Proceedings of the 4th International Workshop on Reading Music Systems

18th November, 2022

Organization

General Chairs

Jorge Calvo-Zaragoza	University of Alicante, Spain
Alexander Pacha	TU Wien, Austria
Elona Shatri	Queen Mary University of London, United Kingdom

Proceedings of the 4th International Workshop on Reading Music Systems, 2022

Edited by Jorge Calvo-Zaragoza, Alexander Pacha, and Elona Shatri

Preface

Dear colleagues!

We are very pleased to present to you the proceedings of the 4$^{\text{th}}$ International Workshop on Reading Music Systems (WoRMS). Following the success of last year's edition in a hybrid format, we decided to have this years edition as an online-only workshop to allow people around the world to easily participate while still being as interactive as possible. We hope that in the next couple of years we will be able to return to an in-person workshop, while maintaining the online option.

When we started the workshop series five years ago we did not now how it would be perceived by the community. Therefore, we are very happy that WoRMS has established a fixed place in the community and is seeing great interest from people all around the world that share a common interest in music reading systems, allowing them to exchange ideas and form relationships with one another.

We would also like to use the opportunity here to mention and promote our public YouTube channel `https://www.youtube.com/OpticalMusicRecognition`, which has recordings for last year's sessions and we plan on adding this year's presentations as well. If you have interesting content that you want to share through this channel, please get in touch with us.

This year's edition features 9 contributions, reaching from topics like dataset generation, via new attempts to tackle music notation assembly to measure detection and drum transcription. We are looking forward to very interesting discussions. We also want to thank the TU Wien for providing Zoom conferencing facilities.

Jorge Calvo-Zaragoza, Alexander Pacha, and Elona Shatri

Contents

Challenging sources: a new dataset for OMR of diverse 19th-century music theory examples

Fabian C. Moss
Institut für Musikforschung
Julius-Maximilians-Universität Würzburg
Würzburg, Germany
fabian.moss@uni-wuerzburg.de

Néstor Nápoles López[§]
Distributed Digital Music Archives and Libraries Lab
McGill University
Montreal, Canada
nestor.napoleslopez@mail.mcgill.ca

Maik Köster[§]
Musikwissenschaftliches Institut
Universität zu Köln
Köln, Germany
mkoest14@uni-koeln.de

David Rizo
Department of Software and Computing Systems
Universidad de Alicante. ISEA.CV
Alicante, Spain
drizo@dlsi.ua.es

***Abstract*—A major limitation of current Optical Music Recognition (OMR) systems is that their performance strongly depends on the variability in the input images. What for human readers seems almost trivial—e.g., reading music in a range of different font types in different contexts—can drastically reduce the output quality of OMR models. This paper introduces the 19MT-OMR corpus that can be used to test OMR models on a diverse set of sources. We illustrate this challenge by discussing several examples from this dataset.**

***Index Terms*—optical music recognition, historical sources, diversity, music theory, digital humanities**

I. Introduction

While Optical Music Recognition (OMR) techniques have advanced in recent years, several challenges remain, and the current state-of-the-art does not always provide satisfactory solutions [1]–[5]. In this report we want to draw the attention of the OMR community to one particular issue, namely the case that the sources themselves are challenging by their inherent diversity. The datasets to which OMR models are applied or on which they are trained are frequently *homogeneous* in the sense that they stem from a single source or collection of more or less uniform sources. This often entails that the images are stylistically similar [7] and that the presence of text is largely limited to lyrics or annotations.

If it is the goal of OMR to perform (at least) at human-level music transcription, it must be able to deal with scores printed with different font types, be capable of understanding which parts of a page contain music and which do not, and distinguish between raw text, lyrics and other textual information such as chords or harmonic analysis indications. This is particularly relevant for research applications with a historical focus, where font types may be less standardized.

