

# Understanding Optical Music Recognition

Jorge Calvo-Zaragoza<sup>\*</sup>, *University of Alicante*

Jan Hajič Jr.<sup>\*</sup>, *Charles University*

Alexander Pacha<sup>\*</sup>, *TU Wien*

## Abstract

For over 50 years, researchers have been trying to teach computers to read music notation, referred to as Optical Music Recognition (OMR). However, this field is still difficult to access for new researchers, especially those without a significant musical background: few introductory materials are available, and furthermore the field has struggled with defining itself and building a shared terminology. In this work, we address these shortcomings by (1) providing a robust definition of OMR and its relationship to related fields, (2) analyzing how OMR inverts the music encoding process to recover the musical notation and the musical semantics from documents, (3) proposing a taxonomy of OMR, with most notably a novel taxonomy of applications. Additionally, we discuss how deep learning affects modern OMR research, as opposed to the traditional pipeline. Based on this work, the reader should be able to attain a basic understanding of OMR: its objectives, its inherent structure, its relationship to other fields, the state of the art, and the research opportunities it affords.

## Index Terms

Optical Music Recognition, Music Notation, Music Scores

## I. INTRODUCTION

Music notation refers to a group of writing systems with which a wide range of music can be visually encoded so that musicians can later perform it. In this way, it is an essential tool for preserving a musical composition, facilitating permanence of the otherwise ephemeral phenomenon of music. In a broad, intuitive sense, it works in the same way that written text may serve as a precursor for speech. Similar to Optical Character Recognition (OCR) technology that has enabled the automatic processing of written texts, reading music notation also invites automation. In an analogy to OCR, the field of Optical Music Recognition (OMR) covers the automation of this task of “reading” in the context of music. However, while musicians can read and interpret very complex music scores, there is still no computer system that is capable of doing so.

We argue that besides the technical challenges, one reason for this state of affairs is also that OMR has not defined its goals with sufficient rigor to formulate its motivating applications clearly, in terms of inputs and outputs. Work on OMR is thus fragmented, and it is difficult for a would-be researcher, and even harder for external stakeholders such as librarians, musicologists, composers, and musicians, to understand and follow up on the aggregated state of the art. The individual contributions are formulated with relatively little regard to each other, although less than 500 works on OMR have been published to date. This makes it hard to combine the numerous contributions and use previous work from other researchers, leading to frequent “reinventions of the wheel.” The field, therefore, has been relatively opaque for newcomers, despite its clear, intuitive appeal.

One reason for the unsatisfactory state of affairs was a lack of practical OMR solutions: when one is hard-pressed to solve basic subproblems like staff detection or symbol classification, it seems far-fetched to define applications and chain subsystems. However, some of these traditional OMR sub-steps, which do have a clear definition and evaluation methodologies, have recently seen great progress, moving from

<sup>\*</sup> Equal contribution

the category of “hard” problems to “close to solved,” or at least clearly solvable [70], [118]. Therefore, the breadth of OMR applications that have long populated merely the introductory sections of articles now comes within practical reach. As the field garners more interest within the document recognition and music information retrieval communities [11], [34], [92], [114], [1], [50], [136], [78], [83], we see further need to clarify how OMR talks about itself.

The primary contributions of this paper are to clearly define what OMR is, what problems it seeks to solve and why.<sup>1</sup> OMR is, unfortunately, a somewhat opaque field due to the fusion of the music-centric and document-centric perspectives. Even for researchers, it is difficult to clearly relate their work to the field, as illustrated in Section II.

Many authors also think of OMR as notoriously difficult to evaluate [29], [84]. However, we show that this clarity also disentangles OMR tasks which are genuinely hard to evaluate, such as full re-typesetting of the score, from those where established methodologies can be applied straightforwardly, such as searching scenarios.

Furthermore, the separation between music notation as a visual language and music as the information it encodes is sometimes not made clear, which leads to a confusing terminology. The way we formulate OMR should provide a framework of thought in which this distinction becomes obvious.

In order to be a proper tutorial on OMR, this paper addresses certain shortcomings in the current literature, specifically by providing:

- A robust definition of what OMR is, and a thorough analysis of its inherent structure;
- Terminological clarifications that should make the field more accessible and easier to survey;
- A review of OMR uses and applications; well-defined in terms of inputs and outputs, and—as much as possible—recommended evaluation methodologies;
- A brief discussion of how OMR was traditionally approached and how modern machine learning techniques (namely deep learning) affect current and future research;
- As supplementary material, an extensive, extensible, accessible, and up-to-date bibliography of OMR (see Appendix A: OMR Bibliography).

The novelty of this paper thus lies in collecting and systematizing the fragments found in the existing literature, all in order to make OMR more approachable, easier to collaborate on, and—hopefully—progress faster.

## II. WHAT IS OPTICAL MUSIC RECOGNITION?

So far, the literature on OMR does not really share a common definition of what OMR is. Most authors agree on some intuitive understanding, which can be sketched out as “computers reading music.” But until now, no rigorous analysis of this question has been carried out, as most of the literature on the field focuses on providing solutions—or, more accurately, solutions to certain subproblems. These solutions are usually justified by a certain envisioned application or by referencing a review paper that elaborates on common motivations, with Rebelo et al. [133] being the most prominent one. However, even these review papers [22], [7], [133], [111] focus almost exclusively on technical OMR solutions and avoid elaborating the scope of the research.

A critical review of the scientific literature reveals a wide variety of definitions for OMR with two extremes: On one end, the proposed definitions are clearly motivated by the (sub)problem which the authors sought to solve (e.g., “transforming images of music scores into MIDI files”) which leads to a definition that is too narrow and does not capture the full spectrum of OMR. On the other end, there are some definitions that are so generic that they fail to outline what OMR actually is and what it tries to achieve. An obvious example would be to define OMR as “OCR for music.” This definition is overly vague, and the authors are—as likewise in many other papers—particularly unspecific when it comes to clarifying what it actually includes and what not. We have observed that the problem statements and definitions in

<sup>1</sup>We assume that readers are familiar with the basics of music notation. While we do revisit some concepts, we recommend Schobrun’s step-by-step guide [144] for a thorough introduction into music notation.

these papers are commonly adapted to fit the provided solution or to demonstrate the relevance to a particular target audience, e.g., computer vision, music information retrieval, document analysis, digital humanities, or artificial intelligence.

While people rely on their intuition to compensate for this lack of accuracy, we would rather prefer to put an umbrella over OMR and name its essence by proposing the following definition.

**Definition 1.** Optical Music Recognition is a field of research that investigates how to computationally read music notation in documents.

The first claim of this definition is that OMR is a *research field*. In the published literature, many authors refer to OMR as “task” or “process,” which is insufficient, as OMR cannot be properly formalized in terms of unique inputs and outputs (as discussed in Section VI). OMR must, therefore, be considered something bigger, like the embracing research field, which investigates how to provide a computer with the ability to read music notation. Within this research field, several tasks can be formulated with specific, unambiguous input-output pairs.

The term “*computationally*” distinguishes OMR from the musicological and paleographic studies of how to decode a particular notation system. It also excludes studying how humans read music. OMR does not study the music notation systems themselves—rather, it builds upon this knowledge, with the goal that a computer should be able to read the music notation as well.

The last part of the definition “*reading music notation in documents*” tries to define OMR in a concise, clear, specific, and inclusive way. To fully understand this part of the definition, the next section clarifies what kind of information is captured in a music notation document and outlines the process by which it gets generated. The subsequent section then elaborates on how OMR attempts to invert this process to read and recover the encoded information.

It should be noted that the output of OMR is omitted intentionally from its definition, as different tasks require different outputs (see Section VI) and specifying any particular output representation would make the definition unnecessarily restrictive.

To conclude this section, Fig. 1 illustrates how various definitions of OMR in the literature relate to our proposed definition and are captured by it.

### III. FROM “MUSIC” TO A DOCUMENT

Music can be conceptualized as a structure of *notes in time*. This is not necessarily the only way to conceptualize music,<sup>2</sup> but it is the only one that has a consistent, broadly accepted visual language used to transmit it in writing, so it is the conceptualization we consider for the purposes of OMR. A note is a musical object that is defined by four parameters: *pitch*, *duration*, *loudness*, and *timbre*. Additionally, it has an *onset*: a placement onto the axis of time, which in music does not mean wall-clock time, but is measured in relative units called beats.<sup>3</sup> Periods of musical time during which no note is supposed to be played are marked by rests, which only have an onset and a duration. Notes and rests can simultaneously be grouped into graphical hierarchies, e.g., multiple eighth notes that are joined by a single beam, as well as logical hierarchies such as phrases or voices. This structure is a vital part of music—it is essential to work it out for making a composition comprehensible.

In order to record this “conceptualization of music” visually, for it to be performed over and over in (roughly) the same way, at least at the relatively coarse level of notes, multiple music notation systems have evolved. A music notation system is a visual language that aims to encode music into a graphical form and enrich it with information on *how to perform* it (e.g., bowing marks, fingerings, or articulations).<sup>4</sup>

<sup>2</sup>As evidenced by either very early music (plainchant) or some later twentieth century compositional styles (mostly spectralism).

<sup>3</sup>Musical time is projected onto wall-clock time with an underlying tempo, which can further be stretched and compressed by the performer. Strictly speaking, the notion of beats might not be entirely applicable to some very early music and some contemporary music, where the rhythmic pulse is not clearly defined. However, the notation used to express such music usually does have beats.

<sup>4</sup>Feist [57] refers to notation whimsically as a “haphazard Frankenstein soup of tangentially related alphabets and hieroglyphics via which music is occasionally discussed amongst its wonkier creators.”

