The CollabScore project – From Optical Recognition to Multimodal Music Sources

Christophe Guillotel-Nothmann, Philippe Rigaux, Bertrand B. Coüasnon, Mathieu Giraud, Aurélie Lemaitre

The CollabScore project – From Optical Recognition to Multimodal Music Sources

Bertrand Coüasnon *Univ. Rennes, CNRS, IRISA, INSA Rennes* bertrand.couasnon@irisa.fr
Mathieu Giraud *CNRS/Univ. Lille* mathieu.giraud@univ-lille.fr
Christophe Guillotel Nothmann CNRS/Sorbonne Univ. christophe.guillotel-nothmann@cnrs.fr
Aurélie Lemaitre *Université Rennes 2, CNRS, IRISA* aurelie.lemaitre@irisa.fr
Philippe Rigaux Cnam, philippe.rigaux@lecnam.net

***Abstract*—We introduce COLLABSCORE, a project funded by the French National Research Agency, devoted to the design and production of tools and methods to improve accesses to large collections of sheet music scans. The new optical music recognition (OMR) approach developed in COLLABSCORE is part of a larger goal, namely that of interlinking multimodal documents related to music works. In this perspective, the music notation obtained from the OMR process is seen as a *pivot* that associates related fragments of images, audio, video, XML, or text sources. As an application of this principle, COLLABSCORE supports the *synchronization* of sources, leveraging the raw content of digital libraries with listening and visualization experiences. The present paper introduces the project and exposes some of its current achievements.**

I. OVERVIEW

The core concept of the project is that of *multimodal music sources* and the main project's efforts aim at creating tools and methods to *interlink* these sources. We begin with an overview of this perspective before surveying some more technical aspects.

A. Multimodal music sources

Given a music work (say, the Goldberg variations) seen as an abstract entity, we can find many concrete documents that provide a specific representation. These documents can be recordings, in audio or video format, images (scans) of score sheets, editable scores in MusicXML or MEI, and even textual sources that comment/annotate/enrich the music. It turns out that each representation is difficult to use beyond its specific purpose. For non specialists, we know it is hard to "hear" the music from a score and, conversely, it is hard to "replay" or analyse the music from a performance, live or recording. Moreover, sources are usually self-contained, independent documents, encoded in some specific format. This keeps from easily mapping music components (a voice, an harmonic sequence, a phrase) from one source to another, at a finer level of granularity than the whole document itself.

In COLLABSCORE, we address these issues with *multimodal music scores* (MMS). A MMS combines an encoding of the music notation (a MEI file) with links that associate the notation elements to the corresponding fragments of multimedia sources, e.g., a region on an image, a time frame in an audio/video source, as section of a textbook. Music notation is thus used as a description language for music content, which serves as a reference, or *pivot* to link heterogeneous sources that encode the same content.

COLLABSCORE implements this model in a data store[1] which provides (i) a management of such *pivot scores*, (ii) a storage of each pivot with external or internal multimedia sources, and (iii) an annotation mechanism that maps the pivot fragments to the corresponding part of each source [1], [2]. Figure 1 shows an example of a MMS: the *pivot score* (here, *La coccinelle*, a melody from Saint-Saëns) stored as a MEI document in Neuma is the central piece that glues together several sources: an image (taken from the Gallica digital library), a video accessible on YouTube, a MIDI file (internal source).

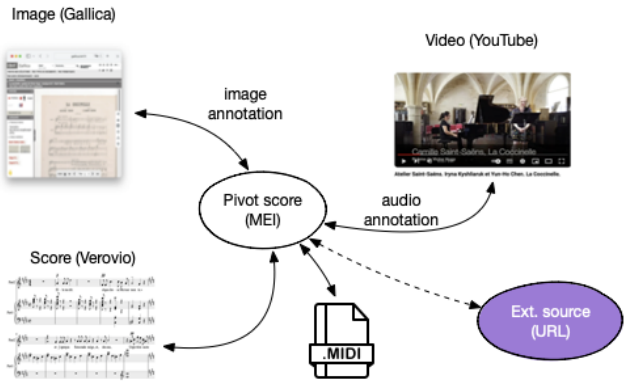

Fig. 1. A multimodal score and its sources

The project's work consists in designing tools to produce and manage MMS, including a powerful OMR system which the privileged mean to obtain a pivot. They are briefly summarized below.

