The CollabScore Dataset -Towards Robust and Generalized OMR Evaluation

Philippe Rigaux, Bertrand Coüasnon, Christophe Guillotel-Nothmann, Fabien Guilloux, Aurélie Lemaitre

The CollabScore Dataset. Towards Robust and Generalized OMR Evaluation

Philippe Rigaux
philippe.rigaux@cnam.fr
Cnam
Paris, France

Bertrand Coüasnon
bertrand.couasnon@irisa.fr
IRISA, Univ. Rennes
Rennes, France

Christophe Guillotel-Nothmann
Fabien Guilloux
Christophe.guillotel-nothmann@cnrs.fr
fabien.guilloux@cnrs.fr
IReMus, CNRS
Paris, France

Aurélie Lemaitre
aurelie.lemaitre@irisa.fr
IRISA, Univ. Rennes 2
Rennes, France

Abstract

We present in this paper the CollabScore dataset, initially built to test and validate a new Optical Music Recognition (OMR) system. This dataset includes 184 pages of scores from an homogeneous corpus which consists of various works from a single composer, Camille Saint-Saëns (1835-1921). In addition to the essential data, namely images and reference scores encoded in MEI, we describe an original approach to content referencing based on the IIIF standard, and we discuss some general principles to conduct OMR evaluation in the proposed framework, including the tuning of comparison metrics.

Keywords

Optical music recognition ; benchmark datasets

ACM Reference Format:
Philippe Rigaux, Bertrand Coüasnon, Christophe Guillotel-Nothmann, Fabien Guilloux, and Aurélie Lemaitre. 2026. The CollabScore Dataset. Towards Robust and Generalized OMR Evaluation. In *Proceedings of International Conference on Digital Libraries for Musicology (DLfM).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Many archives and digital libraries publish their musical collections in the form of digital images of sheet scores. This format does however not allow an access to the score content. This hinders the development of content-based services (such as listening, annotating and searching) and severely limits the usability and these digital libraries. Providing this material in the form of digital scores encoded in MusicXML [6], HumDrum **kern [7] or MEI [17] – which we will call in what follows *editable scores* to distinguish them from *image scores* – would offer much richer perspectives, likely to interest performers, musicologists and publishers as it would allow them to exploit and promote the *content* of digital archives.

Given the difficulty and cost of a manual transcription, the production of editable scores by optical recognition (OMR) seems particularly suitable and desirable, in the wake of text recognition systems (OCR) which are nowadays widely used. OMR has been the subject of much research [1, 2, 11, 15, 16], but does not seem, for the moment, to have produced methods robust and general enough to be used beyond the simplest cases. Optical recognition applied to musical notation raises indeed important challenges [13], due notably to the diversity of documents, to their heterogeneous quality, and to the countless variants of musical notation over the centuries. In addition, OMR research have been exploring many approaches, ranging from pure rule-based systems to recent *end-to-end* approaches based on *deep-learning* methods. This diversity gives rise to serious difficulties when it comes to reliably compare the different approaches, and more particularly to identify their respective strengths and weaknesses.

The CollabScore [3, 14] project (2020-2025), funded by the French national research agency (ANR) focuses on an original hybrid OMR approach [9], associated with a collaborative correction framework and a synchronization process of multimedia musical sources. As with any project of this type, we have been concerned from the beginning to establish the conditions of validation of our work, by building a reference dataset and a test environment allowing us to measure the effectiveness of our methods. We describe in the present paper this dataset and this environment. The contribution to a solid and widespread assessment basis for OMR complements recent works on automated score comparison [4] on the one hand, and on the other hand on the constitution of datasets of editable scores aligned with digital sources, facilitating the production of AI models and their evaluation [5, 8]. A recent and notable publication in this direction is the *Sheet Music Benchmark* [10] which associates a dataset of 600 images with their notation encoded in HumDrum [19]. The same paper also describes a comparison process based on MusicDiff [4].

The description of the testing environment designed in the context of CollabScore aims at completing these recent contributions.

DLfM, Thessaloniki, Greece

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

As mentioned above, the diversity of traditions, styles and notation idioms calls for multiplying this type of proposal in order to strengthen our collective capacity to evaluate OMR solutions in the finest way.

In a nutshell, the supplied dataset presents the following specifics. First of all, it is fully dedicated to a single composer, Camille Saint-Saëns (1835-1921), and proposes a focused sample of Western – and more particularly French – music notation idiosyncrasies in the second half of the 19th century and the early decades of the 20th century. Secondly, we propose a high-quality encoding of each score in MEI, to constitute what will be called *the reference dataset* in the following. Each of these score has been manually produced and aims at giving the finest possible representation that can be obtained by an informed editor with modern engravers. Regarding MEI, it seems difficult at this stage to evaluate which format of predicted editable scores is most suitable for a fair OMR evaluation. MusicDiff is initially designed for MEI scores, but the existence of converters makes it potentially applicable to other formats, as long as biases linked to conversion, or specific aspects of the format itself do not distort the results. Producing reference datasets in various encodings can be a way to identify such biases and improve our understanding of their impact.