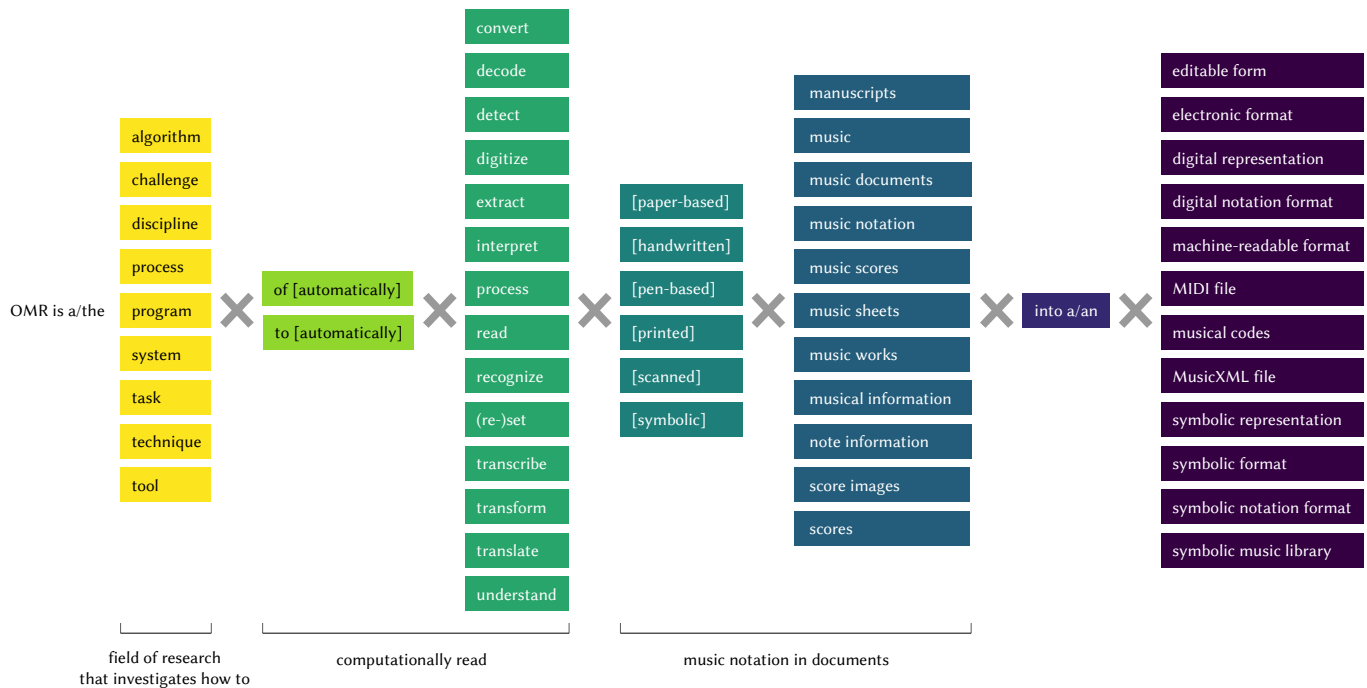


Fig. 1: How OMR tends to be defined or described and how our proposed definition relates to it. For example: “OMR is the challenge of (automatically) converting (handwritten) scores into a digital representation.”

To do that, it uses a set of symbols as its alphabet and specific rules for how to position these symbols to capture a musical idea. Note that all music notation systems entail a certain loss of information as they are designed to preserve the most relevant properties of the composition very accurately, especially the pitches, durations, and onsets of notes, while under-specifying or even intentionally omitting other aspects. Tempo could be one of these aspects, where the composer might have expressed precise metronomic indication, given a verbal hint, or stated nothing at all. It is therefore considered the responsibility of the performer to fill those gaps appropriately. We consider this as a natural boundary of OMR: it ends where musicians start to disagree over the same piece of music.

Arguably the most frequently used notation system is *Common Western Music Notation* (CWMN, also known as modern staff notation), which evolved during the seventeenth century from its mensural notation predecessors and stabilized at the beginning of the nineteenth century. There have been many attempts to replace CWMN, but so far, these have not produced workable alternatives. Apart from CWMN, there exist a wealth of modern tablature scores for guitar, used i.e., to write down popular music as well as a significant body of historical musical documents that use earlier notation systems (e.g., mensural notations, quadratic notation for plainchant, early organum, or a wealth of tablature notations for lutes and other instruments).

Once a music notation system is selected for writing down a piece of music, it is still a challenging task to engrave the music because a single set of notes can be expressed in many ways.<sup>5</sup> The engraving process involves complex decisions [24] that can affect only a local area, like spacings between objects but can also have global effects, like where to insert a page break to make it convenient for the musician to turn the page. For example, one must make sure that the stem directions mark voices consistently and

<sup>5</sup>Normally, music engraving is defined as the process of drawing or typesetting music notation with a high quality for mechanical reproduction. However, we use the term to refer to “planning the page”: selecting music notation elements and planning their layout to most appropriately capture the music, before it is physically (or digitally) written on the page. This is a loose analogy to the actual engraving process, where the publisher would carefully prepare the printing plates from soft metal, and use them to produce many copies of the music; in our case, this “printing process” might not be very accurate, e.g., in manuscripts.



Fig. 2: Excerpt of Robert Schumann's "Kinderszenen", Op. 15 for piano. Properly engraved (a), it has two staves for the left and the right hand with three visible voices, a key signature and phrase markings to assist the musician. In a poor engraving of the same music (b), that logical structure is lost, and it becomes painfully hard to read and comprehend the music, although these two versions contain the same notes.

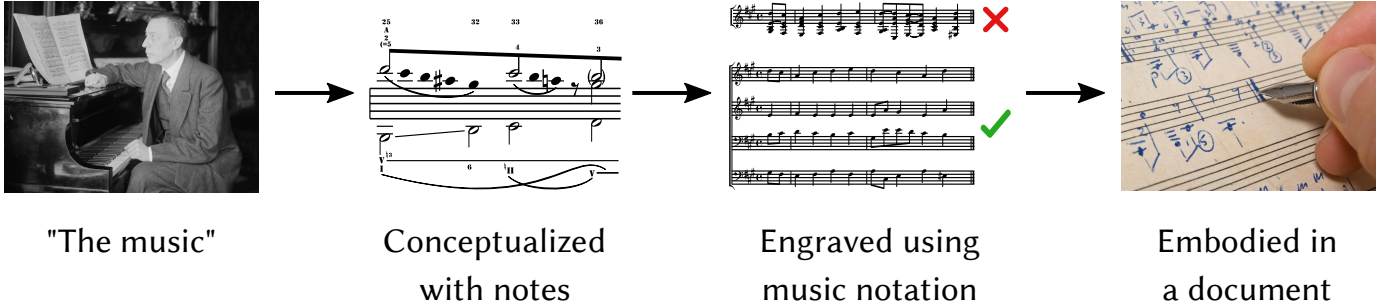


Fig. 3: How music is typically expressed and embodied (written down).

appropriate clefs are used, in order to make the music as readable as possible [89], [145], [79], [57]. These decisions not only affect the visual appearance but also help to preserve the logical structure (see Fig. 2). Afterwards, they can be embodied in a document, whether physically or digitally.

To summarize, music can be formalized as a structured assembly of notes, enriched through additional instructions for the performer that are encoded visually using a music notational language and embodied in a medium such as paper (see Fig. 3). Once this embodiment is digitized, OMR can be understood in terms of inverting this process.

#### IV. INVERTING THE MUSIC ENCODING PROCESS

OMR starts after a musical composition has been expressed visually with music notation in a document.<sup>6</sup> The music notation document serves as a medium, designed to encode and transmit a musical idea from the composer to the performer, enabling the recovery and interpretation of that envisioned music by reading through it. The performer would:

<sup>6</sup>In this context, a document is usually understood as a digital image. See Section VI-A for alternatives.

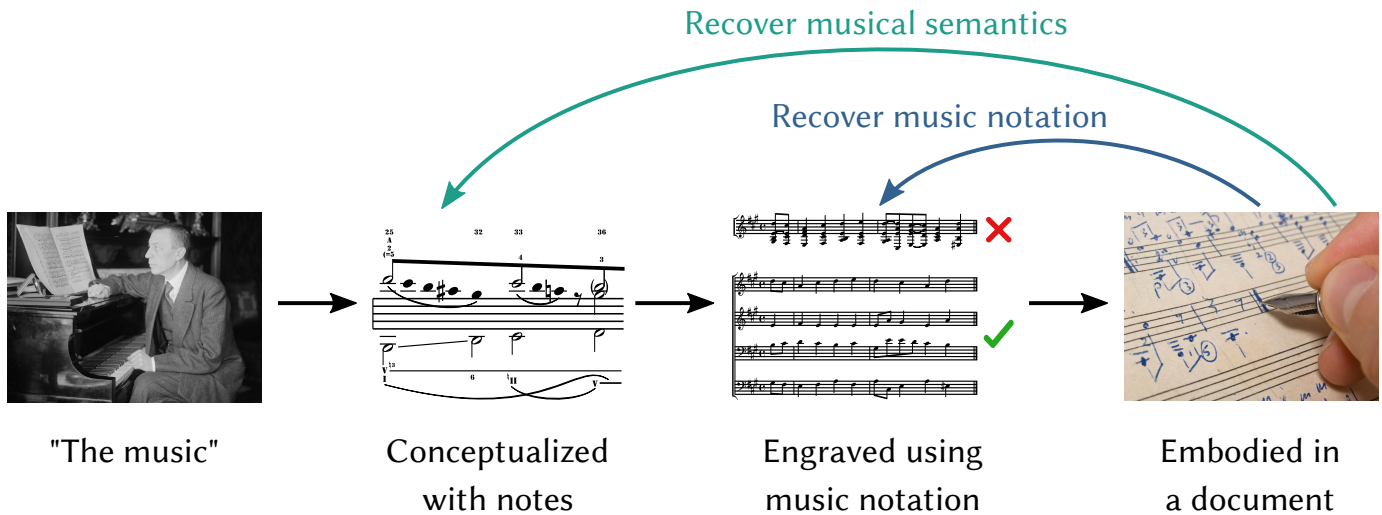


Fig. 4: How “reading” music can be interpreted as the operations of inverting the encoding process.

