From Facsimile to Content Based Retrieval: the Electronic Corpus of Lute Music

Christoph Dalitz
Institut für Mustererkennung
Hochschule Niederrhein
Reinarzstr. 49, 47805 Krefeld

Tim Crawford
Goldsmiths College
University of London
New Cross, London, SE14 6NW

Abstract

In recent years many libraries have made their early music prints and manuscripts available online as 'digital facsimiles'. These are raster images that are not amenable to automatic searches for content, but require a human reader to browse them by eye. For lute music, an international team under the direction of Tim Crawford is currently building a new repository that stores the musical content in addition to the images, thereby opening the door to new levels of music research based on music information retrieval. This article describes the building of the corpus, its content and some of the possible fascinating use cases.

1 Introduction

Until recently, getting hold of a music source was both cumbersome and expensive: after the location of a copy had been identified from a bibliography or a reference book such as [1], the library would need to be contacted, and an order submitted through their reprographic service. After payment of a large sum of money and a delay of several weeks or even months, a microfilm or set of xerox copies would be sent by regular mail. While this is still the normal cause of action for many sources, especially in smaller or private collections, an increasing number of libraries now make scans of their content available on the Internet[1]. Prominent examples are

- the *British Library* ('Early Music Online')
- the *Bavarian State Library* (historic music prints)
- the *Saxon State Library* (music manuscripts from the baroque court of Dresden)
- the *Music Library of Sweden* (autographs of Johann Helmich Roman)

Once a large number of sources has been digitized, it is natural to wonder whether it is possible to find something within them. For this classic document search problem, an important distinction has to be made: a search can be based on *metadata* or on document *content*.

Metadata is the term for information – such as author, year of publication or category keywords – that has been manually entered into an electronic resource based on expert knowledge. Creating useful metadata requires considerable man-power, and, for practical reasons, the amount of metadata is thus typically limited to the data already present in a library catalogue before digitization, i.e. author, publisher, location, year, and verbatim title. Unfortunately, this is rarely sufficient even to find music for a particular instrument (e.g. for generating a list of all lute tablature prints in a library which have been digitized), not to speak of searching for a particular piece of music. A notable exception is the 'Early Music Online' (EMO) collection of the British Library (see Sec. 3, below): in this case the books were newly re-catalogued, so the full contents, giving all titles and composers of the individual pieces of music contained in the prints, are listed. This makes it possible to search for a particular piece of music, provided the words in the search query are spelt in the same way as in the catalogue metadata[2].

Content based music searches could be used to find documents matching a given melodic sequence, harmonic progression, or that are similar to some other document (e.g. different instrumental arrangements of the same vocal model, or divisions over the same

[1] For legal reasons, this is obviously only possible for content that is in the public domain, e.g. for historical sources.

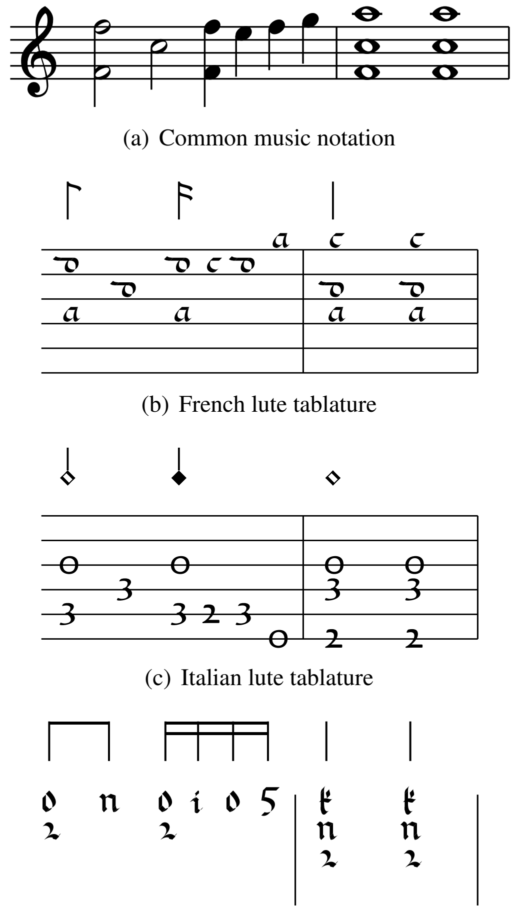
[2] The search in EMO is currently not tolerant to orthographic variations, but it would be possible to make it robust with the techniques described in [2].

This is a self-archived version of an article that appeared in
Phoibos - Zeitschrift für Zupfmusik 2/2013, pp. 167-185 (2013)

ground). Such searches are impossible in most existing ‘digital libraries’, because these store *raster images*, which are merely two-dimensional matrices of colored pixels. These images make musical sense in the eyes of a human viewer, but not to a computer. For real ‘musical’ searches, the content needs to be stored as a symbolic code that represents the musical meaning.

It is the goal of the ECOLM project (the Electronic Corpus of Lute Music) to build a database of such symbolic encodings for historic lute tablatures, which are a special subset of music sources[3]. As this corpus grows, it will offer increasingly useful opportunities for both practical musicians and scholars. The ECOLM project is now in its third stage: the two preceding phases laid the technical foundations, and the database is now assimilating content from EMO.