We supplement the dataset description with that of an image-notation matching method that we believe to be original and likely to interest our colleagues working in this field. Indeed, investigating the output of an OMR process involves identifying the differences between the source (the image) and the prediction (the score after recognition and analysis). These differences are generally established indirectly by comparing the prediction and another score, called the reference score, which is supposed to reflect as closely as possible the analyzed image. This indirect method does however not allow to analyze what, within the image itself, led to a particular interpretation error.

As part of the project, we have designed a one-to-one matching of the reference score elements with the corresponding region of the image. These elements can be pages, systems, staves,or measures. With such a matching, it becomes very simple to associate the encoding of some part of the score, at a given level of granularity with the analyzed region. During the OMR testing campaign, the origin of an error can easily be found through visual analysis, and conversely, a region presenting a particular difficulty can be used as a source for analysis and validation. Our matching mechanism is based on the IIIF recommendation, which has become a *de facto* standard for referencing and sharing image archives.

The rest of this article is organized as follows. Section 2 is devoted to the dataset and its main characteristics. Section 3 presents the matching mechanism and the encoding of correspondences in the form of IIIF annotations. Finally, Section 4 discusses how the whole material can be used to evaluate a novel OMR system.

2 The dataset

The dataset is available at https://github.com/collabscore/dataset. It consists of 24 scores by Camille Saint-Saëns, totaling 184 pages, covering the main genres practiced by the composer with the exception of operas. For each score, the dataset provides (i) images of the original edition, taken from the Gallica digital library, (ii) a reference encoding in MEI format and (iii) a set of annotations linking images and regions in images, to the corresponding notation fragment in the reference score.

2.1 Images

All images come from the original editions, and are published on the Gallica website (https://gallica-bnf.fr) from the National Library of France (BnF). They are free of rights, can be directly viewed online and downloaded. Note that the source of a single score generally consists of a sequence of images, a part of which only a subsequence is devoted to music notation, with the other pages containing editorial information. See for example the link https://gallica.bnf.fr/ark:/12148/bpt6k11620473 to the *mélodie* "*Les coins bleus*" which we will use as an illustrating example in what follows.

Table 1 summarizes the contents of the dataset, with the genre of each piece, a link the source, and the number of regions identifiable as potential inputs for an OMR analysis (pages, systems or measures). Each of these regions is the target of an annotation that links it to the corresponding fragment of the music notation (see Section 3).

2.2 Reference scores

All scores have been transcribed manually from the images with the Sibelius software for the project. This transcription aims at fulfilling two objectives:

- providing a reference of the most precise notation that can be produced by a professional editor working with state-of-the-art engravers;
- faithfully reflecting the source in its editorial choices, including scrupulous respect of divisions in pages and systems, so that it can be considered as the target result of a "perfect" OMR capable of producing a faithful encoding of the source.

The format chosen for encoding is MEI (version 4.0). All MEI documents were obtained by export from the specific module included in the latest versions of Sibelius. A comparison with MusicXML, for which an export Sibelius also exists, would have been possible. However, the MEI format seemed more precise to us in a study context based on rigorous philological principles, and the initial design of MusicDiff as a MEI-based comparison tool seemed suitable for avoiding the distortions occurring during format conversions. The development of our dataset turned out to be an opportunity to observe several limitations or defects of the conversion systems and to suggest improvements.

Beyond these practical considerations, however, nothing attests the relevance of the MEI choice. Recall the that the recent paper [10] is based on another viable format, Humdrum **kern, a choice partly motivated by the availability of many scores [19]. At this point, the impact of encoding on the reliability of comparison results remains unclear, and a study on this point would perhaps deserve to be carried out.

2.3 Main characteristics

We briefly mention some notation characteristics which are of particular interest for an OMR analysis due to their large number of occurrences in our dataset.

Table 1: Summary of the dataset content

Title	Genre	Images	#parts	#pages	#systems	#measures	Lyrics	Clef changes	Ksign changes	Tsign changes	Cross-staff voices
[Dans les coins bleus]	Melody	Link	2	4	15	51	Y	Y	Y	N	Y
[Fière beauté]	Melody	Link	2	5	20	122	Y	Y	Y	N	Y
[God save the King]	National anthem	Link	2	2	7	28	Y	Y	Y	Y	N
[La sérénité]	Melody	Link	2	3	11	51	Y	Y	Y	Y	Y
[Dans ton cœur]	Melody	Link	2	4	15	48	Y	Y	Y	Y	Y
[Avril]	Melody	Link	2	5	19	66	Y	Y	Y	Y	Y
[L'Amour blessé]	Melody	Link	3	5	19	55	Y	Y	Y	Y	Y
[L'Amant malheureux]	Melody	Link	2	5	19	87	Y	Y	Y	Y	Y
[La coccinelle]	Melody	Link	2	2	9	53	Y	Y	Y	Y	Y
[La feuille de peuplier]	Melody	Link	2	4	16	98	Y	Y	Y	Y	Y
[Guitare]	Melody	Link	2	3	13	44	Y	Y	Y	Y	Y
[Quam dilecta]	Motet	Link	5	8	16	89	Y	Y	Y	Y	Y
[Calme des nuits]	Chorus	Link	5	5	10	58	Y	Y	Y	Y	Y
[Les fleurs et les arbres]	Chorus	Link	5	6	12	78	Y	Y	Y	Y	Y
[Madrigal]	Chorus	Link	4	7	21	124	Y	Y	Y	Y	Y
[Sérénade d'hiver]	Chorus	Link	4	14	58	216	Y	Y	Y	Y	N
[Les marins de Kermor]	Chorus	Link	5	15	54	250	Y	Y	Y	N	N
[Ode d'Horace]	Chorus	Link	4	12	47	257	Y	Y	Y	Y	N
[Les soldats de Gédéon]	Chorus	Link	8	38	80	357	Y	Y	Y	N	N
[Suite orchestre, Prélude]	Orchestral suite	Link	10	5	9	50	N	Y	Y	Y	N
[Suite orchestre, Sérénade]	Orchestral suite	Link	9	5	10	54	N	Y	Y	Y	N
[Suite orchestre]	Orchestral suite	Link	7	6	12	85	N	Y	Y	Y	N
[Danse macabre (Melody)]	Melody	Link	2	6	25	137	N	Y	Y	Y	Y
[Danse macabre (Poème symphonique)]	Symphonic poem	Link	3	15	69	471	N	Y	Y	Y	Y
Summary				184	586	2929					