- 1) *Read the visual signal* to determine what symbols are present and what is their configuration,
- 2) Use this information to *parse and decode the notes and their accompanying instructions* (e.g., indications of which technique to use), and
- 3) Apply musical intuition, prior knowledge, and taste to *interpret the music* and fill in the remaining parameters which music notation did not capture.

Note that step (3) is clearly outside of OMR since it needs to deal with information that is not written into the music document—and where human performers start to disagree, although they are reading the very same piece of music [98].<sup>7</sup> Coming back to our definition of OMR, based on the stages of the writing/reading process we outlined above, there are two fundamental ways to interpret the term “read” in *reading music notation* as illustrated in Fig. 4. We may wish to:

- (A) *Recover music notation* and information from the engraving process, i.e. what elements were selected to express the given piece of music and how they were laid out. This corresponds to stage (1) in the analysis above and does not necessarily require specific musical knowledge, but it does require an output representation that is capable of storing music notation, e.g., MusicXML or MEI, which can be quite complex.
- (B) *Recover musical semantics*, which we define as the notes, represented by their pitches, velocities, onsets, and durations. This corresponds to stage (2)—we use the term “semantics” to refer only to the information that can be unambiguously inferred from the music notation document. In practical terms, MIDI would be an appropriate output representation for this goal.

This is a fundamental distinction that dictates further system choices, as we discuss in the next sections. Note that, counter-intuitively, going backwards through this process just one step (A - recover music notation) might be in fact more difficult than going back two steps (B - recover musical semantics) directly. This is because music notation contains a logical structure and more information than simply the notes. Skipping the explicit description of music notation allows this complexity to be bypassed.

There is, of course, a close relationship between recovering music notation and musical semantics. A single computer program may even attempt to solve both at the same time because once the full score with all its notational details is recovered, the musical semantics can be inferred unambiguously. Keep in mind that the other direction does not necessarily work: if only the musical semantics are restored from a document without the engraving information that describes how the notes were arranged, those

<sup>7</sup>Analogously, speech synthesis is not considered a part of optical character recognition. However, there exists expressive performance rendering software that attempts to simulate more authentic playback, addressing step (3) in our analysis. More information can be found in [36].

notes may still be typeset using meaningful engraving defaults, but the result is probably much harder to comprehend (see Fig. 2b for such an example).

#### A. Alternative Names

*Optical Music Recognition* is a well-established term, and we do not seek to establish a new one. We just notice a lack of precision in its definition. Therefore, it is no wonder that people have been interpreting it in many different ways to the extent that even the optical detection of lip motion for identifying the musical genre of a singer [53] has been called OMR. Alternative names that might not exhibit this vagueness are Optical Music Notation Recognition, Optical Score Recognition, or Optical Music Score Recognition.

### V. RELATION TO OTHER FIELDS

Now that we have thoroughly described what *Optical Music Recognition* is, we briefly set it in context of other disciplines, both scientific and general fields of human endeavors.

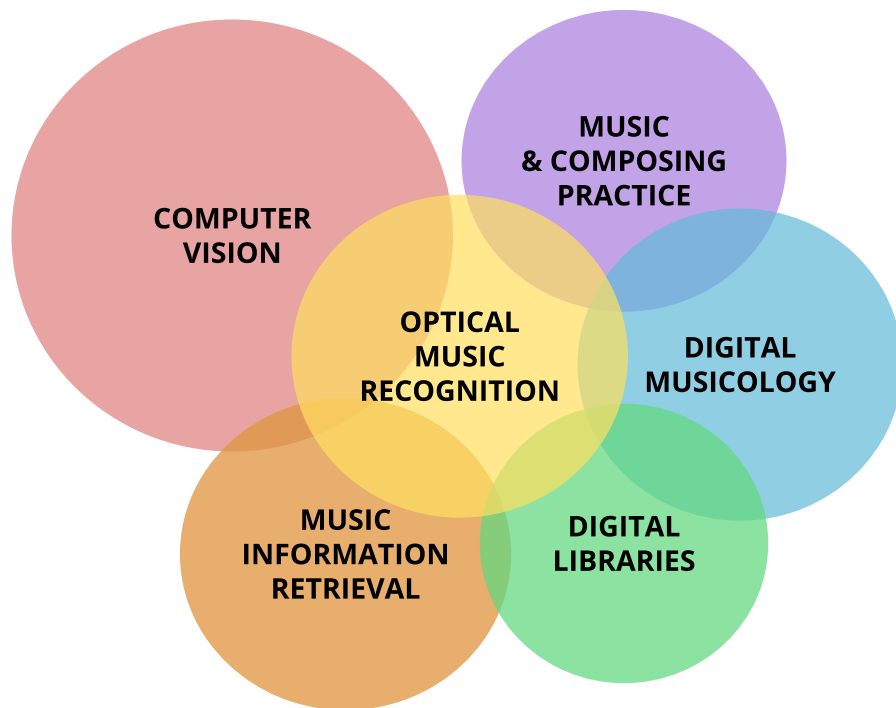


Fig. 5: Optical Music Recognition with its most important related fields, methods, and applications.

Figure 5 lays out the various key areas that are relevant for OMR, both as its tools and the “consumers” of its outputs. From a technical point of view, OMR can be considered a subfield of computer vision and document analysis, with deep learning acting as a catalyst that opens up promising novel approaches. Within the context of Music Information Retrieval (MIR), OMR should enable the application of MIR algorithms that rely on symbolic data and audio inputs (through rendering the recognized scores). It furthermore can enrich digital music score libraries and make them much more searchable and accessible, which broadens the scope of digital musicology to compositions for which we only have the written score (which is probably the majority of Western musical heritage). Finally, OMR has practical implications for composers, conductors, and the performers themselves, as it cuts down the costs of digitizing scores, and therefore bring the benefits of digital formats to their everyday practice.



Fig. 6: How the translation of the graphical concept of a note into a pitch is affected by the clef and accidentals. The effective pitch is written above each note. Accidentals immediately before a note propagate to other notes within the same measure, but not to the next measure. Accidentals at the beginning of a measure indicate a new key signature that affects all subsequent notes.

### A. Optical Music Recognition vs. Text Recognition

One must also address the obvious question: why should OMR be singled out besides Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR), given that they are tightly linked [18], and OMR has frequently been called “OCR for music” [129], [25], [94], [93], [80], [130], [109], [68], [26], [149]?<sup>8</sup> What is the justification of talking specifically about music notation and what differentiates it from other graphics recognition challenges? What are the special considerations in OMR that one does not encounter in other writing systems?

A part of the justification lies in the properties of music notation as a *contextual* writing system. While its alphabet consists of well-defined primitives (e.g., stems, noteheads, or flags) that have a clear interpretation, it is only in their configuration—how they are placed and arranged on the staves, and with respect to each other—that specifies what notes should be played. The properties of music notation that make it a challenge for computational reading have been discussed exhaustively by Byrd and Simonsen [29]; we hypothesize that these difficulties are ultimately caused by this contextual nature of music notation.

Another major reason for considering the field of OMR distinct from text recognition is the application domain itself—music. When processing a document of music notation, there is a natural requirement to recover its musical semantics as well (see Section IV, setting B), as opposed to text recognition, which typically does not have to go beyond recognizing letters or words and ordering them correctly. There is no proper equivalent of this interpretation step in text recognition since there is no definite answer to *how a symbol configuration (=words) should be further interpreted*; therefore, one generally leaves interpretation to humans or to other well-defined tasks from the Natural Language Processing field. However, given that music is overwhelmingly often conceptualized as notes, and notes are well-defined objects that can be inferred from the score, OMR is, not unreasonably, asked to produce this additional level of outputs that text recognition does not. Perhaps the simplest example to illustrate this difference is given by the concept of the pitch of the notes (see Fig. 6). While graphically a note lies on a specific vertical position of the staff, other objects, such as the clefs and accidentals, determine its musical pitch. It is therefore insufficient for the OMR to provide just the results in terms of positions, but it also has to take the context into account, in order to convert positions (graphical concept) into pitches (musical concept). To make matters worse, the correct interpretation depends on conventions, which can change depending on social, historical, or professional context (what a jazz musician accepts as conventional will differ from that understood by a classical player, for example). In this regard, OMR is more ambitious than text recognition, since there is an additional interpretation step specifically for music that has no good analogy in other natural languages.

The character set poses another significant challenge, compared to text recognition. Although writing systems like Chinese have extraordinarily complex character sets, the primitives in music notation have a much greater variability in size, ranging from small elements like a *dot* to big elements spanning an entire page like the *brace*. Many of the primitives may appear at various scales and rotations like *beams* or have a nearly unrestricted appearance like *slurs* that are only defined as more-or-less smooth curves that may be interrupted anywhere. Finally, in contrast to text recognition, music notation involves ubiquitous

<sup>8</sup>Even the English Wikipedia article on OMR has been calling it “Music OCR” for over 13 years.



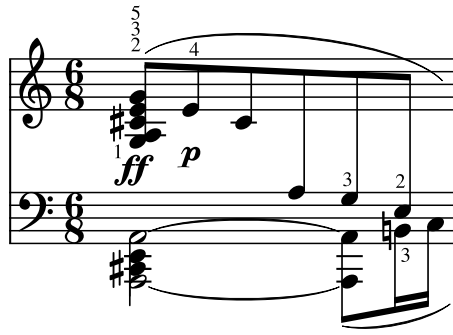


Fig. 7: This excerpt by Ludwig van Beethoven, Piano Sonata op. 2 no. 2, Largo appassionato, m. 31 illustrates some properties of the music notation that distinguish it from other types of writing systems: a wide range of primitive sizes, the same primitives appearing at different scales and rotations, and the ubiquitous two-dimensional spatial relationships.

two-dimensional spatial relationships, which are salient for the symbols' interpretation. Some of these properties are illustrated in Fig. 7.