In this report, we introduce the 19MT-OMR corpus, a multilayered dataset of heterogeneous and multimodal data that may aid researchers in progressing towards this goal, and illustrate the failure of current OMR approaches with a handful of salient examples.

II. Description of the data

The 19MT-OMR corpus was created within the context of the digital-humanities project "Digitizing the dualism debate: a case study in the computational analysis of historical music theory sources" [8]. It consists of scans and transcriptions of 19th-century German music theory textbooks in TEI and MEI formats (see Fig. 1). Common to all books is their music-theoretical content and *not* their typeface and graphical layout. We provide the corpus in three versions of increasing specificity with respect to OMR:

- 19MT-OMR-A: complete segmentations and transcriptions of all sources in the corpus
- 19MT-OMR-B: only pages containing music examples
- 19MT-OMR-C: only the music examples

The dataset is hosted within the *Open Science Framework*[1] [9] under a CC-By Attribution 4.0 International license. To create the corpus, publicly available scans of the books were segmented and transcribed using the *Transkribus* software [10] for segmentation and Optical Character Recognition (OCR), using the ONB_Newseye_GT_M1+ model and trainDataLanguageModel dictionary [11].

During segment markup and transcription, we made several editorial decisions. For instance, some sources contained purely rhythmic notation [12], which we decided to ignore. As a short-hand rule, only music examples with five lines

This research was supported by the *Collaborative Research on Science and Society* (CROSS) program of École Polytechnique Fédérale de Lausanne (EPFL) and Université de Lausanne (UNIL) for the project "Digitizing the Dualism Debate: a case study in the computational analysis of historical music theory sources".

[§]Equal contribution.

[1]https://osf.io/qm9z5/

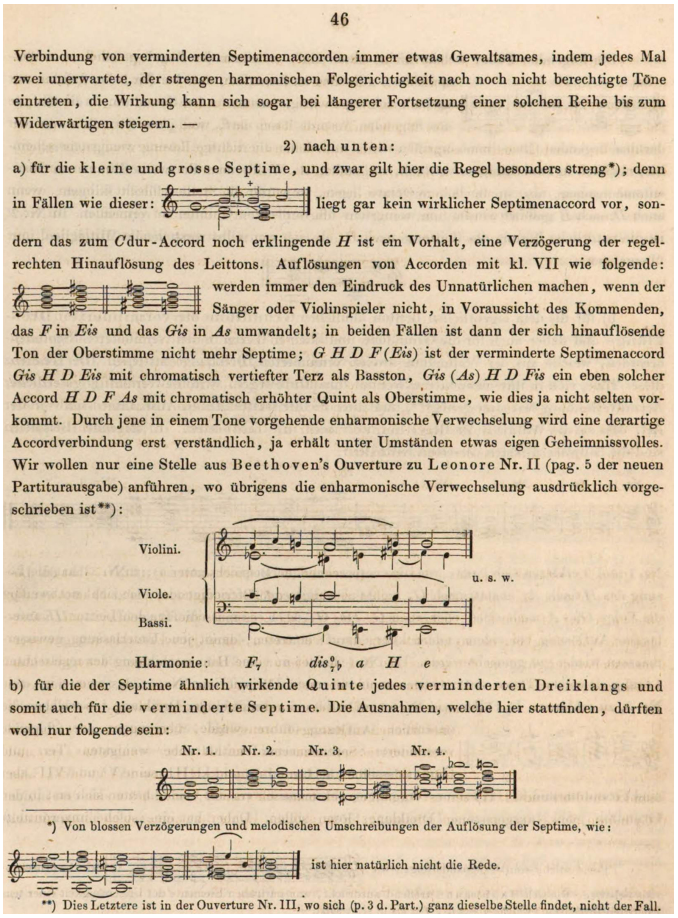


Fig. 1: Crop of page extracted from the corpus.