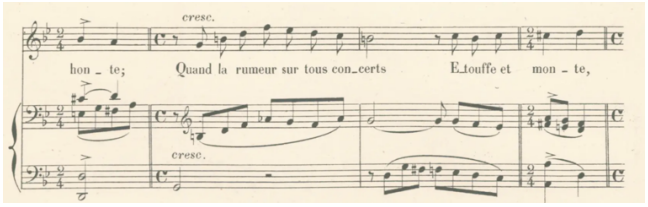

Figure 1: Frequent metric changes (Gallica link)

Music pieces by Saint-Saëns, and in particular his melodies, present very frequent key signature and key time signature changes, as illustrated by Figures 1 and 2.

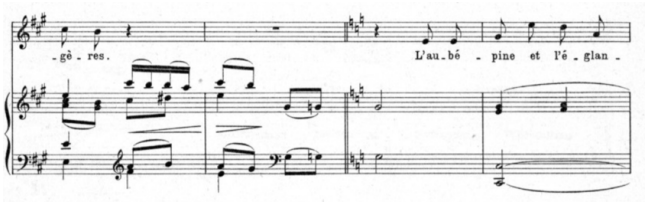

Figure 2: Key and key signature changes (Gallica link)

Another common feature of the dataset is the frequent crossing of voices from one staff to another in piano parts (Figure 3). It is also common to meet "partial" voices which cover only a part of a measure. In general, building the interpretation of a score as a consistent set of voices is one of most difficult OMR sub-tasks, and it turns out to be particularly intricate in our dataset.

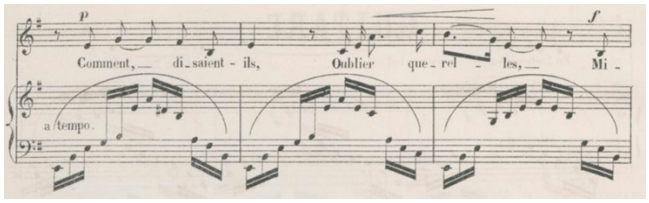

Figure 3: Multi staves voices (Gallica link)

Although they are by no way exceptional, these aspects well represented in CollabScore were not fully handled in the tools used for our experiments (format converters and MusicDiff), which had therefore to be adapted and extended. This remark underlines the interest, in our opinion, of constantly enlarging the variety of sources proposed for OMR validation, since no single dataset can claim to cover on its own all forms of musical notation, even those which appear relatively common.

Finally, our dataset contains a significant number of sung pieces (*mélodies* or *chœurs*, cf. Figure 4), the lyrics of which have been carefully reported into the MEI reference documents. The recognition of the text from the notation is a complex task due to the syllabic division, the reconstruction of words and the sometimes slightly misaligned position of the text with respect to its musical counterpart.

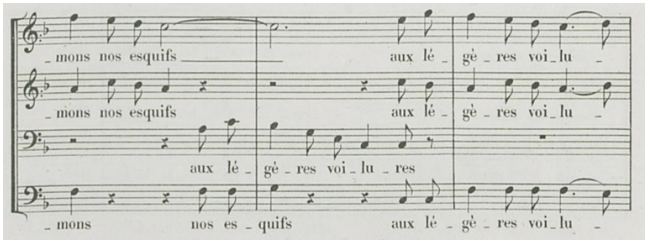

Figure 4: Lyrics (Gallica link)

3 Annotations

We now describe how we represent pairings, at different granularity levels, between the images on the one hand, and the elements of the reference MEI scores on the other hand. Generally speaking, these annotations take the form of bidirectional links associating a region in an image to a fragment in the MEI-XML encoding hierarchy.

3.1 Annotation levels

We distinguish three levels of annotation corresponding to the structure of a printed score, illustrated in Figure 5. To quote the MEI terminology, a "section" in such a score is a sequence of pages, themselves made up of systems, themselves made up of measures encompassing all the musical parts. Each of these levels is the subject of annotations linking each occurrence of a page/system/measure in the MEI document to the corresponding region on the image.