This paper is organized as follows: Sec. 2 gives an introduction to lute tablature notation and its symbolic representation as used by ECOLM. Sec. 3 describes the content from EMO that is digitized. Secs. 4 and 5 describe how the content is automatically recognized and how recognition errors are corrected in a ‘crowd correction’ step. Sec. 6 describes some existing and potential future retrieval possibilities, and the final section draws some conclusions for the future.



(a) Common music notation

(b) French lute tablature

(c) Italian lute tablature

(d) German lute tablature

Figure 1: The same music in different tablature notations.

2 Lute tablature

Tablature is a music notation specifically tailored for an instrument which gives physical playing directions rather than describing the contents of the music. In contrast to common music notation, lute tablature does not specify pitch and duration of individual tones, but indicates the strings (or, rather, ‘courses’, i.e., pairs of strings) to be played, the frets to be stopped, and the relative time between such actions. Nowadays, a form of tablature notation has become enormously popular for sharing rock guitar music on the internet[4]; until the advent of the internet, it was in wider use only for particular niche instruments such as flamenco and folk guitars or mountain dulcimer. In the 16th century however, lute tablature was the most important notation form for instrumental music: more than half of the extant printed instrumental music from the 16th century has come to us in tablature notation [1]. Some of these books are keyboard tablatures, but the vast majority are lute tablature prints and manuscripts.

Compared to mensural notation, lute tablature had the practical benefit of providing a compact reduction of several parts into a single staff, similar to a modern piano reduction of a complex orchestra score. As multi-voiced music was printed in separate part books and not in scores in the 16th century, lute tablature thus allowed for a compact notation of the polyphony. Of particular interest from today’s point of view is that about half of the extant lute repertory is based on vocal models. This allows us to draw conclusions about 16th-century performance practice because in lute tablature many embellishments and *musica ficta* alterations, missing but implicit in the mensural notation, are written out explicitly.

During the 16th century, at least three different nota-

[3] Roughly 8-10% of 16th-century printed music was for lute or related instruments

[4] A recent PhD thesis identified nearly 25,000 online tablature versions of Beatles songs alone. [3]

```
{<rules>
<notation>italian</notation>
</rules>}

Hd2a4
d3
Qd2a4
c2
d2
a1
|
Wc2d3a4
c2d3a4
|
```

Figure 2: TabCode for the tablature from Fig. 1(c).

tional systems were used for lute music, of which the most important are known today as 'Italian', 'French', and 'German' tablature based on their predominant regional use [4]. French tablature uses letters for the frets ('a' = zeroth fret, 'b' = first fret, etc.) with the top line representing the highest-pitched course, while Italian tablature uses numbers for the frets with the bottom line representing the highest-sounding course. German lute tablature on the contrary uses a staffless notational system with symbols that uniquely encode both course and fret (see [5] for details). The rhythm is indicated by 'flags' above the tablature system, which can optionally be beamed. Fig. 1 shows the same piece of music in all three tablature systems. The actual shape of the fret symbols and flags varied from print to print.

The ECOLM data is encoded in *TabCode*, an ASCII text format that was devised by Tim Crawford to be able to represent all graphical elements of the tablature[5]. TabCode does not employ different encodings for different types of tablature, but uses the 'French' convention of letters for the frets and encodes the tablature type in a metadata field before the actual TabCode. This has the effect that the same string/fret 'coordinates' result in the same TabCode which simplifies the technical implementation of content based queries.

Even though TabCode was designed to be compact, intuitive and readable enough for direct human input, an ECOLM end user should rarely come in contact with the actual TabCode, but will use a more intuitive, use-friendly graphical interface instead. It is nevertheless essential that the data is stored in an open, well-documented format, because this facilitates the sharing of data across all computer platforms. Moreover, it allows third parties to implement their own custom search algorithms without relying on proprietary software that can be tweaked only to a limited extent or that is not even available for a particular computer platform.

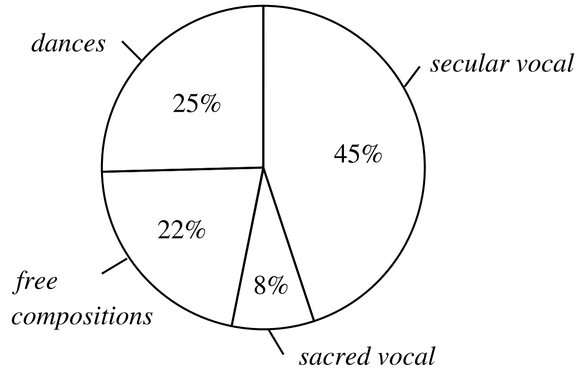


Figure 3: Genre distribution among the 1082 musical items from the EMO tablature prints.