B. Producing the pivot via optical recognition and crowdsourcing

Although pivot scores could be obtained by edition or transcription, COLLABSCORE integrates Optical Music Recognition (OMR) as the primary mean to produce a notation from image sources. In this context, our definition of "OMR" corresponds to the class of "structured encoding" OMR in [3]:

[1] http://neuma.huma-num.fr

we ambition to produce an editable score featuring all the notation elements visible on the sheet scan, along with their proper interpretation. In others words, we implement a process that attempts to invert the production of a printed score from specifications entered in a music notation engraver. Moreover, we combine this process with crowdsourcing phases to achieve a high-quality output, as discussed in [4]. The process is validated on a corpus mostly taken from the BnF Gallica Digital Library. These aspects are covered in Section II.

C. Alignment of sources

Multimedia sources are *aligned* with the pivot as shown on Fig. 2. The XML encoding of the notation (in MEI) identifies each component (here, a chord) with a unique id which is the target of annotations that refer to the corresponding fragments of sources. In the case of image, the annotation specifies a region on the image; in the case of audio/video, a time frame gives the start/end of the fragment.

Fig. 2. Aligning sources: A multimodal score with three documents

The alignment methods depends on the sources. In the case of images, annotations are supplied by the OMR system as a side effect of the recognition process. For other sources, dedicated interfaces have been implemented (Section III).

D. Applications

Finally, through the music description available in the pivot score, the content of two sources can be associated at a fine granularity level. The OMR output for instance can be controlled by a side by side display of both the source image and the pivot score rendering. Textual annotation (e.g., analytic comments) can be added on a score image at precise positions. An interface developed in CollabScore allows to listen an audio/video source while highlighting the music being played on the original image source. Among many other advantages, this is likely to greatly leverage the content of digital libraries with attractive features (details in Section III).

II. The OMR process

Among the various works on OMR [3], [5], two main types of approach can be observed in recent work. One is based on the detection of musical symbols [6], [7], inspired by architectures developed for object detection in natural scenes, with problems specific to OMR related to the large size of the images to be processed and the very small size of some musical symbols. The other is based on end-to-end recognition methods that directly produce a representation of the recognized score, which initially tackled monophonic scores and only very recently have been able to start to handle polyphonic systems [8], [9]. For the moment, these methods do not produce the localization of the recognized information required, for example, for image-sound synchronization.

The OMR process we propose in CollabScore to deal with polyphonic orchestra scores is founded on DMOS method, completed with a collaborative process that aims at clarifying the interpretation of symbols that have been identified as ambiguous. We experiment this combination of a large corpus for which a reference encoding has been produced.

A. Automatic syntactic OMR with DMOS

DMOS [10] relies a grammatical method that enables the combination of visual clues with syntactic rules, in order to describe both the physical and the logical content of the document. The process follows two steps, as shown in Fig. 3.

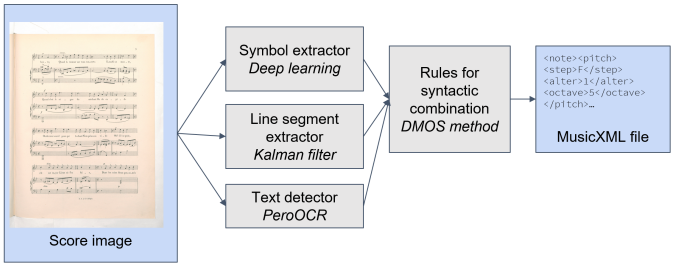

Fig. 3. Overview of DMOS: combination of low level detectors and high level syntactic rules

In a first step, three low level extractors are applied on the image:

- a symbol extractor based on deep learning (Cascade R-CNN - FocalNet architecture), dedicated to the extraction of small musical symbols [11] from high-resolution full-page images;
- an existing line segment extractor, based on Kalman filtering [12], used to extract linear elements, such as staff lines and stems;
- the existing PeroOCR [13], for the extraction of textual elements, such as titles, lyrics, instrument names.