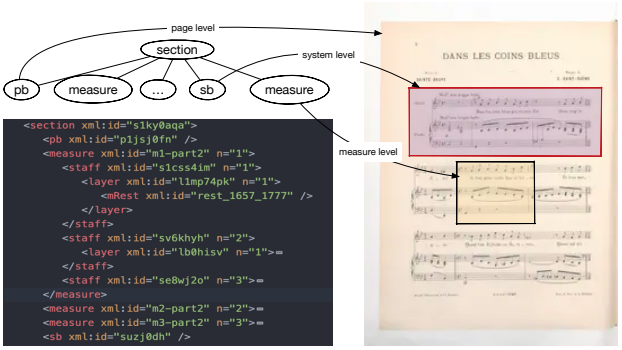

Figure 5: Annotation Levels

It would be possible to also annotate the staves (in a system or in a measure) but this level does not correspond to a well-identified element in the hierarchical structure of the MEI document.

In order to achieve a robust and stable representation of links, we chose to rely on the following standard mechanisms:

- *regions* are URIs conforming to the recommendations of the *International Image Interoperability Framework*, or IIIF[1];
- fragments in the MEI are referenced by XPath expressions [2];
- finally, links between regions and fragments of the notation are encoded in the form of annotations conforming to the W3C [3] *Web annotation model*.

It is probably unnecessary to insist on the advantages of using standards. Let us only mention the existence of a large number of software tools which allow, in the case of images, to browse, visualize and even transform regions, and in the case of XML, to extract, visualize, compare the corresponding musical notation.

3.2 Referencing regions

The *International Image Interoperability Framework* (https://iiif.io/) is an organization whose goal is to facilitate sharing of visual and audio-visual sources. It is supported by large institutions, such as national libraries, which manage large collections of digitized document and wish to make them available to worldwide stakeholders. The IIIF publishes recommendations, based on the W3C model. The two most significant for our goals are:

- The image API (https://iiif.io/api/image/3.0/), which defines a set of services applicable to an image.
- The presentation API (https://iiif.io/api/presentation/3.0/), which describes how to organize a set of visual or audio-visual resources in a structure (the *manifest*) gathering metadata, images references and annotations, thereby defining an organized presentation space for these resources.

The CollabScore dataset relies on these two APIs to describe and reference sheet score images, group them, and link them to music notation. An image API service is called by a URL whose general form (slightly simplified) is:

`{server}/{id}/{region}/{size}/{rotation}/{quality}.{format}`

We can therefore indicate the format, the quality, and specify the region of the image to extract (rotation and size seem to be useless in the case of an OMR extraction). All our images are accessible via the Gallica IIIF server provided by the BnF.[4] This server is available at https://openapi.bnf.fr/iiif/image/v3/. Images are identified in Gallica by an `ark` identifier. Here is for instance the prefix of IIIF services URLs on the first score of the dataset, "*Les coins bleus*":

`https://openapi.bnf.fr/iiif/image/v3/ark:/12148/bpt6k11620473/f2`

Figure 6 shows some examples of referencing regions in a score page, the URL `iiif_url` being the one indicated above. Each page, each system, each measure can be referenced by a IIIF service URL, the server being responsible for computing on the fly the corresponding content.

We provide all these region addresses as complementary data for CollabScore. They are mainly used to link a region with its corresponding reference notation. As such, it supports verification at different granularity levels of the prediction made by the evaluated OMR system .

There are at least two other possible uses of this detailed information on regions. The first is to build a catalog of typical difficult

[1] https://iiif.io/api/
[2] https://www.w3.org/TR/xpath/
[3] https://www.w3.org/TR/annotation-model/
[4] See https://www.bnf.fr/en/gallica-bnf-digital-library for a short presentation.