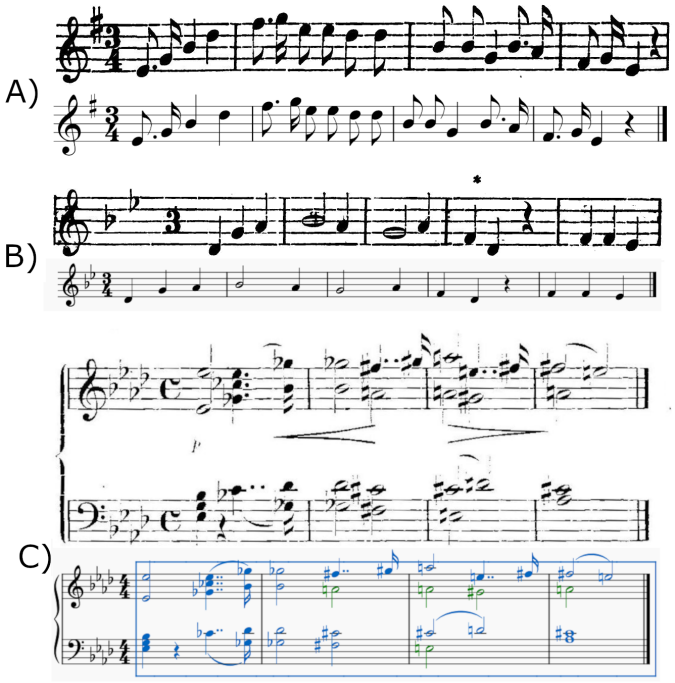


Fig. 2: Particularly challenging cases. Example A) shows an encoding where the original source is missing an augmentation dot (m. 4, first eighth note), which is added to the digital encoding. Example B) shows an unconventional form of the time signature $\frac{3}{4}$ in the original source. Example C) shows a case with more than one voice. In the latter example, the note colors in the encoding highlight the notes that are encoded as the first (blue) and second (green) voices.

were encoded. However, if they were embedded in complex diagrams or annotated with lines and arrows, we likewise excluded them from our transcription. Since the examples were taken from segmented page elements, the dataset excludes the small number of in-line music examples.

The music examples are encoded in the Humdrum (**kern) symbolic music representation The choice of this format is due to its simplicity, unambiguity, and its capability to encode all the desired features of the notation. Thanks to the *Verovio Humdrum Viewer* tool,[2] engraving was fairly simple. Additionally, we have described each image with some features that will be relevant for the OMR process and provide them in an accompanying CSV file (see Section IV for a detailed description).

The transcribed data spans 9 textbooks and a total of 368 musical excerpts. Most of the examples contain between 1 and 32 measures of music. Some examples include single-line-staff rhythmic patterns, however, most examples feature multiple voices per staff with unconventional engravings as described below in Section IV. Musically, many of them feature basic scales and cadences, while others contain short excerpts from known compositions. In all the examples, the direction of stems, number of voices, time signatures, slurs, ties, and accidentals have been meticulously encoded. Other features, such as musical fonts or some special symbols were not taken into consideration. The text in the examples was initially encoded separately from the music notation. Later, in order to have a TEI file that describes both the text and the music, **kern files will be converted into MEI to be linked from the TEI files by using the `<music>` TEI element [13].

III. Challenging examples and encoding decisions

As a corpus of music theory books illustrating specific situations, the dataset contains several unusual musical notations that are difficult to encode. Figure 2 shows three examples from the corpus. The first example, Fig. 2-A, shows a situation where the encoding requires the inclusion of a symbol (augmentation dot) in the last bar that is missing in the source image. This is rare, but it occurs throughout the corpus for some symbols, such as, augmentation dots and triplets. The second example, Fig. 2-B, shows an unconventional notation of the time signature in the source image. Other unconventional notations found in the 19MT-OMR corpus include: music fonts that are hard to read, double augmentation dots that are widely separated from their corresponding note, whole notes located at the center of a measure instead of at

[2]https://verovio.humdrum.org/

the beginning, etc. The third example, Fig. 2-C, showcases the strategy followed for encoding multiple voices. To ensure consistent encoding, a single voice was used whenever possible. However, when two notes have differing stem directions and/or note durations, an additional voice was encoded. The order of the voices was always encoded from top to bottom. That is, the first voice is always the upper voice. Additional (lower) voices were encoded below as needed. The examples in the corpus span up to four voices in one staff.