In a second step, those elements are given as input to a syntactic system, based on DMOS method [10]. It produces a description of the graphical and syntactic content of the musical content of a score image: a score is made of staff systems, containing measures, and each measure contains musical objects (notes, rests, ...) that respect time constraints.

Recognizing a measure involves three steps of analysis. First, the staves and barlines are identified. Then, inside of a score, the graphical content is detected based on the position and assembly constraints of both the symbols detected by the deep object detector and the linear elements extractor: key, notes, rests, dots, accidentals, ties, slurs, dynamics, articulations marks, lyrics... Each detected content is localized in the image, and produced with is associated bounding box (Fig. 4).

Finally, the system organises the content into voices. After the distribution of notes into voices, the system checks the global consistency of the recognition, and produces warning if the detected elements do not follow some given rules. For example, if a eight note is miss-detected, the system will trigger a warning because the time signature is not respected.

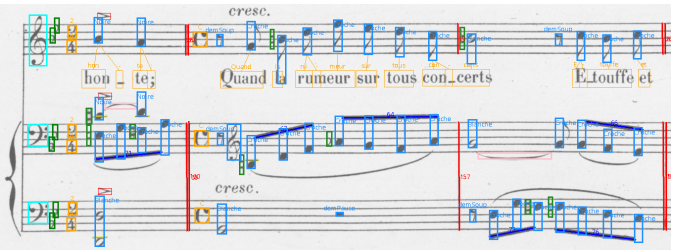

Fig. 4. The OMR process: Detection of graphical content

Moreover, based on the vertical alignment of notes in a system, it is possible to locate or even correct the note with the wrong duration. Applying these rules makes detection more reliable in a context of ancient noisy documents.

All the elements identified by DMOS are organized in a document compliant with the music notation grammar. This document is the main source of information to initiate a multimodal music source in CollabScore. Indeed, from the music notation symbols, a symbolic score in MEI is reconstructed – the pivot, and the source image is aligned with this pivot thanks to the regions identified by DMOS. Moreover, the alerts raised by DMOS are recorded and subsequently submitted to the collaborative process.

B. The collaborative process

The "raw" music score obtained from the OMR process enters in a phase of corrections via a sequence of dedicated interfaces. A design choice of CollabScore is to limit user actions to the list of alerts raised by the DMOS component. While this may seem restricted, we believe that going beyond would ultimately lead to implement a full online score editor[2].

The advantage of considering only the DMOS alerts is that we remain within the scope of an automatic recognition process, augmented with a one-time human assistance to solve difficult cases. This limits the competency expected from users, as well as the complexity of the required actions since they essentially consist in answering a question. This choice also provides a sound basis to evaluate the performance of DMOS: Given a ground truth, we can compare it first to the raw output, and second, to the corrected one, identifying the impact of human interaction on the final quality.

The list of alerts raised by DMOS are classified in three categories, based on how globally the potential error may impact the resulting score. These categories result in three correction phases:

- The first one, called *Instrumentation*, refers to the identification of music parts, and to the correct assignment of staves to the parts. Any error on these structural aspect has a dramatic impact on the whole score. This is the case for instance of a double-staff piano part not recognized as such, or when some parts are introduced/removed from one system to the other (e.g., a solo/melody arriving after an instrumental introduction, resulting in the introduction of a new staff in systems). Special cases difficult to identify automatically (e.g., transposing instrument) can also be solved during this step.
- The second one, *Transcription context*, refers to all the notation element that dictates the transcription of music events: clefs, key signatures and time signatures. Here again, any misinterpretation severely hinders the music notation accuracy.
- Finally, the last phase, *Music objects*, addresses the notation of musical events: notes, chords, rests, ties. At this point, the user cann *locally* correct a property of a faulty music object: duration, height, etc.