Furthermore, Byrd and Simonsen [29] argue that because of the vague limits of what one may want to express using music notation, its syntactic rules can be expected to be bent accordingly; this happens to such an extent that Homenda et al. [90] argued that there is no universal definition of music notation at all. Figure 7 actually contains an instance of such rule-breaking: while one would expect all notes in one chord to share the same duration, the chord on the bottom left contains a mix of white and black noteheads, corresponding to half- and quarter-notes. At the same time, however, the musical intent is yet another: the two quarter-notes in the middle of the chord are actually played as eighth notes, to add to the rich sonority of the fortissimo chord on the first beat.<sup>9</sup> We believe this example succinctly illustrates the intricacies of the relationship between musical comprehension and music notation. This last difference between a written quarter and interpreted eighth note is, however, beyond what one may expect OMR to do, but it serves as further evidence that the domain of music presents its own difficulties, compared to the domains where text recognition normally operates.

### B. Optical Music Recognition vs. Other Graphics Recognition Challenges

Apart from text, documents can contain a wide range of other graphical information, such as engineering drawings, floor plans, mathematical expressions, comics, maps, patents, diagrams, charts or tables [44], [58]. Recognizing any of these comes with its own set of challenges, e.g., comics combine text and other visual information in order to narrate a story, which makes recovering the correct reading order a non-trivial endeavor. Similarly, the arrangement of symbols in engineering drawing and floor plans can be very complex with rather arbitrary shapes. Even tasks that are seemingly easy, such as the recognition of tables, must not be underestimated and are still subject to ongoing research [146], [132]. The hardest aspects of OMR are much closer to these challenges than to text recognition: the ubiquitous two-dimensionality, long-distance spatial relationships, and the permissive way of how individual elements can be arranged and appear at different scales and rotations.

One thing that makes CWMN more complex than many graphics recognition challenges like mathematical formulae recognition is the complex typographical alignment of objects [7], [29] that is dictated by the content, e.g., simultaneous events should have the same horizontal position, or each space between multiple notes of the same length should be equal. Interactions between individual objects can force other elements to move around, potentially breaking those principles (see Fig. 8, 9 and 10).

<sup>9</sup>This effect would be especially prominent on the Hammerklavier instruments prevalent around the time Beethoven was composing this sonata.

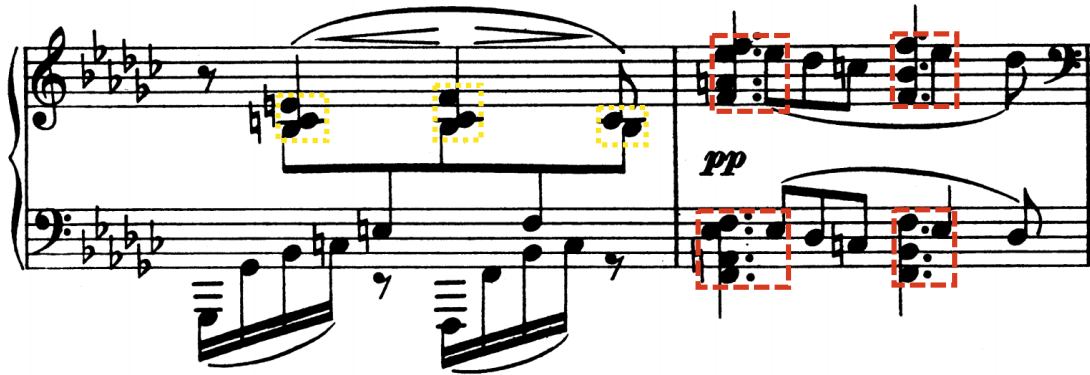


Fig. 8: Brahms Intermezzo, Op. 117 no. 1. Adjacent notes of the chords in the first bar in the top staff are shifted to the right to avoid overlappings (yellow dotted boxes). The moving eighths in the second bar are forced even further to the right, although being played simultaneously with the chord (red dashed boxes).

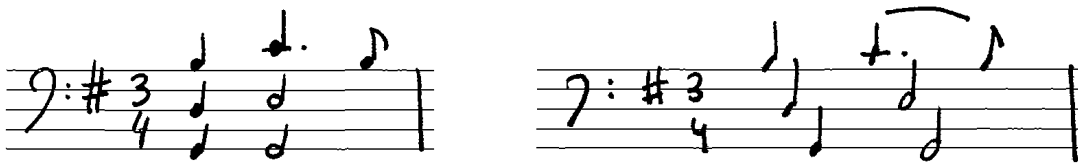


Fig. 9: Sample from the CVC-MUSCIMA dataset [60] with the same bar transcribed by two different writers. The first three notes and the second three notes form a chord and should be played simultaneously (left figure) but sometimes scores are written in a poor musical orthography (right figure) which can create the impression that the notes should be played after each other instead of together.

Apart from the typographical challenges, music notation also has extremely complex semantics, with many implicit rules. To handle this complexity, researchers started a long time ago to leverage the rules that govern music notation and formulate them into grammars [124], [4]. For instance, the fact that the note durations (in each notated voice) have to sum up to the length of a measure has been integrated into OMR as a post-processing step [120]. Fujinaga [67] even states that music notation can be recognized by an LL(k) grammar. Nevertheless, the following citation from Blostein and Baird [22] (p.425) is still mostly true:

“Various methods have been suggested for extending grammatical methods which were developed for one-dimensional languages. While many authors suggest using grammars for music notation, their ideas are only illustrated by small grammars that capture a tiny subset of music notation.” [22] (p.425; sec. 7 - Syntactic Methods).

There has been progress on enlarging the subset of music notation captured by these grammars, most notably in the DMOS system [49], but there are still no tractable 2-D parsing algorithms that are powerful enough for recognizing music notation without relying on fragile segmentation heuristics. It is not clear whether current parsers used to recognize mathematical expressions [3] are applicable to music notation or simply have not been applied yet—at least we are not aware of any such works.

## VI. A TAXONOMY OF OMR

Now that we have progressed in our effort to define *Optical Music Recognition*, we can turn our attention to systematizing the field with respect to motivating applications, subtasks, and their interfaces. We reiterate that our objective is not to review the methods by which others have attempted to reach the goals of their OMR work; rather, we are proposing a taxonomy of the field’s goals themselves. Our



Fig. 10: Sample from the Songbook of Romeo & Julia by Gerard Presgurvic [125] with uneven spacing between multiple sixteenth notes in the middle voice to align them with the lyrics.

motivation is to find natural groups of OMR applications and tasks for which we can expect, among other things, shared evaluation protocols. The need for such systematization has long been felt [23], [30], but subsequent reviews [133], [111] have focused almost entirely on technical solutions.

#### A. OMR Inputs

The taxonomy of *inputs* of OMR systems is generally established. The first fundamental difference can be drawn between *offline* and *online* OMR:<sup>10</sup> offline OMR operates on a static image, while online OMR processes music notation as it is being written, e.g., by capturing the stylus input on an electronic tablet device [72], [73], [152], [31]. Online OMR is generally considered easier since the decomposition into strokes provides a high-quality over-segmentation essentially for free. Offline OMR can be further subdivided by the engraving mechanism that has been used, which can be either *typeset* by a machine, often inaccurately referred to as *printed*,<sup>11</sup> or *handwritten* by a human, with an intermediate, yet common scenario of handwritten notation on pre-printed staff paper. Throughout the rest of the paper, we mainly elaborate on the case of offline OMR, because it represents the most common and considered scenario.

Importantly, music can be written down in many different notation systems that can be seen as different languages to express musical concepts (see Fig. 11). *CWMN* is probably the most prominent one. Before *CWMN* was established, other notations such as mensural or neumes preceded it, so we refer to them as *historical notations*. Although this may seem like a tangential issue, the recognition of manuscripts in ancient notations has motivated a large number of works in OMR that facilitate the preservation and analysis of the cultural heritage as well as enabling digital musicological research of early music at scale [51], [160], [69], [50]. Another category of notations that are still being actively used today are *instrument-specific notations*, such as tablature for string instruments or percussion notation. The final category captures all *other notations* including, e.g., modern graphic notation, braille music or numbered notation that are only rarely used and for which the existing body of music is much smaller than for the other notations.

To get an idea of how diverse the visual expression of music is, the Standard Music Font Layout [150] currently lists over 2440 recommended characters, plus several hundred optional glyphs.

Byrd and Simonsen [29] further characterize OMR inputs by the *complexity* of the notated music itself, ranging from simple monophonic music to “pianoform.” They use both the presence of multiple staves as well as the number of notated voices inside a single staff as a dimension of notational complexity. In

<sup>10</sup>Although it might sound ambiguous, the term “online recognition” has been used systematically in the handwritten recognition community [123]. Sometimes, this scenario is also referred to as pen-based recognition.

<sup>11</sup>Handwritten manuscripts can also be printed out, if they were scanned previously, therefore we prefer the word typeset.