IV. OMR CHALLENGES

In order to understand the possible difficulties that OMR systems may encounter when approaching the corpus, we have created a set of features to qualitatively describe the source images that may also be used to filter subsets of our corpus (see Section II).

The first feature is the *polyphonic or monodic nature* of the individual staves in the image. For instance, the image in Fig. 3a is not tagged as polyphonic because both staves are monodic. As mentioned above, some music examples contain purely *rhythmic notation*, see Fig. 3b.

One of the most challenging features of the dataset is the presence of *harmonic indications* that are difficult for an OMR system to distinguish from lyrics or other text annotations (see Fig. 3a (figures bass), 3c) (reference labels), and 3d (groupings).

The corpus contains different *layouts*, from single staves containing just one voice such as in Fig. 2-A and -B, grand staff examples containing multiple voices (Fig. 2-C), as well as small ensemble scores or several staves that should be read aligned (Fig. 3a). Consequently, those images that contain more than one staff are tagged using either the "grand staff" or "several staves" feature. A feature has also been created for those images containing several *systems*, such as Fig. 3d.

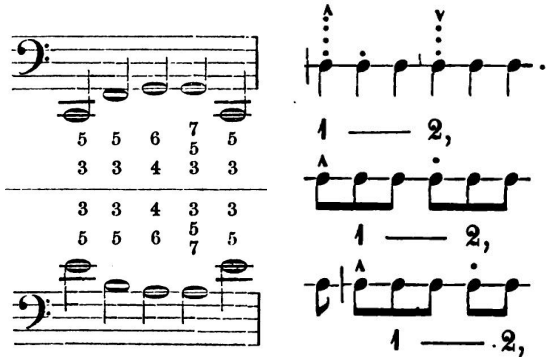
As the musical excerpts are meant to illustrate the content of the treatises, a single image may in many cases contain several separate examples (Fig. 3e), or have numbers or section labels naming the different examples in the image (see staff above footnote in page shown in Fig. 1, or Fig. 3c).

The corpus contains a range of symbols and engravings for which there are no standard encodings, such as dots indicating metrical strength (top staff in Figure 3b), braces around chords (Fig. 3d), duplicate note heads, bar lines broken by slurs, movable types (Fig. 3h), or whole note horizontally displaced (Fig. 3h, fourth measure), to name just a few.

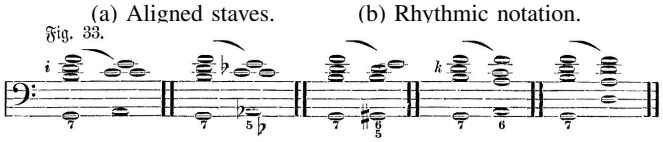
Finally, the corpus contains several cases of unusual fonts or engravings such as elliptical note heads (Figs. 2-B and 3), beams, slurs and stems made of movable types that are visible (Fig. 3f), overlapping notes (see last measure in Fig. 3h), and notes that seem to be printed over other contents (Fig. 3e).