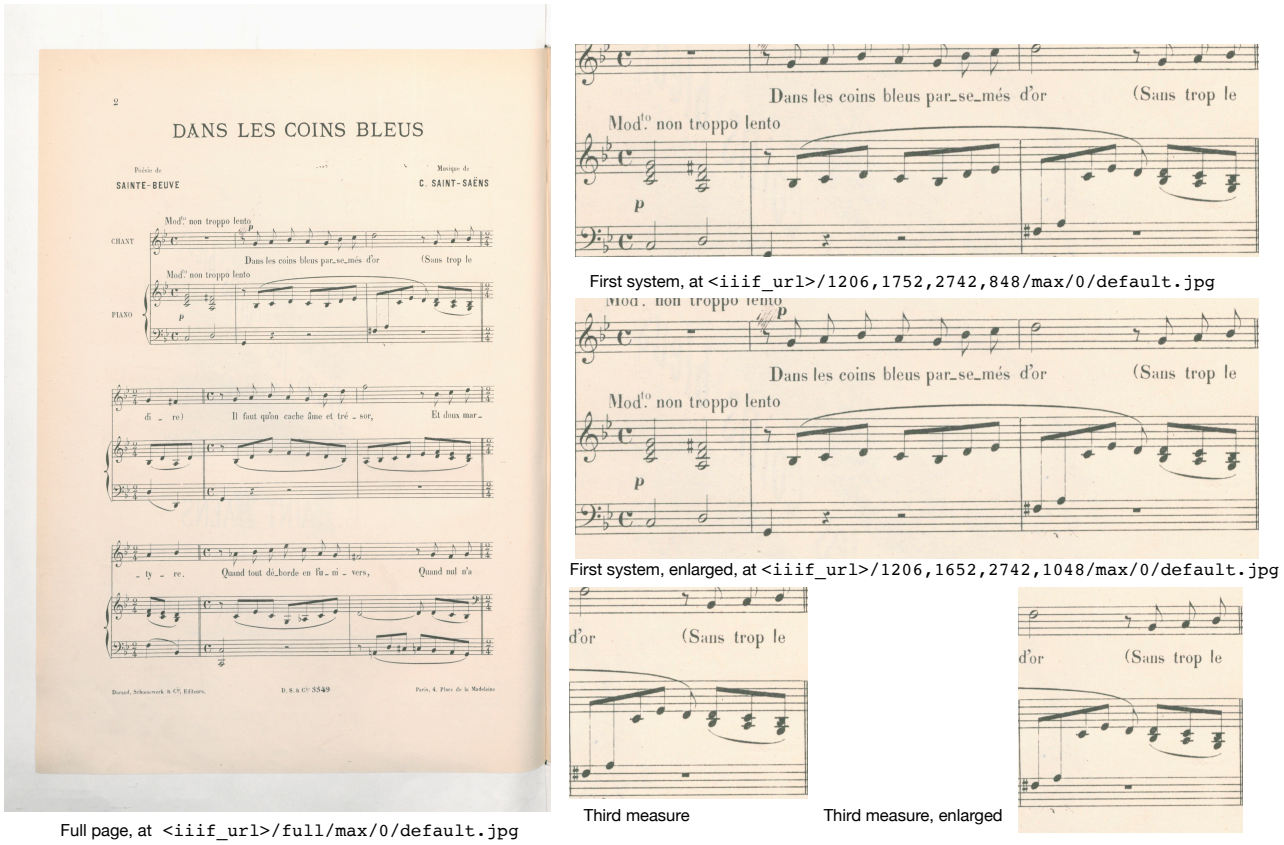

Figure 6: Examples of IIIF references (link to the full page)

OMR issues in order to test the ability of the system to process an isolated region or to cope with a specific difficulty. Figure 6 shows for example measure 3 of the first system, in which one of the voices moves from one staff to the other. This excerpt can be integrated into a unit testing system dedicated to this specific type of problem. One second use which naturally derives from it is to use these regions (and their associated notation) as a source to train a learning system.

Finally, note that it is very easy to slightly enlarge/distort/move regions in order to test the robustness of the recognition. The system and measures in Figure 6 are shown with two distinct regions: the first is strictly aligned on the staves, whereas the second is a vertical widening +/- 100 pixels.

3.3 Multi-page sources and IIIF manifests

A music score generally extends over several pages, and we must therefore provide a structure defining the sequence of pages. This sequence is to be given as input to the OMR whose goal is to rebuilt a complete score and not a list of independent pages interpretations.

Such a structure already exists in the IIIF recommendations under the name *manifest*. It takes the form of a JSON document whose structure is specified in the presentation API. We provide simply with the dataset a manifest for each score, as well as a Python utility allowing to extract any useful information, including the sequence of image URLs constituting the source.

Among other advantages, the use of a manifest allows to benefit from the management tools of the IIIF ecosystem, and in particular visualizers. Figure 7 shows the display of a score overview with the Mirador viewer[5].

3.4 Referencing fragments of musical notation

Notational elements (pages, systems, measures) in the MEI can be referenced by standard XML tools (e.g., XPath), as long as we assign them a unique and stable identifier in the source document.

The following identifiers are reported in the reference MEI:

- pages are identified by `p1`, `p2`, etc.
- systems are identified by their position in the page (e.g., `p1-s1`, `p1-s3`, etc.)
- finally measures are identified by their sequential number, (e.g. `m1`, `m2`, etc.)

Figure 8 shows an MEI fragment illustrating this identification system. It is important to note that the pages are numbered in relation to the first source page, i.e., the one *containing the notated music*, which may differ from the first "physical" page, often devoted