3 EMO sources

'Early Music Online' (EMO) is a digital collection of images of 300 volumes of music printed before 1600 from holdings at the British Library. It was created in 2011 and is freely available online[6]. The project has focused on anthologies of printed music, because these provide a more realistic cross-section of 16th century music than prints devoted to individual composers and because items hidden in anthologies are more difficult to find in practice. As both titles and composer attributions were newly transcribed by expert cataloguers from the sources during the digitization and are now searchable as metadata in EMO, this provides a treasure trove for both practical musicians and scholars to track down sources or alternative versions of particular pieces of music. It should be noted however that composer attributions are often wrong or missing in the original prints and that spelling was far from normalized in any country in the 16th century: there are, for example, more than eight different spellings of the name of the composer 'Willaert' in the EMO sources, with this modern spelling notably absent.[7] The ortho-

[5] See `http://www.ecolm.org/` for a detailed specification of TabCode.

[6] See `http://www.earlymusiconline.org/`.

[7] The EMO metadata helpfully where possible also supplies standardized person-names, following normal library practice, us-

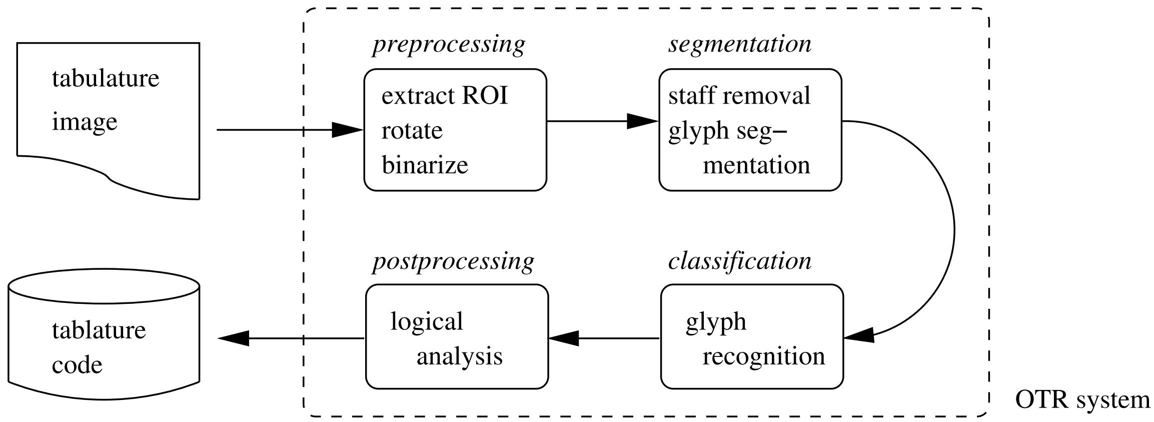


Figure 4: The processing steps of optical tablature recognition (OTR).

graphic variety for item titles is even greater.

Amongst the 300 volumes in EMO which contain about 10,000 musical compositions, there are 27 lute tablature prints containing in total 1082 musical items. As can be seen in Fig. 3, more than half of these items are transcriptions from vocal models. In some cases, the vocal models are also contained in EMO, which provides an interesting starting point for musicological studies, eg. with respect to embellishments or musica ficta. Finding concordances within EMO can be based on the EMO metadata in combination with the piece-level metadata from Brown's monumental work *Instrumental Music before 1600* [1], which provides titles and concordances for the lute music. By approximate string matching techniques [2], or geometric methods [6], candidates for concordances can be found for which the music needs to be compared in each case to verify actual concordances.

For studying known concordances with modern retrieval techniques and also for detecting concordances automatically, it is necessary to encode not only the tablature content of the lute music prints, but also the music from the vocal music prints in mensural notation. With the aid of 'Aruspix', a program developed by Laurent Pugin for the recognition of 16th century mensural notation [7], vocal music prints are also being recognized and encoded in the ECOLM project. Because of the musicological motivation behind its design, the cross-platform ASCII music-encoding XML format, MEI, is being used to store this data [8]. Since vocal music in mensural notation was printed in separate part books, it cannot be directly compared with a tablature version of the same piece. Some preprocessing is required, which could either be automatic 'intabulation' from the mensural parts, or automatic part extraction from the tablature. Concerning the latter problem, ECOLM researcher Reinier de Valk is currently working on different machine learning approaches for voice extraction from lute tablature [9].

4 Automatic recognition

The software for automatic recognition of the tablature content from the scanned images and its conversion to TabCode was developed under the direction of Christoph Dalitz and is freely available under an Open Source license[8]. It is based on the Gamera framework for document image analysis, a cross-platform Python library for building custom recognition systems [10]. Gamera is in worldwide use for the recognition of various kinds of non-standard document types like non-western language texts, medieval neumes, multiple choice tests, etc., and was thus especially well-suited for building a tablature recognition system [11].

Fig. 4 shows the individual steps of the recognition process. At the beginning of the current third stage of ECOLM, the software could already recognize French, Italian, and German lute tablature [11] [12]. Due to the peculiarities of the EMO sources, a number of extensions were necessary, however:

- in the *preprocessing* step, a region of interest (ROI) extraction became necessary.

ing the Library of Congress authority files from `http://id.loc.gov/authorities/names`

[8] See the section 'Addons' on the Gamera home page `http://gamera.sf.net/`.