V. EVALUATION

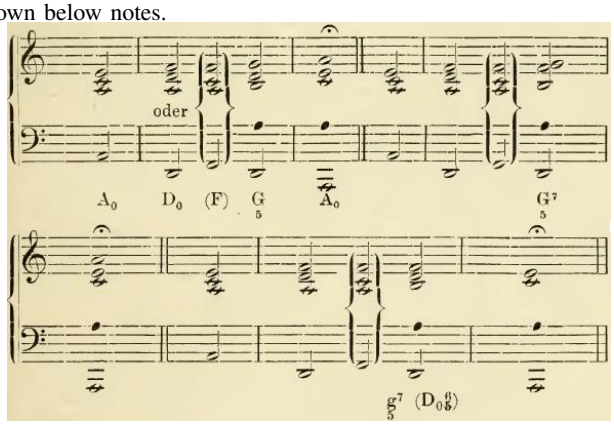
Having described the type of content to be handled by OMR, it was *a priori* expected that no current system approach would be able to correctly handle the entire corpus. No rule-based system is built with the nature of this corpus in mind. Given the



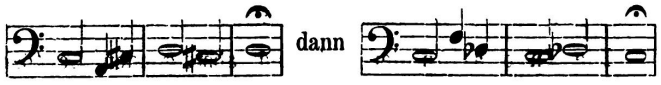
(a) Aligned staves. (b) Rhythmic notation.



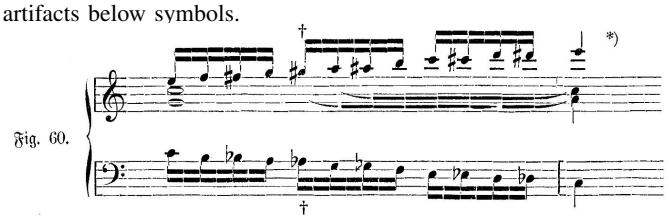
(c) Two examples, *i* on the first bar and *k* on the third one. Inversions shown below notes.



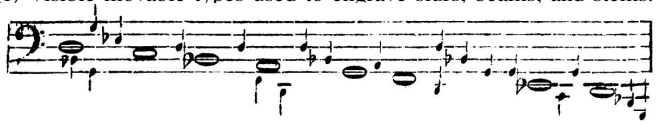
(d) Braces enclosing chords. See also black note heads.



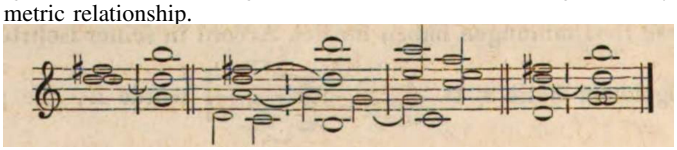
(e) Images containing two separated examples. Note the messy artifacts below symbols.



(f) Visible movable types used to engrave slurs, beams, and stems.



(g) Cue-sized notes along with whole notes not following at all any metric relationship.



(h) Bar lines cut by slurs. Double note heads and overlapped ones (see last measure).

Fig. 3: Examples of special cases found in the corpus.

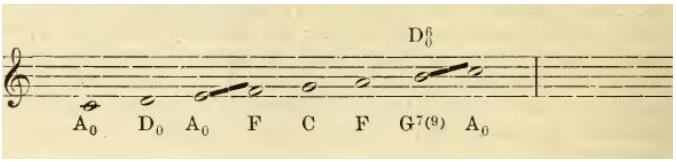
small size of the dataset, on the other hand, machine learning algorithms do not have sufficient samples to build a model capable of recognizing the diversity of situations found.

To qualitatively evaluate this 'bad-performance' hypothesis for current OMR systems with the 19MT-OMR corpus, we have chosen some salient examples that we consider representative of the different characteristics introduced above. We used trial or demo versions of the most popular commercial systems PhotoScore by Neuratron,[3] SmartScore by Musitek,[4] Maestria by Newzik that claims to use state-of-the-art neural-network techniques,[5] Capella-Scan, Version 9.0-10,[6], FORTE 12 Premium Edition that contains ScanScore 2 Ensemble,[7] and finally the open-source system Audiveris.[8]

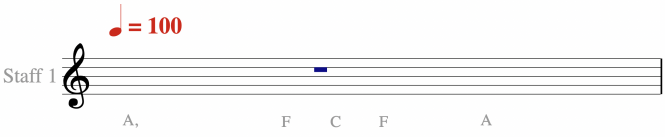
Each of the systems has different requirements for the input images in terms of file formats and resolutions that have been met in all cases by using an image processing program, in some cases by resampling the source images. As mentioned in the introduction, most OMR systems have been designed with a different type of repertoire in mind, with content and image size requirements that are, in most cases, far from 19MT-OMR. This may be the main reason why the recognition accuracy is extremely low in all tested systems.

