file per source
- Incipits embedded in the CSV files, in ABC format, rendered to SVG from the PHP scripts for serving to browser
- Separate index CSV files list manuscript metadata and concordances
- Version-controlled since 2013
- Organisation seems tidy and easy to deal with

3.1.4 Lute Society

- Lute News facsimiles and transcriptions organised by issue
- Organisation is on filesystem, with PDFs and transcriptions of both text and tablature
- Simple front-end added by Tim Crawford[7] provides a web index via Javascript requests from client
- Seems well arranged, the biggest problem looks like the wide variety of types of material present and the original linear organisation for readers and players

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

3.1.5 Phalèse

3.2 Other Related Sites

3.2.1 RISM

3.2.2 Vihuela Database

3.2.3 Josquin Research Project

3.2.4 DIAMM

3.2.5 earlymusicsources.com

3.2.6 IMSLP

4 Desirable Qualities of a Solution

4.1 Social

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

4.2 Technical

- ability to continue to absorb upstream changes, for adaptations of datasets that are also being maintained elsewhere
- possibly “immutable pipeline” - rebuildable from source format that is friendly for humans to work with
  - 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
- stable identifiers for works *and* for transcriptions, so that the latter can be used in principle by other services e.g. similarity
- clear relation to identifiers in other sources (online or offline) where available

4.3 User Experience

5 Proposed Path

6 Data representation in existing systems

6.1 ECOLM II

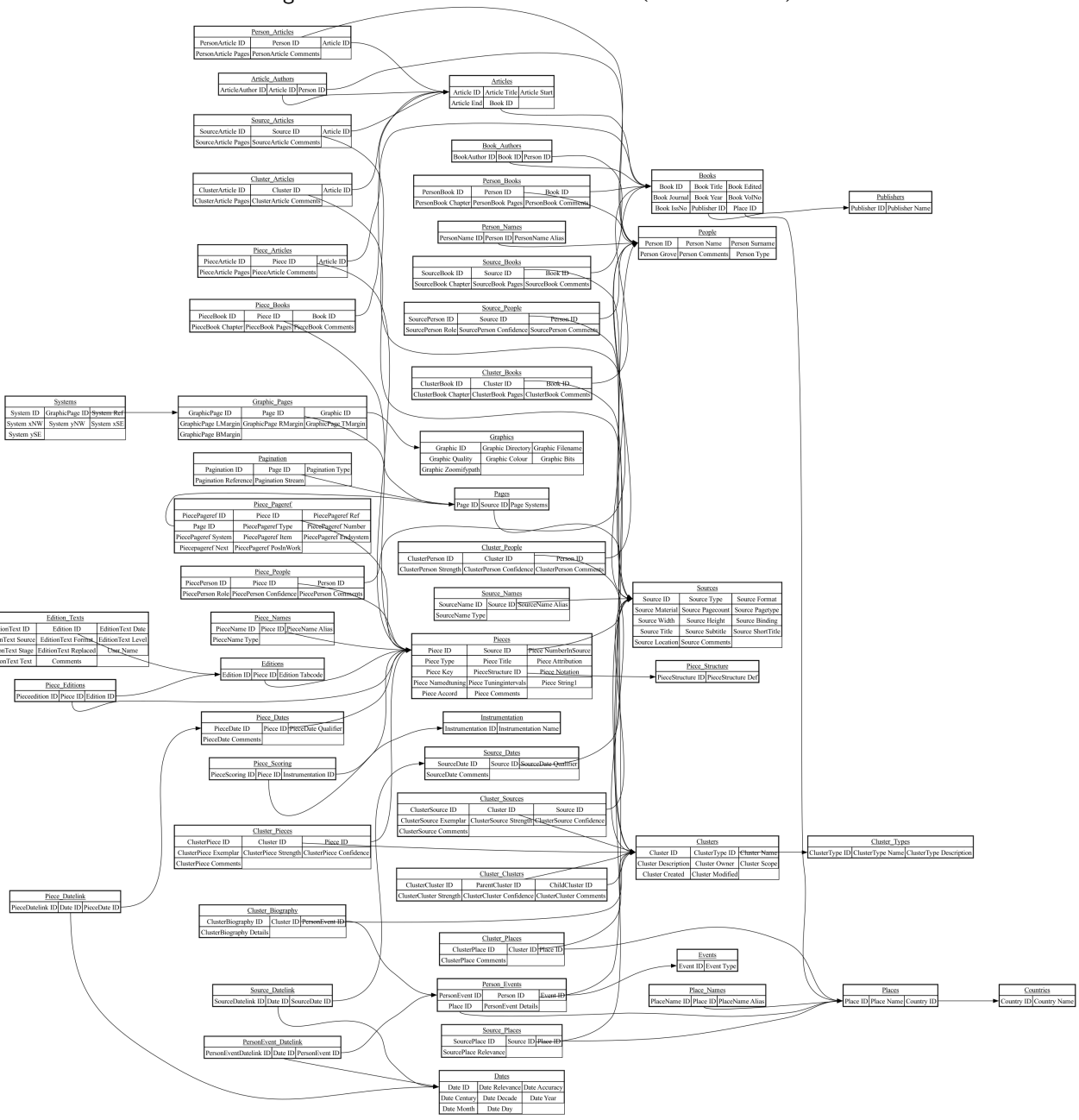
ECOLM I and II store all data directly in a relational database. The database contains both editorial tables, about users of ECOLM and their contributions, and record tables, about the works in the dataset. Figure 1 shows the record tables.

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.

Fig. 1: ECOLM II database schema (record tables)



6.2 RISM Muscat web application

Muscat[8] is a web application published by RISM for cataloguing musical sources, written using Ruby on Rails. Figure 2 shows the main tables.

Muscat uses a hybrid schema, in that each table has a single `marc_source` column containing an authoritative record in MARC21[9] 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 2) 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.

[8] https://rism.info/community/muscat.html
[9] https://www.loc.gov/marc/

Fig. 2: RISM Muscat schema summary