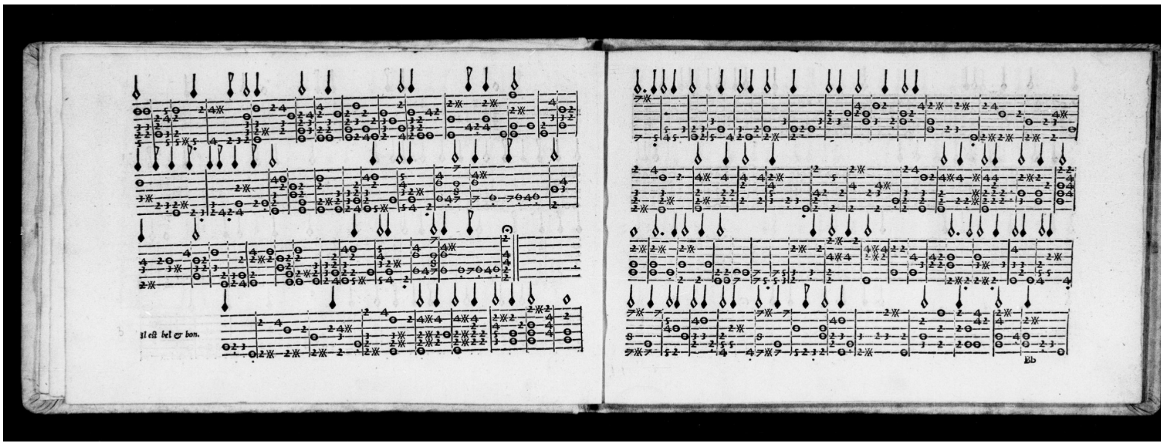


Figure 5: An example image from EMO showing the peculiarities of the image capturing process during digitization by the British Library; from Melchior Barberiis, *Intabolatura di Lauto ... Libro Decimo* (Venice: Scotto, 1549).

- the *glyph recognition* step had to be extended to deal with beamed flags, ledger lines for the bass strings ('diapasons'), and special signs indicating hold ('tenuto') fingerings or ornaments. Moreover, the existing barline recognition was not very good on the EMO sources and needed to be improved.
- the *postprocessing* step had to be modified to write TabCode, and to generate additional layout information like staff borders and glyph locations that can be utilized by the web interface of ECOLM.

In the following two subsections, we will describe the ROI extraction and the barline recognition, because these have been newly developed for ECOLM and have not yet been described elsewhere.

4.1 Region of interest extraction

Due to the specific digitization process in the British Library, the EMO images, which were derived from archival microfilms, have some peculiarities that can be seen in Fig. 5: each image shows a two-page opening with the individual pages in general having different skews, and there are black borders around the image and a shadow at the binding in the middle between the two pages. While some algorithms for ROI extraction are described in the research literature [13] [14], these are specifically designed for text documents. We have therefore devised a new algorithm for the EMO images that works as follows:

1) The image is converted to black and white with Otsu's threshold [15] and all black regions touching the image border are extracted.

2) All remaining connected components are extended at the top and bottom by some large value (e.g. one third of the image height), and thereafter overlapping segments are merged into single rectangles.

3) Of the resulting rectangles, the two largest represent the left and right side of the image. This splits the pages and automatically cuts out the shadow in the middle, but the rectangles still include the black borders at top and bottom.

4) Each rectangle is extracted from the original grey scale image and individually skew corrected with the projection profile method described in [16].

5) Inside the images containing only the black border for each side, the maximal empty rectangle is determined with the algorithm by Vandevoorde [17]. This rectangle is the resulting region of interest.

On 897 double sided tablature images from EMO, this algorithm resulted in only six errors, which is less than 0.6%: three double sides could not be split becausee the binding is too tight, and three half pages were too small because content touched the image border (i.e. the pages were trimmed too close by the binder), causing the largest empty rectangle inside the black border to become too small in the last processing step.

printer	number of barlines	new missed	new falsepos	new touching	old missed	old falsepos	old touching
Girolamo Scotto	1581	7.0%	0.0%	0.6%	14.4%	0.3%	16.2%
Scotto heirs	445	1.1%	0.2%	0.5%	0.9%	1.4%	11.0%
Antonio Gardane	2257	0.5%	0.0%	0.1%	5.1%	0.1%	1.6%
Gerhard Grevenbroich	379	0.0%	0.0%	0.0%	1.6%	0.0%	0.3%
Ricciardo Amadino	508	0.0%	0.0%	0.0%	8.3%	0.0%	0.0%
Pierre Phalese	1623	9.1%	0.0%	0.2%	23.3%	0.0%	7.8%
William Barley	611	7.4%	5.4%	0.3%	19.0%	2.6%	7.4%

Table 1: Comparison of the old, connected component based, with the new, runlength based, barline recognition on all different printers of Italian and French lute tablature in the EMO sources. *falsepos* are detected barlines that were no actual barlines, and *touching* are barlines that were joined with an adjacent symbol.

4.2 Barline recognition

As previously observed in [11], barlines are often so severely broken in historic lute tablature prints that they cannot be detected by classifying connected glyphs. The barline detection in the old recognition system before ECOLM III was done by a rule based grouping that looked for and joined barline fragments according to aspect ratio, width and total height of a group of barline fragments. On the EMO images, this approach was less than satisfactory, in particular because in some sources barlines were frequently connected with adjacent tablature letters, which had the effect that often neither the barline nor the letter touching it was recognized.