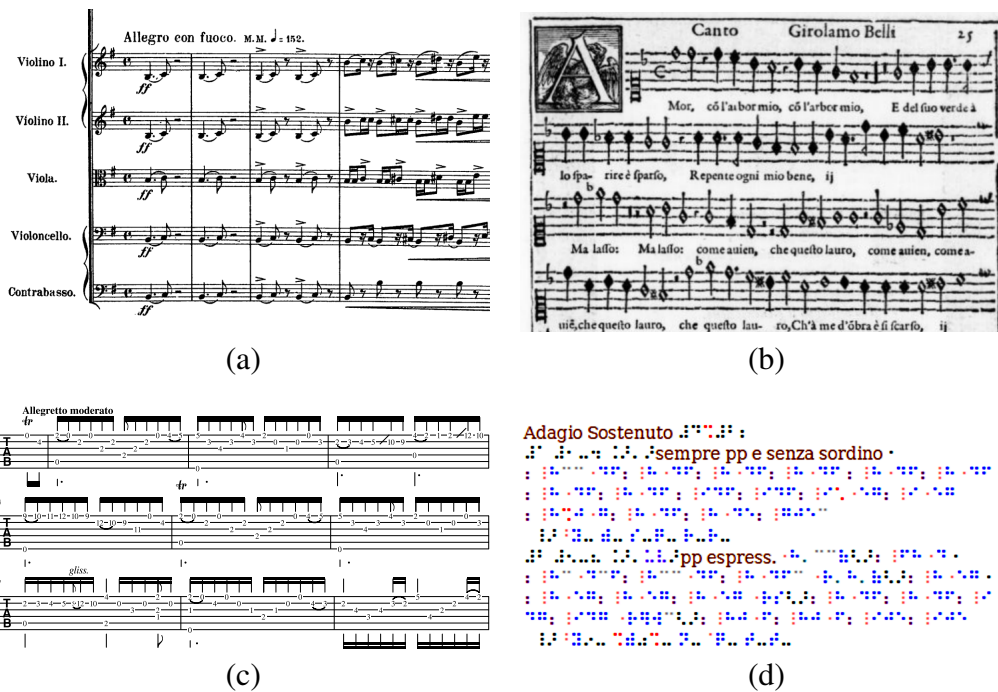


Fig. 11: Examples of scores written in various notations: (a) Common Western Music Notation (Dvorak Symphony No.9, IV), (b) White Mensural Notation (Belli [121]), (c) Tabulature (Regondi, Etude No.10) and (d) Braille (Beethoven, Sonata No.14 Op.27 No.2).

contrast, we do not see the number of staves as a driver of complexity since a page typically contains many staves and a decision on how to group them into systems has to be made anyway. Additionally, we explicitly add a category for *homophonic* music that only has a single logical voice, even though that voice may contain chords with multiple notes being played simultaneously. The reason for singling out homophonic music is that once notes are grouped into chords, inferring onsets can dramatically be simplified into simply reading them left-to-right, as opposed to polyphonic music which requires the assignment of multiple logical voices first.

Therefore, we adapt Byrd and Simonsen’s taxonomy [29] into the following four categories of *structural complexity* (see Fig. 12):

- (a) *Monophonic*: only one note (per staff) is played at a time.
- (b) *Homophonic*: multiple notes can occur at the same time to build up a chord, but only as a single voice.
- (c) *Polyphonic*: multiple voices can appear in a single staff.
- (d) *Pianoform*: scores with multiple staves and multiple voices that exhibit significant structural interactions. They can be much more complex than polyphonic scores and cannot be disassembled into a series of monophonic scores, such as in polyphonic renaissance vocal part books.

This complexity of the encoded music has significant implications on the whole OMR system because the various levels translate into different sets of constraints and requirements. Furthermore, the structural complexity cannot simply be adjusted or simulated like the visual complexity by applying an image operation on a perfect image [95] because it represents an intrinsic property of the music.

Finally, as with other digital document processing, OMR inputs can be classified according to their image quality which is determined by two independent factors: the underlying *document quality*, and the *digital imaging acquisition* mode. The underlying document quality is a continuum on a scale from perfect or nearly flawless (e.g., if the document was born-digital and printed) to heavily degraded or defaced documents (e.g., ancient manuscripts that deteriorated over time and exhibit faded ink, ink blots, stains, or bleedthrough) [29]. The image acquisition mode is also a continuum that can extend from

(a) Monophonic: A single melodic line in 4/4 time, featuring eighth and sixteenth notes, with triplets in the final measure.

(b) Homophonic: A single melodic line with block chords in 4/4 time, featuring eighth and sixteenth notes, with a triplet in the final measure.

(c) Polyphonic: Two independent melodic lines in 4/4 time, featuring eighth and sixteenth notes, with triplets in the final measure.

(d) Pianoform: A two-staff piece in 4/4 time, featuring a complex interplay of melodic and harmonic lines, with triplets in the final measure.

Fig. 12: Examples of the four categories of structural complexity.

born-digital images, over scans of varying quality to low-quality, distorted photos that originate from camera-based scenarios with handheld cameras, such as smartphones [2], [162].

### B. OMR Outputs

The taxonomy of OMR *outputs*, on the other hand, has not been treated as systematically in the OMR literature. Lists of potential or hypothetical applications are typically given in introductory sections [67], [22], [38], [111]. While this may not seem like a serious issue, it makes it hard to categorize different works and compare their results with each other because one often ends up comparing apples to oranges [7].

The need for a more principled treatment is probably best illustrated by the unsatisfactory state of OMR evaluation. As pointed out by [29], [84], [81], there is still no good way at the moment of how to measure and compare the performance of OMR systems. The lack of such evaluation methods is best illustrated by the way OMR literature presents the state of the field: Some consider it a mature area that works well (at least for typeset music) [61], [62], [5], [12], [135]. Others describe their systems with reports of very high accuracies of up to nearly 100% [147], [99], [163], [91], [104], [110], [122], [162], [33], giving an impression of success; however, many of these numbers are symbol detection scores on a small corpus with a limited vocabulary that are not straightforward to interpret in terms of actual usefulness, since

they do not generalize [19], [29].<sup>12</sup> The existence of commercial applications [106], [108], [107], [151], [112], [71], [131] is also sometimes used to support the claim that OMR “works” [13]. On the other hand, many researchers think otherwise [28], [19], [133], [109], [40], [46], [118], [134], [83], [82], emphasizing that OMR does not provide satisfactory solutions in general—not even for typeset music. Some indirect evidence of this can be gleaned from the fact that even for high-quality scans of typeset music, only a few projects rely on OMR,<sup>13</sup> while other projects still prefer to crowdsource manual transcriptions instead of using systems for automatic recognition [78], or at least crowdsource the correction of the errors produced by OMR systems [142]. Given the long-standing absence of OMR evaluation standards, this ambivalence is not surprising. However, a scientific field should be able to communicate its results in comprehensible terms to external stakeholders—something OMR is currently unable to do.

We feel that to a great extent this confusion stems from the fact that the question “Does OMR work?” is an overly vague question. As our analysis in Section II shows, OMR is not a monolithic problem—therefore, asking about the “state of OMR” is *under-specified*. “Does OMR work?” must be followed by “... as a tool for X,” where X is some application, in order for such questions to be answerable. There is, again, evidence for this in the OMR literature. OMR systems have been properly evaluated in retrieval scenarios [10], [66], [1] or in the context of digitally replicating a musicological study [83]. It has, in fact, been explicitly asserted [81] that evaluation methodologies are only missing for a **limited** subset of OMR applications. In particular, there is no known meaningful edit distance between two scores, i.e., an algorithm that computes the similarity between two scores and produces the steps that are needed to transform one score into the other. If such a distance function existed, one could at least count the number of required edits to assess the performance of an OMR system quantitatively. However, this does not necessarily provide a good measure of quality because it is unclear how to weight the costs of different edit operations, e.g., getting a note duration wrong vs. missing an articulation mark.

At the same time, the granularity at which we define the various tasks should not be too fine, otherwise one risks entering a different swamp: instead of no evaluation at all, each individual work is evaluated on the merits of a narrowly defined (and often merely hypothetical) application scenario, which also leads to contributions that cannot be compared. In fact, this risk has already been illustrated on the subtask of symbol detection, which seems like a well-defined problem where the comparison should be trivial. In 2018, multiple music notation object detection papers were published [117], [116], [82], [154], but each reported results in a different way while presenting a good argument for choosing that kind of evaluation, so significant effort was necessary in order to make these contributions directly comparable [119]. A compromise is therefore necessary between fully specifying the question of whether OMR “works” by asking for a specific application scenario, and on the other hand retaining sufficiently general categories of such tasks.

Having put forward the reasoning why systematizing the field of OMR with respect to its outputs is desirable, we proceed to do so. For defining meaningful categories of outputs for OMR, we come back to the fundamentals of how OMR inverts the music encoding process to recover the musical semantics and musical notation (see Section II). These two prongs of reading musical documents roughly correspond to two broad areas of OMR applications [105] that overlap to a certain extent:

- *Replayability*: recovering the encoded music itself in terms of pitch, velocity, onset, and duration. This application area sees OMR as a component inside a bigger music processing pipeline that enables the system to operate on music notation documents as just another input. Notice that readability by humans is not required for these applications, as long as the computer can process and “play” the symbolic data.
- *Structured Encoding*: recovering the music along with the information on how it was encoded using elements of music notation. This avenue is oriented towards providing the score for music

<sup>12</sup>The problem of results that cannot be compared was already noted in the very first review of OMR in 1972 by Kassler [96], when he reviewed the first two OMR theses by Pruslin [127] and Prerau [124].

<sup>13</sup>Some users of the Choral Public Domain Library (CPDL) project use commercial applications such as SharpEye or PhotoScore Ultimate: <http://forums.cpdll.org/phpBB3/viewtopic.php?f=9&t=9392>, but without recording the time spent in correcting and editing the output.



performance, which requires a (lossless) re-encoding of the score, assuming that humans read the OMR output directly. Recovering the musical semantics might not in fact be strictly necessary, but in practice, one often wishes to obtain that information too, in order to enable digitally manipulating the music in a way that would be easiest done with the semantics being recovered (e.g., transposing a part to make it suitable for another instrument).

In other words, the output of an application that targets replayability is typically processed by a machine, whereas humans usually demand the complete recognition of the structured encoding to allow for a readable output (see Fig. 2).

While the distinction between replayability and structured encoding is already useful, there are other reasons that make it interesting to read musical notation from a document. For example, to search for specific content or to draw paleographic conclusions about the document itself. Therefore, we need to broaden the scope of OMR to actually capture these applications. We realized that some use-cases require much less comprehension of the input and music notation than others. To account for this, we propose the following four categories that demand an increasing level of comprehension: *Document Metadata Extraction*, *Search*, *Replayability*, and *Structured Encoding* (see Fig. 13).



Fig. 13: Taxonomy of four categories of OMR applications that require an increasing level of comprehension, starting with metadata extraction where a minimal understanding might be sufficient, up to structured encoding that requires a complete understanding of music notation with all its intricacies.

Depending on the goal, applications differ quite drastically in terms of requirements—foremost in the choice of output representation. Furthermore, this taxonomy allows us to use different evaluation strategies.