[5] https://projectmirador.org

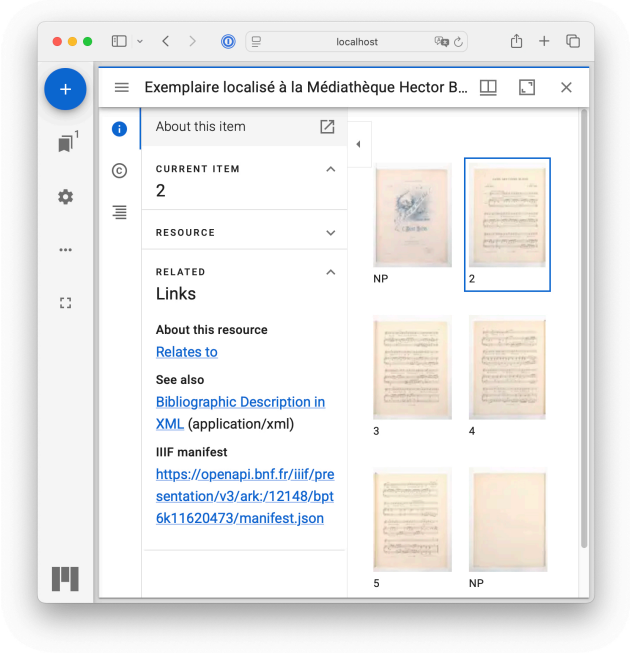

Figure 7: Inspecting a source

```
<music>
  <body>
    <mdiv xml:id="m1jx284v">
      <score xml:id="sgr3err">
        <scoreDef xml:id="s1m90xso" midi.bpm="120.000000">
        <section xml:id="s1n80u7l">
          <pb xml:id="p1" />
          <measure xml:id="m1" n="1">
          <measure xml:id="m2" n="2">
          <measure xml:id="m3" n="3">
            <sb xml:id="p1-s2" />
          <measure xml:id="m4" n="4">
          <measure xml:id="m5" n="5">
          <measure xml:id="m6" n="6">
          <sb xml:id="p1-s3" />
```

Figure 8: Idenfifiers in MEI files

to a title page and/or to an editorial introduction. We assume that any OMR system should know how to identify these music pages from the complete sequence, but we also provide this information in the metadata.

3.5 Linking data with annotations

Finally, we provide the links between images, regions (at the different considered levels) and XML fragments corresponding to the MEI notation as *annotations* conforming to the W3C recommendation [18]. Here is an example that will suffice to understand the principle.

```
{
  "motivation": "linking",
  "body": {
     "source": "<mei_source>",
     "selector": {
       "type": "FragmentSelector",
       "conformsTo": "http://tools.ietf.org/rfc/rfc3023",
       "value": "p1-s2"
     }
  },
  "target": "<iiif_url>"
}
```

The annotated elements, or *target*, are image regions, referenced by service IIIF URLs. The *body* of the annotation is a link to the XML fragment containing the notation. In the W3C annotation model, such a link associates a source (the MEI document) and the expression selecting the fragment. Here, by convention, it is the page Id, system Id or measure Id.

An important issue is to determine the correct place to insert these annotations. Since they represent bidirectional links, several options can be (non exclusively) considered: (i) along with the score encoding, (ii) in the manifest together with image references, or (iii) as an independent resource.

A first option would be to insert annotations directly in the MEI document, using a specific namespace. We also considered using the `facsimile` module mentioned in the MEI recommendation to indicate the sources, and areas within a source. We however, gave up this solution which seems redundant with the already adopted IIIF approach.

In MEI V5.0, an integration with IIIF seems envisaged, and we believe indeed that it would be the right approach to take. However, the topic is not at all elaborated at the time writing (April 2026).

In summary, we adopted for the time being a neutral solution by supplying an independent document containing all annotations. This does not preclude to witness, in the future, a strong integration of external references with music score encoding, and possibly its support for display in tools like Verovio [12]. Such an evolution would lead us to the decision of enriching the dataset accordingly.

4 Conducting a comparison

We end with a section commenting an OMR evaluation process supported by the CollabScore dataset. The following remarks are based on our experience and mainly aim at contributing to advances in the current methodologies that permit an evaluation of a specific OMR system.

A description of the OMR developed in the CollabScore project is beyond the scope of the present paper, but the interested reader is referred to [9] which also includes a detailed evaluation based on the dataset. The GitHub site also contains additional instructions on how an evaluation campaign can be carried out based on the method described thereafter.

4.1 Evaluating an OMR system

Optical music recognition is the inverse of the common creation of a score sheet. Traditionally, one produces the notation, and the engraving software generates the graphical representation. If one wants to change something, add or correct the score, the notation must be edited and the new representation is derived from the updated notation.

OMR works just the opposite: it observes the symbols and their position in the input image, and the graphical/geometric information is the domain which must be checked and possibly corrected.

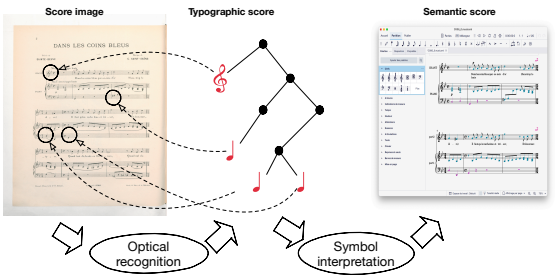

Figure 9: OMR digitization workflow

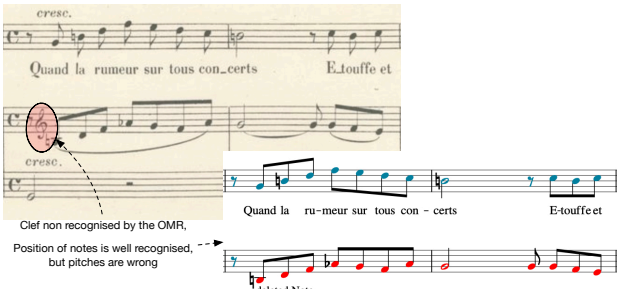

Figure 10: Propagating a key signature recognition error

In the case of a note for example, the pitch is derived from staff position and other contextual information, not given directly.