As runlength filtering[9] has proved useful for the detection of both horizontal and vertical lines in staffless German lute tablature [12], we deployed a similar idea for staff based tablature. The new bar detection algorithm assumes that the stafflines have been removed, e.g. with one of the methods from [18], and works as follows:

1) To find candidates for barline fragments, all horizontal black runs shorter than 2.5 times the staffline height[10] are extracted. The runlengths are segmented into connected components, and only those that intersect the staff regions and have a height/width ratio greater than 2.5 are kept.

2) The candidate fragments are sorted by staff system and x-position, and fragments varying by a small distance in the x-direction are grouped as barline candidates. A barline candidate is eventually accepted if at least one of its fragments is higher than 1.2 times the staffspace height[11] and the entire barline is higher than 2.5 times the staffspace height.

The first step has the side effect of separating barlines from symbols possibly touching them.

To evaluate the new barline recognition algorithm (method *TabPage.remove_barlines* in the Gamera OTR toolkit) and to compare it with the old rule based algorithm (method *TabGlyphList.classify_bars*), we have manually counted the errors of the two algorithms on ten pages of each French or Italian tablature print in EMO. The results, grouped by printer, are shown in Table 1. The total number of barlines varies not only due to different book formats and print density, but also because each printer was represented with a different number of books in EMO.

The numbers clearly show that the new runlength-based approach results in fewer missed bars and furthermore almost always separates barlines from touching symbols. A particular problem, however, occurs in two prints by Barley (1596). These use vertical lines not only for barlines, but also to indicate simultaneously plucked chords, thereby resulting in a high percentage of false positives. Even though this problem could be reduced by introducing additional rules or some context based classification, we did not investigate this further because Barley's prints are highly unusual in using woodcut rather than typesetting, and are therefore closer to engraved or manuscript tablatures.

[9] A *runlength* is the count of subsequent black image points until the next white point occurs.

[10] The *staffline height* can be measured quite reliably as the most frequent vertical black runlength in the image [18].

[11] The *staffspace height* can be measured quite reliably as the most frequent vertical white runlength in the image [18].

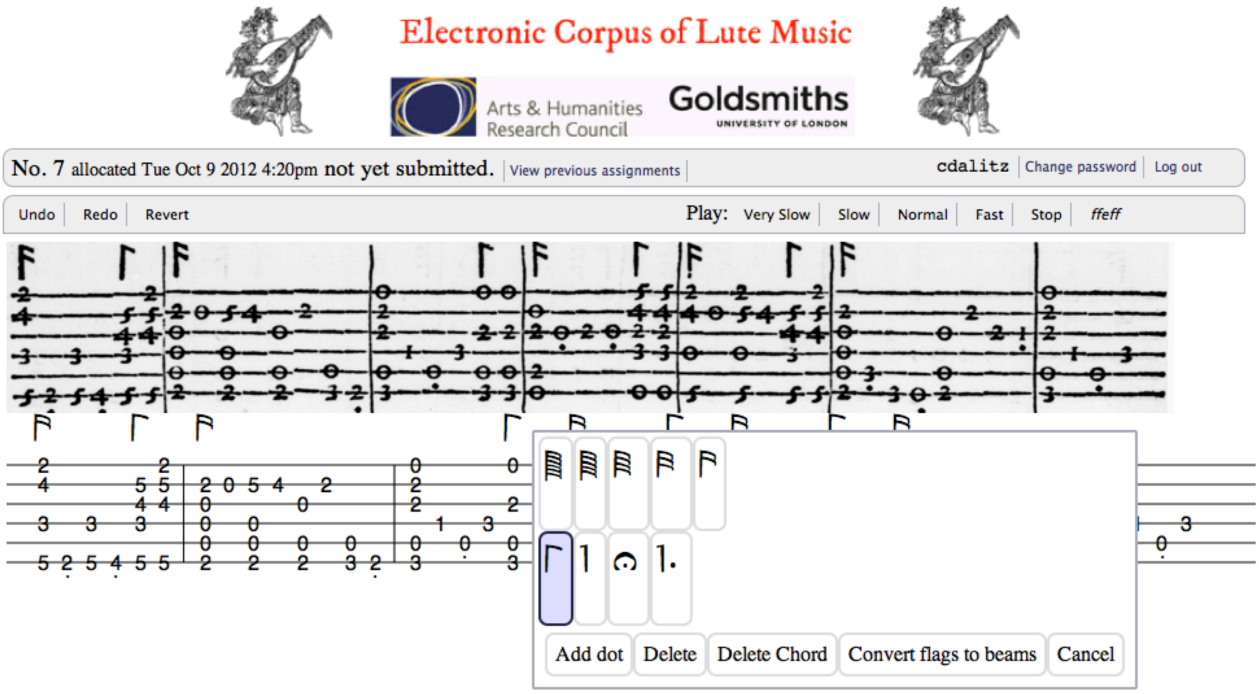


Figure 6: The ECOLM III webeditor for interactive correction of the automatically recognized tablature.