1) *Document Metadata Extraction*: The first application area requires only a partial understanding of the entire document and attempts to answer specific questions about it. These can be very primitive ones, like whether a document contains music scores or not, but the questions can also be more elaborate, for example:

- In which period was the piece written?
- What notation was used?
- How many instruments are involved?
- Are two segments written by the same copyist?

All of the aforementioned tasks entail a different level of underlying computational complexity. However, we are not organizing applications according to their difficulty but instead by the type of answer they provide. In that sense, all of these tasks can be formulated as classification or regression problems, for which the output is either a discrete category or a continuous value, respectively.

**Definition 2.** Document metadata extraction refers to a class of Optical Music Recognition applications that answer questions about the music notation document.

The output representation for document metadata extraction tasks are scalar values or category labels. Again, this does not imply that extracting the target values is necessarily easy, but that the difficulties are not related to the output representation, as is the case for other uses.

Although this type of application has not been very popular in the OMR literature, there is some work that approaches this scenario. In [9] and [118] the authors describe systems that classify images according

to whether they depict music scores or not. While the former one used a basic computer vision approach with a Hough transform and run-length ratios, the latter uses a deep convolutional neural network. Such systems can come in handy if one has to automatically classify a very large number of documents [114]. Perhaps the most prominent application is identifying the writer of a document [63], [64], [77], [140] (which can be different from the composer). This task was one of the main motivations behind the construction of the CVC-MUSCIMA dataset [60] and was featured in the ICDAR 2011 Music Score Competition [59].

The document metadata extraction scenario has the advantage of its unequivocal evaluation protocols. Tasks are formulated as either classification or regression problems, and these have well-defined metrics such as accuracy, F-measure, or mean squared error.

2) *Search*: Nowadays we have access to millions of musical documents. Libraries and communities have taken considerable efforts to catalog and digitize music scores, by scanning them and freely providing users access to them, e.g., IMSLP [126], SLUB [141], DIAMM [20] or CPDL [113], to name a few. Here is a fast growing interest in automated methods which would allow users to search for relevant musical content inside these sources systematically. Unfortunately, searching for specific content often remains elusive because many projects only provide the images and manually entered metadata. We capture all applications that enable such lookups under the category *Search*. Examples of search questions could be:

- Do I have this piece of music in my library?
- On which page can I find this melody?
- Where does this sequence of notes (e.g., a theme) repeat itself?
- Was a melody copied from another composition?
- Find the same measure in different editions for comparing them.

**Definition 3.** Search refers to a class of Optical Music Recognition applications that, given a collection of sheet music and a musical query, compute the relevance of individual items of the collection with respect to the given query.

Applications from this class share a direct analogy with keyword spotting (KWS) in the text domain [74] and a common formulation: the input is a query as well as the collection of documents on which the search is to be carried out; the output is the selection of elements from that collection that match the query. However, that selection of elements is a loose concept and can refer to a complete music piece, a page, or in the most specific cases, a particular bounding-box or even a pixel-level location. In the context of OMR, the musical query must convey musical semantics (as opposed to general search queries, e.g., by title or composer; hence the term “musical” query in Definition 3). The musical query is typically represented in a symbolic way, interpretable unambiguously by the computer (similar to query-by-string in KWS), yet it is also interesting to consider queries that involve other modalities, such as image queries (query-by-example in KWS) or audio queries (query-by-humming in audio information retrieval or query-by-speech in KWS). Additionally, it makes sense to establish different domain-specific types of matching, as it is useful to perform searches restricted to specific music concepts such as melodies, sequences of intervals, or contours, in addition to exact matching.

A direct approach for search within music collections is to use OMR technology to transform the documents into symbolic pieces of information, over which classical content-based or symbolic retrieval methods can be used [47], [14], [55], [153], [88], [97], [1], [52]. The problem is that these transformations require a more comprehensive understanding of the processed documents (see Sections VI-B3 and VI-B4 below). To avoid the need for an accurate symbol-by-symbol transcription, search applications can resort to other methods to determine whether (or how likely) a given query is in a document or not. For instance, in cross-modal settings, where one searches a database of sheet music using a MIDI file [66], [10] or a melodic fragment that is given by the user on the fly [1], OMR can be used as a hash function. When the queries and documents are both projected into the search space by the same OMR system, some limitations of the system may even cancel out (e.g., ignoring key signatures), so that retrieval performance might



deteriorate less than one would expect. Unfortunately, if either the query or the database contains the true musical semantics, such errors do become critical [83].

A few more works have also approached the direct search of music content without the need to convert the documents into a symbolic format first. Examples comprise the works by Malik et al. [100] dealing with a query-by-example task in the CVC-MUSCIMA dataset, and by Calvo-Zaragoza et al. [35], considering a classical query-by-string formulation over early handwritten scores. In the cross-modal setting, the audio-sheet music retrieval contributions of Dorfer et al. [54] are an example of a system that explicitly attempts to gain only the minimum level of comprehension of music notation necessary for performing its retrieval job.

Search systems usually retrieve not just a single result but all those that match the input query, typically sorted by confidence. This setting can re-use general information retrieval methodologies for evaluating performance, such as precision and recall as well as encompassing metrics like average precision and mean average precision [101], [87].

3) *Replayability*: Replayability applications are concerned with reconstructing the notes encoded in the music notation document. Notice that producing an actual audio file is not considered to be part of OMR, despite being one of the most frequent use-cases of OMR. In any case, OMR can enable these applications by recovering the pitches, velocities, onsets, and durations of notes. This symbolic representation, usually stored as a MIDI file, is already a very useful abstraction of the music itself and allows for plugging in a vast range of computational tools such as:

- synthesis software to produce an audio representation of the composition
- music information retrieval tools that operate on symbolic data
- tools that perform large-scale music-theoretical analysis
- creativity-focused applications [164]

**Definition 4.** Replayability refers to a class of Optical Music Recognition applications that recover sufficient information to create an audible version of the written music.

Producing a MIDI (or an equivalent) representation is one key goal for OMR—at least for the foreseeable future since MIDI is a representation of music that has a long tradition of computational processing for a vast variety of purposes. Many applications have been envisioned that only require replayability. For example, applications that can sight-read the scores to assist practicing musicians by providing missing accompaniment.

Replayability is also a major concern for digital musicology. Historically, the majority of compositions have probably never been recorded, and therefore are only available in written form as scores; of these, most compositions have also never been typeset, since typesetting has been a very expensive endeavor, reserved essentially either for works with assured commercial success, or composers with substantial backing by wealthy patrons. Given the price of manual transcription, it is prohibitive to transcribe large historical archives. At the present time, OMR that produces MIDI, especially if it can do so for manuscripts, is probably the only tool that could open up the vast amount of compositions to quantitative musicological research, which, in turn, could perhaps finally begin to answer broad questions about the evolutions of the average musical styles, instead of just relying on the works of the relatively few well-known composers.

Systems designed for the goal of replayability traditionally seek first to obtain the structured encoding of the score (see Section VI-B4), from which the sequences of notes can be straightforwardly retrieved [82]. However, if the specific goal is to obtain something equivalent to a MIDI representation, it is possible to simplify the recognition and ignore many of the elements of musical notation, as demonstrated by numerous research projects [102], [16], [90], [65], [139], [91], [116]. An even clearer example of this distinction can be observed in the works of Shi et al. [148] as well as van der Wel and Ullrich [159]; both focus only on obtaining the sequence of note pairs (duration, pitch) that are depicted in single-staff images, regardless of how these notes were actually expressed in the document. Another instance of a replay-oriented application is the Gocen system [5] that reads handwritten notes with a specially designed

device with the goal of producing a musical performance while ignoring the majority of music notation syntax.

Once a system is able to arrive at a MIDI-like representation, evaluating the results is a matter of comparing sequences of pitch-onset-duration-triplets. Velocities may optionally be compared too, once the note-by-note correspondence has been established, but can be seen as secondary for many applications. Note, however, that even on the level of describing music as configurations of pitch-velocity-onset-duration-quadruples, MIDI is a further simplification that is heavily influenced by its origin as a digital representation of performance, rather than a composition: the most obvious inadequacy of MIDI is its inability to distinguish pitches that sound equivalent but are named differently, e.g., F-sharp and G-flat.<sup>14</sup>

Multiple similarity metrics for comparing MIDI files have been proposed during the Symbolic Melodic Similarity track of the Music Information Retrieval Evaluation eXchange (MIREX),<sup>15</sup> e.g., by determining the local alignment between the geometric representations of the melodies [156], [157], [158], [155]. Other options could be multi-pitch estimation evaluation metrics [17], Dynamic Time Warping [54], or edit distances between two time-ordered sequences of pitch-duration pairs [165], [33].

4) *Structured Encoding*: It can be reasonably stated that digitizing music scores for “human consumption” and score manipulation tasks that a *vollkommener Capellmeister*<sup>16</sup> [103] routinely performs, such as part exporting, merging, or transposing for available instruments is the original motivation of OMR ever since it started [127], [124], [67], [6] and the one that appeals to the widest audience. Given that typesetting music is troublesome and time-consuming, OMR technology represents an attractive alternative to obtain a digital version of music scores on which these operations can be performed efficiently with the assistance of the computer.

This brings us to our last category that requires the highest level of comprehension, called structured encoding. Structured encoding aims to recognize the entire music score while retaining all the engraving information available to a human reader. Since there is no viable alternative to music notation, the system has to fully transcribe the document into a structured digital format with the ultimate goal of keeping the same musical information that could be retrieved from the physical score itself.

**Definition 5.** Structured Encoding refers to a class of Optical Music Recognition applications that fully decode the musical content, along with the information of ‘how’ it was encoded by means of music notation.