[2]Note that it always remain possible to import the MEI or MusicXML output in a standard score engraver

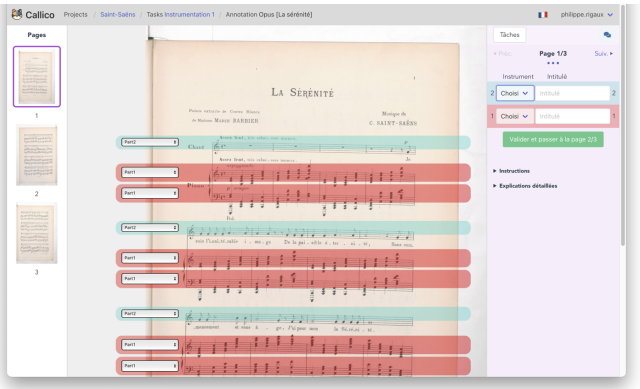

Fig. 5. The collaborative process, phase 1: checking parts and their staves

For each phase, a list of microtasks is produced, and submitted to a group of users. At the end of each phase, the list of validated corrections is applied to the score, and this corrected version is proposed to the following phase.

Fig 5 shows an example of the user interface dedicated to the first phase (Instrumentation). It heavily relies on information obtained from the DMOS analysis which comes as the default interpretation. Here, the list of parts (chant and piano) has been identified, and each staff (or pair of staves) assigned to a part. The user can correct this information if needed.

The subsequent phases imply a display of both the initial image and the score for comparison purpose (see Fig. 6 for phase 2). Elements to be controlled (here, clefs and signatures) can be highlighted on both the image and the target score, thanks to the regions provided by the OMR and to the links between both sources. We implemented an interface that lets the user directly correct an object (a clef, Fig. 6), each action being immediately reported on the score.

At the time of writing, we are finalizing the implementation of the collaborative system. It is based on the Open-source Callico system [14] and available at https://collabscore.cnam.fr. An experiment will be conducted in early 2025 with a group of users on a large corpus to be described next.

C. The reference corpus

The reference corpus comprises all the works by Camille Saint-Saëns (1835-1921) with the exception of dramatic works (operas, oratorios, incidental music). Aside from considerations relating to the BnF's promotion policy – CollabScore coincided with a project to promote the composer's work on

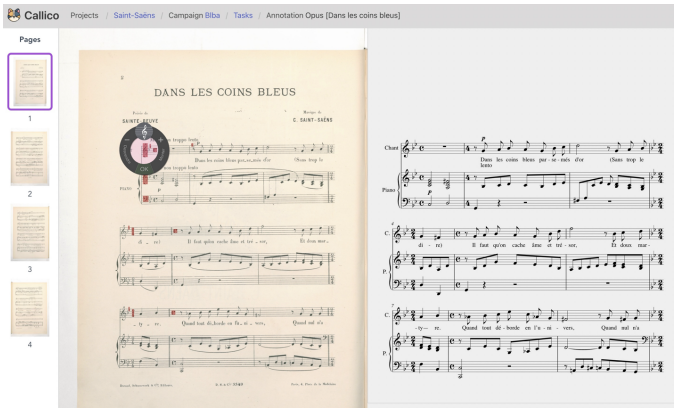

Fig. 6. The collaborative process, phase 2: checking the transcription context

the occasion of the hundredth anniversary of his death in 2021 – two criteria prevailed in the selection of this corpus, which totals more than 500 compositions.

1) **Variety of genres and instrumentation.** The compositions include sacred and secular works for a *capella* choir, chamber music, melodies for voice and piano, compositions for brass or military bands, keyboard repertoire and symphonic works with or without solo instruments. This diversity allows the software solution to be tested in different situations that present particular challenges, such as cross-staff notation in piano works, transposing instruments in orchestral works, or syllable positioning in melodies, etc.
2) **Particularities of French printed music from the period 1850-1920.** These scores, which have been made available by the BnF on Gallica, differ from modern, standardised notation, with regard to their implicit features (e.g. triplet notation), special signs (crochet rests), complexities relating to the placement of the text and the presence of artifacts in the preserved scores. Thus, we see this case study as an appropriate starting point for follow-up projects dedicated to printed music from earlier periods and handwritten notation.

For all the items, MEI files were created containing mei-headers with metadata extracted from Gallica including title, date of creation, genre, authorial attribution(s), historical print identifier, location and physical description. A sample of 18 scores was then transcribed in full, either manually or using commercial software (PhotoScore) with post-correction.