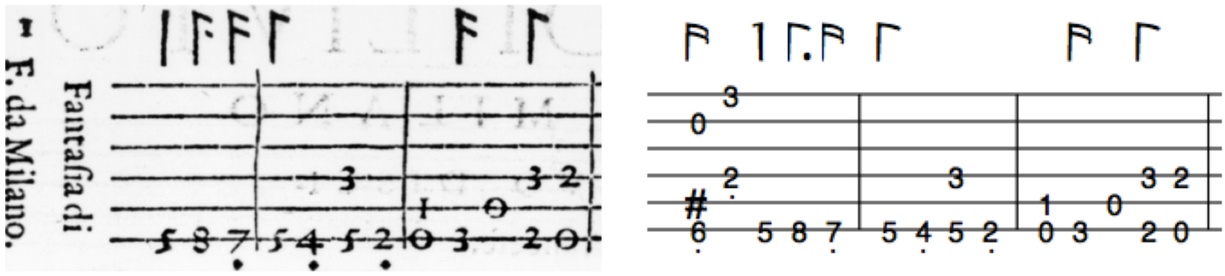
5 Crowd correction

Despite the low error rate of the automatic recognition software, there are still inevitable errors in the resulting code, e.g. due to show-through, bleed-through, broken or poorly-printed glyphs, overlap of titles with tablature, or due to tablature letters which are hard to distinguish, like ‘f’ and ‘l’. Correcting such errors requires experts in tablature, that is music scholars or practical lute players. As these are the people who benefit most from the outcomes of the ECOLM project, it was natural to ask them to volunteer to do the correction. While this might seem idealistic at first sight, projects like Wikipedia have shown that ‘crowd sourcing’ can be a successful approach that helps all affected parties. The call for volunteers at the British and German lute societies attracted over 50 participants within 3 months of the system going live.

As experts in tablature cannot be expected to be experts in computer science too, it is essential that the interactive user interface does not expose the end user to the complexities of TabCode. Moreover, it should run on a variety of computer platforms, which is most easily achieved with a web based interface. This has the additional benefit that the end user does not even need to install a particular software program, but can do the correction on any computer with a web browser and Internet access. The interface that was was developed by David Lewis under the direction of Tim Crawford can be seen in Fig. 6. It displays each tablature line in two forms: the original image and the automatically recognized tablature. In the latter, the user can click on errors and correct them through popup menus.

In the present phase of ECOLM, it is important that the correction does not introduce deviations from the sources by tacitly correcting apparent printing errors. The encoding stored in ECOLM is not meant to be a ‘practical performance edition’, but a faithful representation of the source. This does not preclude the addition at a later point of ‘creative corrections’ and critical comments, but for musicological studies based on the sources it is essential that the code is uncontaminated by well-meaning ‘editors’. To ensure this, each line of tablature is presented to two independent correctors and the entire proof-reading history of each encoding in the ECOLM database is stored as metadata. This has the side effect that even the uncorrected data can be used immediately for studies that do not need perfect reliability.

We have run our OTR system on a first set of fourteen complete volumes of 16th-century lute music from the EMO collection and started to put the output



(a) Original.

(b) As recognized by the OTR system.

Figure 7: Beginning of a fantasia by Francesco da Milano; from Francesco da Milano & Perino Fiorentino, *Intabolatura di Liuto ... Libro Terzo* (Venice: Gardane, 1562), fol. A1v. The OTR system tries to interpret the sideways-printed title as tablature, but otherwise produces a perfect result.

through the online error-correction process. Although correction was not complete at the time of writing, we have sufficient data to make a provisional assessment of both the accuracy of the recognition system itself and the effectiveness of the dual-correction process. The books for which we have dual-correction results are eight volumes of Italian lute tablature printed in Venice by Gardane (See [1], items: 1546/5, 1546/6, 1546/7, 1546/10, 1547/3, 1562/1, 1566/2 and 1566/3) and three of French tablature printed in Antwerp by Phalèse (1549/8, 1573/8 and 1574/7). While the figures presented below are by no means definitive, they are certainly encouraging. (It should also be stressed, however, that the test-set was selected on the basis that preliminary tests suggested a high likelihood of successful recognition.)

Of the 642 systems (lines) of tablature which have so far been corrected twice, 128 (about 20%) were judged to be ‘perfect’ by both correctors (that is, no corrections were made by either). A further 25 contained a single glyph error found by one of the correctors, and in another 55 a single glyph error was found by each corrector (we have not checked that this is the same error in both cases). So approximately 32% of systems (208) contained no more than a single incorrect glyph.

An initially puzzling finding was that the error-rate is consistently worse for the first system of each piece than for all of the others. It soon became clear that the OTR system was in fact attempting to read the textual titles for pieces as tablature where these are printed in alignment with the music systems, which they usually are. See Fig. 7. In future, we anticipate that this effect can be much reduced by a preprocessing phase of region selection similar to that described in Section 4.1, above.

The average error-rate (in terms of recognized glyphs) for the first systems in each piece (often containing titles, as shown in Fig. 7) was 2.39%; for systems other than the first, the error-rate drops to 1.60%, giving an overall error-rate of 1.69%. The worst-performing system (of those subjected so far to dual correction) showed an error-rate of about 18%; this suggests that even in this worst case, only about 40 glyphs need correction in an average system containing 223 glyphs.