This is illustrated in Fig. 4. A score example that is easily recognized by a human reader (Fig. 4a) comes with a single symbol that can confuse the OMR: the portamento-like lines showing where the semitones are. The remaining sub-figures show the output of the used OMR software (the output from Newzik is missing because it just generated an empty MusicXML file). In addition to the 'garbage' obtained regarding the scale in the original, none of the approaches is able to recognize the harmonic annotation as such.

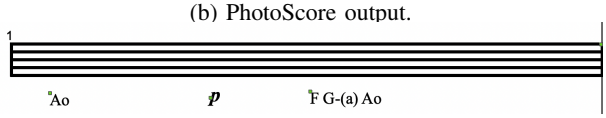
The results for the examples in Fig.3 are similar. For Fig. 3f, PhotoScore does not detect anything but a list of whole rests, warning us in a dialog box that the image contains a high number of reading errors. SmartScore fails in the operation and Newzik just generates an empty MusicXML file. Audiveris detects some notes and strange symbols, and both Capella-Scan and Forte 12 just output noise. For Fig. 3e the situation is not better. Photoscore and Newzil identify one example with many errors. Smartscore crashes and Forte 12 does not export anything. Both Audiveris and Capella-Scan correctly detect two separated regions, but these contain just clusters of unrelated symbols. The output for all other examples is similarly problematic.



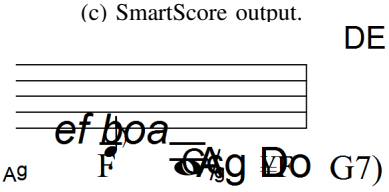
(a) See lines between notes E–F and B–C (semitones).



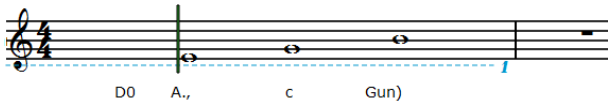
(b) PhotoScore output.



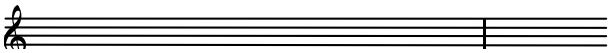
(c) SmartScore output.



(d) Capella-Scan output.



(e) FORTE 12 Premium Edition output.



(f) Audiveris output.

Fig. 4: Music recognized by different systems from a single image.

VI. Discussion

Most efforts of the OMR community are now focused on whole scores containing predominantly music notation with only few text elements. In this work, we have introduced a different kind of purpose that likewise needs the improvement to render OMR technologies that we illustrated by introducing the 19MT-OMR corpus and discussing a number of salient examples of where current OMR approaches and encoding formats fail. We believe that progress is particularly needed in order to exploit the full potential of OMR techniques also for traditional musicologists and digital humanists.

We consider 19MT-OMR and the discussed examples to be a valuable contribution to OMR research. What may look like odd idiosyncrasies from a modern engraving perspective were actually common historical engraving techniques. Elliptical white note heads, for instance, also feature in several important eighteenth century treatises [14]. Recognizing an object, such as a half note, across various historical printing conventions is a significant long-term challenge, comparable to what has recently been achieved in OCR models handling black letter as well as Roman fonts. Finally, the correct identification of different harmonic annotations is a pending feature that should be addressed in the future by the music encoding and OMR communities.

[3]https://www.neuratron.com/photoscore.htm (Accessed Sept. 20, 2022).
[4]https://www.musitek.com/ (Accessed Sept. 20, 2022).
[5]https://newzik.com/en/maestria/ (Accessed Sept. 20, 2022).
[6]https://www.capella-software.com (Accessed Sept. 20, 2022).
[7]https://www.fortenotation.com (Accessed Sept. 20, 2022).
[8]https://audiveris.github.io/audiveris/ (Accessed Sept. 20, 2022).