A conceptual processing pipeline that models the steps involved in the optical recognition of a musical source is illustrated on Fig. 9. It consists in two steps. *Optical recognition* takes as input an image (or, more generally, a sequence of images) and produces an encoding of the typographical layer of music notation, reflecting what has been perceived on the source image. Let us call this object *Typographic score* (Tscore for short). A Tscore combines music notation symbols (e.g., clefs, signatures, notes, rests, chords, etc.) and their graphical relationship in a specific structure which is not (yet) a music notation format such as MusicXML or Kern.

Each symbol in an Tscore refers to a graphical element identified in the image. For instance the G clef *symbol* in Fig. 9 refers to the first clef of the score, the three quarter notes *symbols* refer to notes located on various systems and staves, etc. At this point, no interpretation is done, and we do not know whether a note symbol correspond to a G, B, A#, etc, since such an interpretation depends on a (possibly complex) context made of clefs, signatures and former accidentals.

The second phase produces the *semantic score*, i.e., an encoding in some standard format (say, MEI) of the corrected Tscore where all symbols have been interpreted as valued music objects in the music notation domain.

It seems therefore clear that the OMR evaluation must apply to the Tscore, and *not* to the semantic score. The major argument is that, in the semantic score, object valuations are interdependent. The values assigned to a notational element depend on a context which may have influenced by events located very far away from the current location. It follows that a single recognition error *propagates* to dependent elements.

Figure 10 shows a key change occurring on one of the staves. Let us assume that this key is not being recognized by the OMR system. If we take, as input of the comparison process, the score after full interpretation, the pitches that depend on the missing key will all be considered as false (and reported in red in the comparison tool) because interpreted on the basis of a wrong context.

The symbol detection did not actually make any errors on the *position* of these notes in the staff, nor on their other graphical properties. A fair evaluation should therefore report a single error, the one made on the missing key, and consider as correct the note symbols recognition since their graphical properties (including the position on staff) have been correctly identified.

4.2 The metric

The comparison method proposed with our dataset relies on `MusicDiff`, first published in [4], and now available as a Python package actively maintained and supported[6]. MusicDiff is designed to compute the edit distance between two music score documents, in the spirit of the Unix diff tool that compares two text files. It produces a list of the differences between the two scores, each "diff" being a pseudo-edition applied to a target element (a note, a rest, a symbol, etc.). The diff list is the minimal set of editions necessary to transform one score (in our case, the OMR output or "predicted" score) into the other (in our case the ground-truth or "reference" score).

This permits comparisons, but does not directly yield a normalized metric. Recently, the authors of [10] the OMR-NED, a global metric that divides the number of editions by the total number of elements in the notation. Intuitively, this represents the amount of the predicted score that needs to be modified to obtain the reference score.

We adopt a similar approach, but introduced two important improvements. First, as advocated above, the comparison must be made at the typographic level, not the semantic one. Symbols must be compared with respect to their graphical properties, and not on their musical domain interpretation. New comparators based on these graphical properties have been introduced in MusicDiff. Secondly, a single comparison indicator such as OMR-NED glues together many aspects which may or may not be significant depending on the context. We chose to distinguish three categories of element, as illustrated by Fig. 11.

Voice elements. (in red, Fig. 11) cover notes, rests, and chords. We define the (voice elements) *edition rate* as the number of required editions (i.e., the number of differences) divided by the total number of musical elements (note, rest, chord).

Contextual elements. (in blue in Fig. 11) ensure the correct interpretation of voice elements for a performing musician. Those elements are either a key signature, time signature, or a clef. We define the (context elements) *edition rate* as the number of differences concerning contextual elements divided by the total number of context elements in the score (clef, key signature, time signature, in blue on the Figure), excluding redundancies (in grey).

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

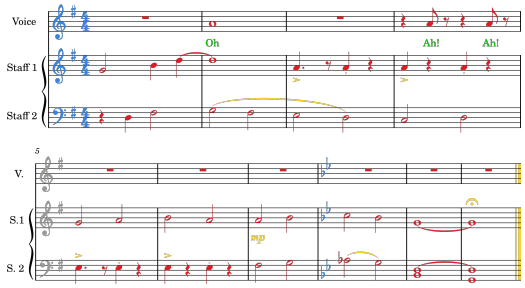

Figure 11: Elements evaluated by MusicDiff: Voice elements (in red), Contextual elements (in blue), and Lyrics (in green).

Lyrics. (in green in Fig.11) recognition is also evaluated. A lyric is correct if: (i) it is associated to the correct note, (ii) it is related to the correct verse, (iii) the syllable position in the current word (beginning, middle or end) is correctly reported and (iv) its transcription is identical (case-sensitive).

The code of this metric can be downloaded with the CollabScore dataset, and the evaluation report can be reproduced. The interested reader is referred to [9] and to the GitHub documentation for an in-depth presentation of our OMR system, its results and how it compares to the ground-truth on the one hand, and to a commercial system, namely PhotoScore [7]. Note that both the CollabScore and Photoscore predictions are shipped with the dataset and can be found on the GitHub, allowing reproducibility and comparison with other systems or future versions.