Note that the difference between replayability and structured encoding can seem vague: for instance, imagine a system that detects all notes and all other symbols and exports them into a MusicXML file. The result, however, is not the structured encoding unless the system also attempts to preserve the information on how the scores were laid out. That does not mean it has to store the bounding box and exact location of every single symbol, but the engraving information that *conveys musical semantics*, like whether the stem of a note went up or down. To illustrate this, consider the following musical snippet in Fig. 14. If a system like the one described in [33] recognized this, it would remain restricted to replayability. Not because of the current limitations to monophonic, single-staff music, but due to the selected output representation, which does not store engraving information such as the simplifications that start in the second measure of the top staff (the grayed out 3s that would be omitted in the printing) or the stem directions of the notes in the bottom staff (green and blue) that depict two different voices. In summary, any system discarding engraving information that conveys musical semantics, by definition, cannot reach the structured encoding goal.

To help understand why structured encoding poses such a difficult challenge, we would like to avail ourselves of the intuitive comparison given by Donald Byrd:<sup>17</sup> representing music as time-stamped events

<sup>14</sup>This is the combined heritage of equal temperament, where these two pitches do correspond to the same fundamental frequency, and of the origins of MIDI in genres dominated by fretted and keyboard instruments.

<sup>15</sup> [https://www.music-ir.org/mirex/wiki/MIREX\\_HOME](https://www.music-ir.org/mirex/wiki/MIREX_HOME)

<sup>16</sup>roughly translated from German as “ideal conductor”

<sup>17</sup>[http://music.informatics.indiana.edu/don\\_notation.html](http://music.informatics.indiana.edu/don_notation.html)

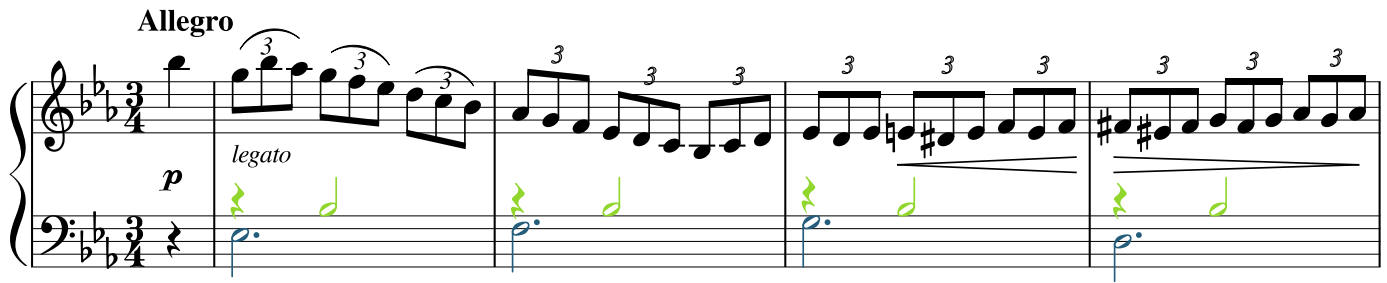


Fig. 14: Beginning of Franz Schubert, Impromptu D.899 No. 2. The triplet marks starting in the second measure of the top staff are typically omitted in printed editions (here depicted in gray for visualization). The two distinct voices in the bottom staff are color-coded in green and blue.

(e.g., with MIDI) is similar to storing a piece of writing in a plain text file; whereas representing music with music notation (e.g., with MusicXML) is similar to a structured description like an HTML website. By analogy, obtaining the structured encoding from the image of a music score can be as challenging as recovering the HTML source code from the screenshot of a website.

Since this use-case appeals to the widest audience, it has seen development both from the scientific research community and commercial vendors. Notable products that attempt full structured encoding include SmartScore [106], Capella Scan [37], PhotoScore [108] as well as the open-source application Audiveris [21]. While the projects described in many scientific publications seem to be striving for structured encoding to enable interesting applications such as the preservation of the cultural heritage [39], music renotation [41], or transcriptions between different music notation languages [136], we are not aware of any systems in academia that would actually produce structured encoding.

A major stumbling block for structured encoding applications has for a long time been the lack of practical formats for representing music notation that would be powerful enough to retain the information from the input score, and at the same time be a natural endpoint for OMR. This is illustrated by papers that propose OMR-specific representations, both before the emergence of MusicXML [75], [76] as a viable interchange format [105] and after [86]. At the same time, however, even without regard for OMR, there are ongoing efforts to improve music notation file formats: further development of MusicXML has moved into the W3C Music Notation Community Group,<sup>18</sup> and there is an ongoing effort in the development of the Music Encoding Initiative format [138], best illustrated by the annual Music Encoding Conference.<sup>19</sup> Supporting the whole spectrum of music notation situations that arise in a reasonably-sized archive is already a difficult task. This can be evidenced by the extensive catalog of requirements for music notation formats that Byrd and Isaacson [27] list for a multi-purpose digital archive of music scores. Incidentally, the same paper also mentions support for syntactically incorrect scores among the requirements, which is one of the major problems that OMR has with outputting to existing formats directly. Although these formats are becoming more precise and descriptive, they are not designed to store information about how the content was automatically recognized from the document. This kind of information is actually relevant for evaluation, as it allows, for example, determining if a pitch was misclassified because of either a wrongly detected position in the staff or a wrongly detected clef.

The imperfections of representation standards for music notation is also reflected in a lack of evaluation standards for structured encoding. Given the ground truth representation of a score and the output of a recognition system, there is currently no automatic method that is capable of reliably computing how well the recognition system performed. Ideally, such a method would be rigorously described and evaluated, have a public implementation, and give meaningful results. Within the traditional OMR pipeline, the partial steps (such as symbol detection) can use rather general evaluation metrics. However, when OMR is applied for getting the structured encoding of the score, no evaluation metric is available, or at least generally

<sup>18</sup><https://www.w3.org/community/music-notation/>

<sup>19</sup><https://music-encoding.org/conference/past.html>

accepted, partially because of the lack of a standard representation for OMR output, as mentioned earlier. The notion of “edit cost” or “recognition gain” that defines success in terms of how much time a human editor saves by using an OMR system is yet more problematic, as it depends on the editor and on the specific toolchain [19].

Hajič [81] argues that a good OMR evaluation metric should be intrinsic<sup>20</sup> and independent of a certain use-case. The benefits would be the independence from the selected score editing toolchain as well as the music notation format and a clearly interpretable automatic metric for guiding OMR development (which could ideally be used as a differentiable loss function for training full-pipeline end-to-end machine learning-based systems).

There is no reason why a proper evaluation should not be possible since there is only a finite amount of information that a music document retains, which can be exhaustively enumerated. It follows that we should be able to measure what proportion of this information our systems recovered correctly. One reason why this is still such a hard problem is the lack of an underlying formal model of music notation for OMR. Such a model could support structured encoding evaluation by being:

- *Comprehensive*: integrating naturally both the “reprintability” and “replayability” level (also called graphical and semantic level in the literature), by being capable of describing the various corner cases (which implies extensibility);
- *Useful*: enabling tractable inference (at least approximate) and an adequate distance function; and
- *Sufficiently supported* through open-source software.

Historically, context-free grammars have been the most explored avenue for a unified formal description of music notation, both with an explicit grammar [4], [49] and implicitly using a modified stack automaton [8]: this feels natural, given that music notation has strict syntactic rules and hierarchical structures that invite such descriptions. The 2-D nature of music notation also inspired graph grammars [56] and attributed graph grammars [15]. Recently, modeling music notation as a directed acyclic graph has been proposed as an alternative [86], [82]. However, none of these formalisms has yet been adopted: the notation graph is too recent and does not have sufficient software and community support, and the older grammar-based approaches lack up-to-date open-source implementations altogether (and are insufficiently detailed in the respective publications for re-implementation). Without an appropriate formalism and the corresponding tooling, the evaluation of structured encoding can hardly hope to move beyond ad-hoc methods.

It is important to distinguish between the formal model of music notation on which OMR systems operate, and the output representation they generate. MusicXML and MEI are good choices as output format because of their widespread tool support and acceptance, but are not suitable as internal representation of OMR systems. For example, they describe music scores top-down, something like “page-staff-measure-note” with semantic terms such as notes that have a pitch and a duration whereas an OMR system typically detects music bottom-up, often starting with abstract graphical shapes at certain positions in the image that need to be stored and processed somehow before the actual semantics can be inferred. MusicXML and MEI were simply not designed to capture these low-level graphical concepts but the final, well-defined musical concepts. That is why almost all OMR systems have their own internal model before encoding the results into standard formats.

## VII. APPROACHES TO OMR

In order to complete our journey through the landscape of *Optical Music Recognition*, we yet have to visit the arena of OMR techniques. These have recently seen a strong shift towards machine learning that has brought about a need to revisit the way that OMR methods have traditionally been systematized. As opposed to OMR applications, the vocabulary of OMR methods and subtasks already exists [133] and only needs to be updated to reflect the new reality of the field.

<sup>20</sup>Extrinsic evaluation means evaluating the system in an application context: “How good is this system for application X?” Intrinsic evaluation attempts to evaluate a system without reference to a specific use-case, asking how much of the encoded information has been recovered. In the case of OMR, this essentially reduces evaluation to error counting.

As mentioned before, obtaining the structured encoding of the scores has been the main motivation for developing the OMR field. Given the difficulty of such objective, the process was usually approached by dividing it into smaller stages that could represent challenges within reach with the available technologies and resources. Over the years, the pipeline described by Bainbridge and Bell [7], refined by Rebelo et al. in 2012 [133] became the de-facto standard. That pipeline is traditionally organized into the following four blocks, sometimes with slightly varying names and scopes of the individual stages:

- 1) *Preprocessing*: Standard techniques to ease further steps, e.g., contrast enhancement, binarization, skew-correction or noise removal. Additionally, the layout should be analyzed to allow subsequent steps to focus on actual content and ignore the background.
- 2) *Music Object Detection*: Finding and classifying all relevant symbols or glyphs in the image.
- 3) *Notation Assembly*: Recovering the music notation semantics from the detected and classified symbols. The output is a symbolic representation of the symbols and their relationships, typically as a graph.
- 4) *Encoding*: Encoding the music into any output format unambiguously, e.g., into MIDI for playback or MusicXML/MEI for further editing in a music notation program.