As stated above, our initial test set was selected on the basis that it could perform well, but these figures are above our expectations, and suggest that, for a significant subset of the printed lute repertory, a fully-automatic process could quickly extract reasonably accurate musical encodings from microfilmed images. We intend to widen the scope of testing to include the products of more 16th-century printers of lute tablature, such as Scotto, the printer of the book shown in Fig. 5, though we foresee difficulties due to the lower quality of printing (poor registration and uneven spacing), and to physical problems such as poor paper quality, which can allow ‘bleed-through’ from characters printed on the reverse side of each page. A further crucial factor in any optical recognition is the quality of photography; in the case of EMO, however, the digitizations were done from the British Library’s archival films, taken at a time when standards were generally high in this respect.

6 Musical searching in ECOLM

The current ECOLM project web-site includes a ‘Query Builder’ interface which controls the assembly and initiation of SQL queries to the underlying MySQL database of ECOLM metadata. This includes a simple TabCode string-matching query system whereby any short tablature extract (with or without wildcards) may be exactly matched within the encoded music in the database. This can be made to work well, once the use of wildcards has been mastered, for locating exact occurrences of a certain pattern of tablature (a given chord, or an unaccompanied melodic sequence) in the database.

However, exact matching is of very limited value in this repertory for most purposes, owing to the possibility of an indefinite number of extraneous symbols which, though important in tablature terms, do not affect the essential 'musical' content (fingering and ornament markings, in particular). Furthermore, since matching is carried out on strings rather than a more musical structure, comments from the encoder or OTR system embedded within the TabCode can interrupt the musical flow. Using wildcards to make the search robust to such insertions risks permitting arbitrary musical insertions, reducing the reliability of search results. Such searches are also vulnerable to disruption caused by transposition or changes in barring.

A further problem is that normal string-matching techniques are only useful for monophonic music; if a single-line melody is sought in a piece containing a harmonised version (with a simple bass line, for example) it will not in general be matched. For this to work, a polyphonic matching method is needed.

If the relative tuning of the strings of the lute for which a tablature was intended is known, we can easily extract the chromatic pitch of each 'sounding' tablature glyph; the relative time-intervals between vertically aligned tablature 'chords' can also be determined easily from the rhythm signs above the staff (the absence of a rhythm sign implying the continued enforcement of the last time-value). In this way, we can build a pitch/onset-time representation of the music which we can use for MIDI playback or, indeed, for searching.

A family of geometric algorithms for finding matching patterns in a set of points in a multi-dimensional space, based on one called SIA, has been designed with music in mind. [6] These have several desirable features, including the ability to find transposed versions of a pattern as easily as those at the same pitch; they can even find relationships where notes are missing or altered. A problem with many melodic search tools is that they require the music to be separated into distinct voices – something that lute tablatures and piano music seldom provide. SIA, because it explores *all* relationships between notes, can easily locate melodies concealed within complex polyphonic textures.

We have implemented the SIA(M)ESE algorithm for searching the ECOLM database (which includes both the tablature and the tuning for over 1,000 pieces), using such a pitch/onset-time representation of the music. This implementation runs as a Common Lisp web service, responding to appropriately-framed HTTP requests.[12] Our implementation currently exists in test form and is not yet incorporated into the public ECOLM interface, but we believe that the advantages that it offers will make it an extremely useful component of the web resource, and a strong illustration of the need for symbolic, rather than purely graphical, resources.

This opens up the possibility of using tablature queries to search external databases of music as well as the reverse, that is, searching the ECOLM database for music from external resources. This is a field of great promise for musicology, and we are fortunate that the EMO collection of printed mensural notation (containing much of the same repertory as that in the lute books, much of which is arranged from vocal music) can largely be automatically encoded by a similar optical recognition method ([7]). Once this has been done, it will be possible to search across the lute and vocal music of EMO in order to discover hitherto undiscovered parallels between the two repertories for the first time.

As well as such work on matching and searching for patterns in symbolic representations of music a great deal of research activity has been devoted since about 2000 to the field of music information retrieval from audio recordings.[13] While this is largely motivated by commercial factors connected with the rapid expansion of online digital music resources in recent decades, the same techniques have great potential for musicological applications. We conducted some preliminary experiments on a collection of c. 80 audio CDs of lute (and related) music using a state-of-the-art audio search engine.[14] This requires that a 'feature-sequence' be extracted from the audio file, usually a sequence of multi-dimensional vectors of values for

[12]Our implementation of SIA(M)ESE benefitted greatly from the assistance of Jamie Forth and David Meredith.

[13]See: http://www.ismir.net

[14]www.omras2.org/audioDB

each position of a 'window' moved through the file by a known time-step. Searches are conducted by matching a query-sequence in the same format, usually constructed by extracting features from an audio file. However, depending on the nature of the feature used for matching (for many musical tasks the timbre-independent 'chromagram' feature is especially useful; see [19]), it is sometimes possible to construct suitable feature-sequences from symbolic, rather than audio, data, such as that in the pitch/time-onset internal tablature representation discussed above.