The reference corpus will serve as a ground truth to evaluate the performance of DMOS (for raw output) and of the collaborative phases (for users-corrected output). OMR evaluation is a notably difficult task [15]–[17] and we hope to contribute to progresses in this field. We started using the MusicDiff tool, designed by one of the project's partners [18] and now available as a Python package [3], but additional work is required with the OMR community to achieve a commonly accepted yardstick.

[3]https://github.com/gregchapman-dev/musicdiff

III. Sources alignment and synchronisation

Once obtained, the pivot score can be aligned with multimedia sources. We tailored the Dezrann platform [19] of our partner Algomus to propose tools for synchronization and synchronized score playback. Regarding images, as shown on Fig. 4, we can rely on the bounding box supplied by DMOS for each detected symbol, but also for all the measures, staves and systems. We link this region to the corresponding element ID in the pivot document.

Aligning with recordings (audio or video) involves identifying the time frame at the finest possible temporal granularity (we target the beat level). The fields of audio-score alignment and score following are actively researched [20]–[22]. Common methods involve dynamic time-warping algorithms or, more recently, deep learning approaches. In particular when sections are repeated. user interaction is often necessary to achieve a satisfying correspondence. We designed a simple interface to let users add and update alignment timestamps [19].

Finally, as a demonstration of the potential of our work to promote the content of digital libraries to a wide audience, CollabScore proposes an interface where the sources of a multimodal score can be displayed simultaneously for an improved user experience. Fig 7 shows how the original Gallica image, the pivot score and a YouTube recording can be associated, exhibiting at any moment a close correspondence between the performance, the notation, and the original image.

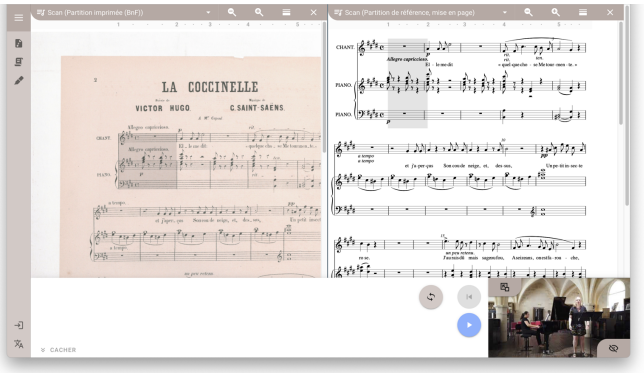

Fig. 7. CollabScore interface showing three synchronized sources on *La Coccinelle* with the Dezrann libraries: the original image, the pivot score, and a YouTube performance.

IV. Conclusion

The CollabScore project addresses many challenges in modeling and interlinking multimodal documents related to music, and has already required a lot of efforts to achieve its current state in OMR, collaborative process, score synchronization and playback. Each aspect would obviously deserve a much more detailed presentation and require further research and development, but we believe the the results obtained so far seem very promising. We are keen to showcase the CollabScore project with the community, and obtain in return an informed feedback.

[3]https://gallica.bnf.fr/ark:/12148/bpt6k1162049x

REFERENCES

[1] S. Cherfi, C. Guillotel, F. Hamdi, P. Rigaux, and N. Travers, "Ontology-Based Annotation of Music Scores," in *Intl. Conf. on Knowledge Capture (K-CAP'17)*, 2017, austin, Texas, Dec. 4-6 2017.

[2] R. Sanderson, P. Ciccarese, and B. Young, "Web annotation data model," Technical report, W3C Recommendation, 23 February, Tech. Rep., 2017.

[3] J. Calvo-Zaragoza, J. Hajic, and A. Pacha, "Understanding optical music recognition," *ACM Computing Surveys (CSUR)*, vol. 53, pp. 1 – 35, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:199543265

[4] C. Saitis, A. Hankinson, and I. Fujinaga, "Correcting large-scale OMR data with crowdsourcing," in *1st International Workshop on Digital Libraries for Musicology*. ACM, 2014, pp. 1–3.

[5] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. S. Marçal, C. Guedes, and J. S. Cardoso, "Optical music recognition: state-of-the-art and open issues," *International Journal of Multimedia Information Retrieval*, vol. 1, pp. 173–190, 2012.