REFERENCES

[1] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. S. Marcal, C. Guedes, and J. S. Cardoso, "Optical music recognition: state-of-the-art and open issues," International Journal of Multimedia Information Retrieval, vol. 1, no. 3, pp. 173–190, Oct. 2012, doi: 10.1007/s13735-012-0004-6.

[2] D. Byrd, 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, Jan. 2015.

[3] J. Calvo-Zaragoza, J. H. Jr., and A. Pacha, "Understanding Optical Music Recognition," ACM Comput. Surv., vol. 53, no. 4, p. 77:1-77:35, Jul. 2020, doi: 10.1145/3397499.

[4] E. Shatri and G. Fazekas, "Optical Music Recognition: State of the Art and Major Challenges," arXiv:2006.07885 [cs, eess], Jun. 2020, Accessed: Jun. 21, 2020. [Online]. Available: http://arxiv.org/abs/2006.07885

[5] J. deGroot-Maggetti, T. R. de Reuse, L. Feisthauer, S. Howes, Y. Ju, S. Kokubu, S. Margot, N. Nápoles López, F. Upham "Data Quality Matters: Iterative Corrections on a Corpus of Mendelssohn String Quartets and Implications for MIR Analysis" in International Society for Music Information Retrieval Conference (ISMIR 2020), 2020, Montréal, Canada, pp. 432–438.

[6] A. Ríos-Vila, J. Calvo-Zaragoza, and D. Rizo, "Evaluating Simultaneous Recognition and Encoding for Optical Music Recognition," in 7th International Conference on Digital Libraries for Musicology, New York, NY, USA, Oct. 2020, pp. 10–17. doi: 10.1145/3424911.3425512.

[7] J. Calvo-Zaragoza, , D. Rizo, "End-to-End Neural Optical Music Recognition of Monophonic Scores", Applied Sciences, vol. 8, no. 4, doi: 10.3390/app8040606, 2018.

[8] F. C. Moss, M. Köster, M. Femminis, C. Métrailler, and F. Bavaud, "Digitizing a 19th-century music theory debate for computational analysis," in CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands, 2021, pp. 159-170.

[9] E. D. Foster and A. Deardorff, "Open Science Framework (OSF)," J Med Libr Assoc, vol. 105, no. 2, pp. 203–206, Apr. 2017, doi: 10.5195/jmla.2017.88.

[10] G. Muehlberger et al., "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study," Journal of Documentation, vol. 75, no. 5, pp. 954–976, Jan. 2019, doi: 10.1108/JD-07-2018-0114.

[11] Antoine Doucet, "NewsEye: A digital investigator for historical newspapers", presented at the Digital Humanities 2020 (DH 2020), Ottawa, Canada, Jul. 2020. doi: 10.5281/zenodo.3895269.

[12] M. Hauptmann, Die Natur der Harmonik und der Metrik. Leipzig: Breitkopf und Härtel, 1853.

[13] G. di Bacco, D. Ried. "A very brief introduction to MEI - the Music Encoding Initiative. And case studies dealing with mixed verbal-musical content for TEI-Publisher", presented at the e-editiones online event "Music is in the air – MEI and TEI Publisher", Jul. 8th, 2020. https://e-editiones.org/music-is-in-the-air/.

[14] H. C. Koch, Versuch einer Anleitung zur Composition, Leipzig: A. F. Böhme, 1782.

[15] D. Rizo, J. Calvo-Zaragoza, J.M. Iñesta, "MuRET: a music recognition, encoding, and transcription tool", Proceedings of the 5th International Conference on Digital Libraries for Musicology, September 2018, pp 52-56. doi: doi.org/10.1145/3273024.3273029