Preliminary experiments show that tablature-to-audio matching is indeed possible, although search results are not yet as good as those for audio-to-audio matching. This raises the further exciting possibility that tablature-originated searches could in future be conducted on mixed collections of symbolic (score) and recorded (audio) music, of which the latter is vastly more plentiful owing to the labour-intensive nature of score-encoding, even when aided by automatic methods as discussed in this article. Such research sets the scene for a bright future for computational musicology, which has suffered from the lack of suitable medium-to-large-scale resources on which to test its methods. It also suggests the possibility of moving from entries in a musicological corpus such as ECOLM to search for corresponding music in large resources such as YouTube and Spotify.

7 Conclusions

The 'Electronic Corpus of Lute Music' provides a valuable resource for lute players and musicologists as well as for computer scientists interested in music information retrieval. The error rates measured during the crowd correction step show that little manual correction of the automatic recognition is necessary. Building a large corpus is thus feasible in reasonable time. Even though the content based search facilities within ECOLM are currently very basic, they are already quite useful. The reader can try them out on the ECOLM website[15]. The new material from 'Early Music Online" discussed in this article will be incorporated shortly. Future enhancements of ECOLM will include the implementation and evaluation of different content based retrieval methods as described in Sec. 6 and the encoding of additional sources.

The next phase of development will focus on the provision of ECOLM's metadata and musical content as Linked Data, using the techniques of the Semantic Web to make the entire corpus, its methods and even the results of scholars' investigations using it available as a truly 'discoverable' resource[16]. Not only will this enable the easy linking of external resources (such as the British Library's, and other, online catalogues) to ECOLM, but it will allow ECOLM's data to be provided with a potentially limitless amount of contextual information (e.g. about people, places, musical instruments or historical events) from the 'infinite archive' of the Internet.

Acknowledgements

ECOLM was funded by the U.K. Arts and Humanities Research Council (AHRC). Parts of the tablature recognition software were written by Thomas Karsten, Christine Pranzas, and Christian Brandt. The ECOLM webeditor for the crowd correction was developed by David Lewis, to whom we are also grateful for valuable comments.

References

[1] H. M. Brown, *Instrumental Music Printed Before 1600: A Bibliography*. iUniverse, 2nd ed., 1999.

[2] G. Navarro, "A guided tour to approximate string matching," *ACM Computing Surveys*, vol. 33, pp. 31–88, 2001.

[3] R. Macrae and S. Dixon, "Guitar tab mining, analysis and ranking," in *ISMIR 2011*, pp. 453–458, 2011.

[4] D. A. Smith, *A History of the Lute from Antiquity to the Renaissance*. The Lute Society of America, 2002.

[5] D. Poulton, *A Tutor for the Renaissance Lute*. Schott Music Ltd., 1991.

[6] G. A. Wiggins, K. Lemström, and D. Meredith, "SIA(M)ESE: An algorithm for transposition invariant, polyphonic content-based music retrieval," in *ISMIR 2002*, pp. 283–284, 2002.

[15] `http://ecolm.org/` - click on 'Search the Database"

[16] http://linkeddata.org/

[7] L. Pugin, "Optical music recognition of early typographic prints using hidden Markov models," in *ISMIR 2006*, pp. 53–56, 2006.

[8] A. Hankinson, P. Roland, and I. Fujinaga, "The music encoding initiative as a document encoding framework," in *ISMIR 2011*, pp. 293–298, 2011.

[9] R. de Valk, *Towards Automatic Transcription of Sixteenth-Century Lute Intabulations: Designing an Algorithm for Automatic Voice Extraction*. Master Thesis, University of Utrecht (NL), 2008.

[10] M. Droettboom, K. MacMillan, and I. Fujinaga, "The Gamera framework for building custom recognition systems," in *Symposium on Document Image Understanding Technologies*, pp. 275–286, 2003.

[11] C. Dalitz and T. Karsten, "Using the Gamera framework for building a lute tablature recognition system," in *ISMIR 2005*, pp. 478–481, 2005.

[12] C. Dalitz and C. Pranzas, "German lute tablature recognition," in *ICDAR 2009*, pp. 371–375, 2009.

[13] D. Le, G. Thoma, and H. Wechsler, "Automated borders detection and adaptive segmentation for binary document images," in *ICPR 1996*, pp. 737–741, 1996.

[14] F. Shafait, J. van Beusekom, D. Keysers, and T. M. Breuel, "Document cleanup using page frame detection," *International Journal of Document Analysis and Recognition*, vol. 11, pp. 81–96, 2008.

[15] N. Otsu, "A threshold selection method from grey-level histograms," *IEEE Transactions os Systems, Man, and Cybernetics*, vol. 9, pp. 62–66, 1979.

[16] C. Dalitz, G. K. Michalakis, and C. Pranzas, "Optical recognition of psaltic byzantine chant notation," *International Journal of Document Analysis and Recognition*, vol. 11, pp. 143–158, 2008.

[17] D. Vandevoorde, "The maximal rectangle problem," *Dr. Dobb's Journal*, vol. 23, April, pp. 28, 30–32, 100, 1998.

[18] C. Dalitz, M. Droettboom, B. Pranzas, and I. Fujinaga, "A comparative study of staff removal algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[19] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 15–18, 2001.