[6] L. Tuggener, Y. P. Satyawan, A. Pacha, J. Schmidhuber, and T. Stadelmann, "The DeepScoresV2 dataset and benchmark for music object detection," in *Proc. ICPR*, 2021, pp. 9188–9195.

[7] Y. Zhang, Z. Huang, Y. Zhang, and K. Ren, "A detector for page-level handwritten music object recognition based on deep learning," *Neural Comput. Appl.*, 2023.

[8] J. Mayer, M. Straka, J. Hajič, and P. Pecina, "Practical end-to-end optical music recognition for pianoform music," in *Document Analysis and Recognition - ICDAR 2024*, E. H. Barney Smith, M. Liwicki, and L. Peng, Eds. Cham: Springer Nature Switzerland, 2024, pp. 55–73.

[9] A. Ríos-Vila, J. Calvo-Zaragoza, and T. Paquet, "Sheet music transformer: End-to-end optical music recognition beyond monophonic transcription," in *Document Analysis and Recognition - ICDAR 2024*, E. H. Barney Smith, M. Liwicki, and L. Peng, Eds. Cham: Springer Nature Switzerland, 2024, pp. 20–37.

[10] B. Coüasnon, "DMOS, a generic document recognition method: Application to table structure analysis in a general and in a specific way," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 8(2), pp. 111–122, 2006.

[11] A. Yesilkanat, Y. Soullard, B. Coüasnon, and N. Girard, "Full-page music symbols recognition: state-of-the-art deep models comparison for handwritten and printed music scores," in *DAS 2024 Workshop on Document Analysis System*, Sep. 2024.

[12] C. Queguiner, J. Camillerapp, and I. Leplumey, "Kalman Filter Contributions Towards Document Segmentation," in *ICDAR 1995 Third International Conference on Document Analysis and Recognition*, Montreal, Canada, Aug. 1995, pp. 765–769.

[13] O. Kodym and M. Hradis, "Page layout analysis system for unconstrained historic documents," *CoRR*, vol. abs/2102.11838, 2021.

[14] C. Kermorvant, E. Bardou, M. Blanco, and B. Abadie, "Callico: A versatile open-source document image annotation platform," in *Document Analysis and Recognition - ICDAR 2024: 18th International Conference, Athens, Greece, August 30 – September 4, 2024, Proceedings, Part III*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 338–353. [Online]. Available: https://doi.org/10.1007/978-3-031-70543-4_20

[15] D. Byrd and J. G. Simonsen, "Towards a standard testbed for optical music recognition: Definitions, metrics, and page images," *Journal of New Music Research*, vol. 44, no. 3, pp. 169–195, 2015.

[16] J. j. Hajič, "A case for intrinsic evaluation of optical music recognition," in *1st International Workshop on Reading Music Systems*, J. Calvo-Zaragoza, J. H. jr., and A. Pacha, Eds., Paris, France, 2018, pp. 15–16. [Online]. Available: https://sites.google.com/view/worms2018/proceedings

[17] P. Torras, S. Biswas, and A. Fornés, "A unified representation framework for the evaluation of Optical Music Recognition systems," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 27, no. 3, pp. 379–393, 2024.

[18] F. Foscarin, F. Jacquemard, and R. Fournier-S'niehotta, "A diff procedure for music score files," in *6th International Conference on Digital Libraries for Musicology*, 2019, pp. 58–64.

[19] L. Garczynski, M. Giraud, E. Leguy, and P. Rigaux, "Modeling and editing cross-modal synchronization on a label web canvas," 2022.

[20] M. Dorfer, F. Henkel, and G. Widmer, "Learning to listen, read, and follow: Score following as a reinforcement learning game," in *Proceeding of International Conference on Music Information Retrieval (ISMIR)*, 2018.

[21] J. Thickstun, J. Brennan, and H. Verma, "Rethinking evaluation methodology for audio-to-score alignment," *arXiv preprint arXiv:2009.14374*, 2020.

[22] M. Müller, Y. Özer, M. Krause, T. Prätzlich, and J. Driedger, "Sync Toolbox: A Python package for efficient, robust, and accurate music synchronization," *Journal of Open Source Software (JOSS)*, vol. 6, no. 64, pp. 3434:1–4, 2021.