5 Conclusion

We propose a new dataset which complements recent contributions aimed at constituting a robust and documented basis supporting OMR validation and evaluation. The CollabScore dataset has strong specificities. It is centered on a single composer and extensively presents the particularities of the musical notation in the second half of the 19th century and the early decades of the 20th century. As such, it cannot be used for general validation of the OMR, but it seems clear to us that the variety of traditions regarding music notation makes a general validation illusory, and calls for multiplying specialized corpora. Our contribution is therefore a step in this direction.

We provide 184 pages of sheet music, associating reference images and their sub-regions with the corresponding notation. Beyond this essential content, we describe our testing environment. Images of the same score are grouped together so as not to be limited to single-page recognition. We also offer a set of data and tools that aim at facilitating an evaluation process. The analytical information on the dataset, and in particular the specification in the form of annotations of the regions containing the notation, can help to control the advantages and limitations of a particular OMR solution, and in any case facilitate their understanding.

We are moving towards a well-accepted practice of OMR evaluation. Our work is indeed fully consistent with recent proposals. We employ a common metholology, and rely on common tools (which benefit from constant improvements and extensions). This suggests that we will, in the future, be able to assess the results of OMR systems with respect to a common yardstick. We hope therefore that the present work will contribute to the emergence of robust transcription tools and to the larger adoption of OMR as a powerful and reliable method in the future.

[7] See https://www.avid.com/products

Acknowledgments

This work has been partially funded by the French national research agency (ANR). We are also extremely grateful to Greg Chapman for his availability and willingness to adapt the MusicDiff package to our needs, and to the anonymous reviewers for their comments that helped considerably to improve the content of the present paper.

References

[1] Calvo-Zaragoza, J., Hajic, J., and Pacha, A. Understanding optical music recognition. *ACM Computing Surveys (CSUR) 53* (2019), 1 – 35.

[2] Calvo-Zaragoza, J., Martinez-Sevilla, J. C., Penarrubia, C., and Rios-Vila, A. Optical music recognition: Recent advances, current challenges, and future directions. In *Document Analysis and Recognition – ICDAR 2023 Workshops* (2023), pp. 94–104.

[3] Coüasnon, B. B., Guillotel-Nothmann, C., Giraud, M., Lemaitre, A., and Rigaux, P. The CollabScore project – From Optical Recognition to Multimodal Music Sources. In *Proc. of Intl. Workshop on Reading Music Systems* (2024), pp. 33–37.

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

[5] Foscarin, F., McLeod, A., Rigaux, P., Jacquemard, F., and Sakai, M. ASAP: a dataset of aligned scores and performances for piano transcription. In *Proc. Intl. Society for Music Information Retrieval Conference (ISMIR)* (2020), pp. 534–541.

[6] Good, M. *The Virtual Score: Representation, Retrieval, Restoration.* W. B. Hewlett and E. Selfridge-Field, MIT Press, 2001, ch. "MusicXML for Notation and Analysis", pp. 113–124.

[7] Huron, D. *Beyond MIDI : The Handbook of Musical Codes*, vol. 1. The MIT Press, 1997, ch. Humdrum and Kern: Selective Feature Encoding.

[8] jr., J. H., and Pecina, P. The MUSCIMA++ dataset for handwritten optical music recognition. In *proc. Intl. Conf. on Document Analysis and Recognition, (ICDAR)* (2017), pp. 39–46.

[9] Lemaitre, A., Coüasnon, B., and Rigaux, P. CollabScore OMR: A Hybrid System for Music Score Recognition. Submitted to ICDAR, 2026.

[10] Martinez-Sevilla, J. C., Cerveto-Serrano, J., Luna-Barahona, N. N., Chapman, G., Sapp, C., Rizo, D., and Calvo-Zaragoza, J. Sheet music benchmark: Standardized optical music recognition evaluation. In *Proc. Intl. Society for Music Information Retrieval Conference, (ISMIR)* (2025), pp. 604–611.

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

[12] Pugin, L., Zitellini, R., and Roland, P. Verovio: A library for engraving mei music notation into svg. In *ISMIR* (2014), pp. 107–112.

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

[14] Rigaux, P., Coüasnon, B., Cretin, S., Étienne, V., Giraud, M., Leguy, E., Lemaitre, A., Guillotel-Nothmann, C., Guilloux, F., Gurrieri, M., and Vernet, T. Linking and Combining Digital Music Sources in Digital Archives: The CollabScore approach. In *IAML 2026 - International Association of Music Libraries Congress* (2026).

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

[16] Ríos-Vila, A., Rizo, D., Iñesta, J. M., and Calvo-Zaragoza, J. End-to-end optical music recognition for pianoform sheet music. *International Journal of Document Analysis and Recognition (IJDAR) 26*, 3 (2023), 347–362.

[17] Rolland, P. The Music Encoding Initiative (MEI). In *Proceedings of the International Conference on Musical Applications Using XML* (2002), pp. 55–59.

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

[19] Sapp, C. S. Online Database of Scores in the Humdrum File Format . In *Proceeding of International Conference on Music Information Retrieval (ISMIR)* (2005).