With the appearance of deep learning in OMR, many steps that traditionally produced suboptimal results, such as the staff-line removal or symbol classification have seen drastic improvements [70], [118] and are nowadays considered solved or at least clearly solvable. This caused some steps to become obsolete or collapse into a single (bigger) stage. For instance, the music object detection stage was traditionally separated into a segmentation stage and a classification stage. Since staff lines make it hard to separate isolated symbols through connected component analysis, they typically were removed first, using a separate method. However, deep learning models with convolutional neural networks have been shown to be able to deal with the music object detection stage holistically without having to remove staff lines at all. In addition to the performance gains, a compelling advantage is the capability of these models to be trained in a single step by merely providing pairs of images and positions of the music objects to be found, eliminating the preprocessing step altogether. A baseline of competing approaches on several datasets containing both handwritten and typeset music can be found in the work of Pacha et al. [119].

The recent advances also diversified the way OMR is approached altogether: there are alternative pipelines with their own ongoing research that attempt to face the whole process in a single step. This holistic paradigm, also referred to as end-to-end systems, has been dominating the current state of the art in other tasks such as text, speech, or mathematical formula recognition [48], [45], [165]. However, due to the complexity of inferring musical semantics from the image, it is difficult (for now) to formulate it as a learnable optimization problem. While end-to-end systems for OMR do exist, they are still limited to a subset of music notation, at best. Pugin pioneered this approach utilizing hidden Markov models for the recognition of typeset mensural notation [128], and some recent works have considered deep recurrent neural networks for monophonic music written in both typeset [148], [159], [33], [32] and handwritten [13] modern notation. However, polyphonic and pianoform scores are currently out of reach because there is simply no appropriate model formulation. Another limitation of these approaches is that they do not allow inspecting intermediate results to identify errors.

Along with the shift towards machine learning, several public datasets have emerged, such as MUSCIMA++ [86], DeepScores [154] or Camera-PrIMuS [32].<sup>21</sup> There are also significant efforts to develop tools by which training data for OMR systems can be obtained including MUSCIMarker [85], Pixel.js [143], and MuRET [136]. On the other hand, while the machine learning paradigm has undeniably brought significant progress, it has shifted the costs onto data acquisition. This means that while machine learning can be more general and deliver state-of-the-art results when appropriate data is available, it does not necessarily drive down the costs of applying OMR. Still, we would say—tentatively—that once these resources are spent, the chances of OMR yielding useful results for the specific use-case are higher compared to earlier paradigms.

<sup>21</sup>A full list of all available datasets can be found at <https://apacha.github.io/OMR-Datasets/>

Tangentially to the way of dealing with the process itself, there has been continuous research on interactive systems for years. The idea behind such systems is based on the insight that OMR systems might always make some errors, and if no errors can be tolerated, the user is essential to correct the output. These systems attempt to incorporate user feedback into the OMR process in a more efficient way than just post-processing system output. Most notably is the interactive system developed by Chen et al. [42], [43], where users directly interact with the OMR system by specifying which constraints to take into account while visually recognizing the scores. Users can then iteratively add or remove constraints before re-recognizing individual measures until they are satisfied. The most powerful feature of interactive systems is probably the displaying of recognition results superimposed on top of the original image, allowing errors to be spotted quickly [37], [21], [161], [136].

## VIII. CONCLUSIONS

In this article, we have first addressed what *Optical Music Recognition* is and proposed to define it as research field that investigates how to computationally read music notation in documents—a definition that should adequately delimit the field, and set it in relation to other fields such as OCR, graphics recognition, computer vision, and fields that await OMR results. We furthermore analyzed in depth the inverse relation of OMR to the process of writing down a musical composition and highlighted the relevance of engraving music properly—something that must also be recognized to ensure readability for humans. The investigation of what OMR is, revealed why this seemingly easy task of reading music notation has turned out to be such a hard problem: besides the technical difficulties associated with document analysis, many fundamental challenges arise from the way how music is expressed and captured in music notation. By providing a sound, concise, and inclusive definition, we capture how the field sees and talks about itself.

We have then reviewed and improved the taxonomy of OMR, which should help systematize the current and future contributions to the field. While the inputs of OMR systems have been described systematically and established throughout the field, a taxonomy of OMR outputs and applications has not been proposed before. An overview of this taxonomy is given in Fig. 15.

Finally, we have also updated the general breakdown of OMR systems into separate subtasks in order to reflect the recent advances with machine learning methods and discussed alternative approaches such as end-to-end systems and interactive scenarios.

One of the key points we wanted to stress is the internal diversity of the field: OMR is not a monolithic task. As analyzed in Section IV, it enables various use-cases that require fundamentally different system designs, as discussed in Section VI-B. So before creating an OMR system, one should be clear about the goals and the associated challenges.

The sensitivity to errors is another relevant issue that needs to be taken into account. As long as errors are inevitable [43], [50], it is important to consider the impact of those errors to the envisioned application. If someone wants to transcribe a score with an OMR system, but the effort needed for correcting the errors is greater than the effort for directly entering the notes into a music notation program, such an OMR system would obviously be useless [19]. Existing literature on error-tolerance is inconclusive: while we tend to believe that users—especially practicing musicians—would not tolerate false recognitions [137], we also see systems that can handle a substantial amount of errors [50], [1], [83] and still produce meaningful results, e.g., when searching in a large database of scores. Therefore, it cannot be decided in advance how severe errors are, as it is always the end user who sets the extent of tolerable errors.

Readers should now comprehend the spectrum of what OMR might do, understand the challenges that reading music notation entails, and have a solid basis for further exploring the field on their own—in other words, be equipped to address the issues described in the next section.

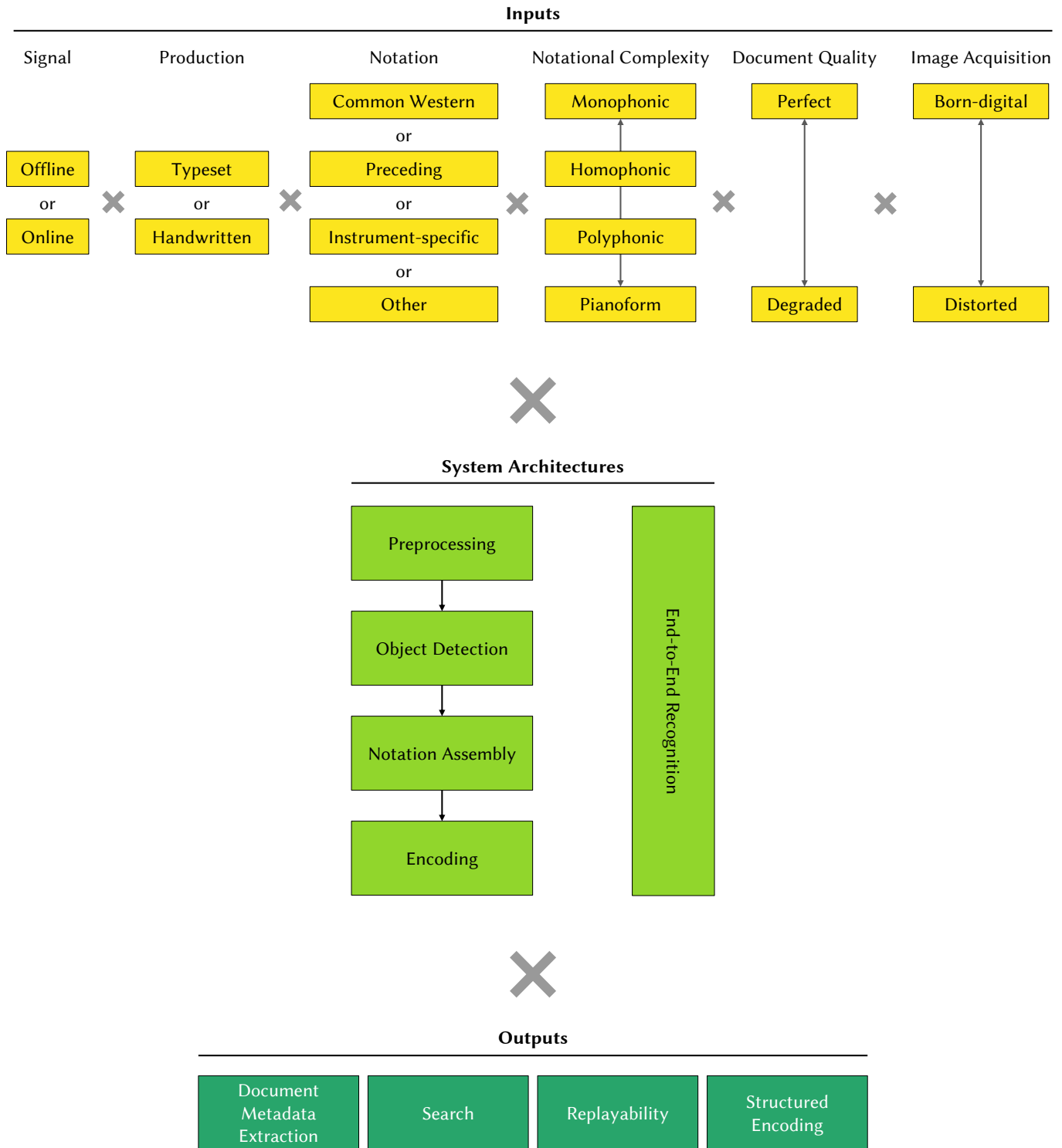


Fig. 15: An overview of the taxonomy of OMR inputs, architectures, and outputs. A fairly simple OMR system could, for example, read high-quality scans (offline) of well-preserved documents that contain typeset, monophonic, mensural notation, process it in a traditional, modular pipeline and write the results into a MIDI file to achieve replayability. A more sophisticated system, on the other hand, would allow images (offline) of handwritten music in common western notation from degraded documents as input and strive to recognize the full structured encoding in an end-to-end fashion. Note that the distinction between modular systems and end-to-end recognition is not binary as there can be hybrid systems that only merge a subset of the stages.

### A. Open Issues and Perspectives for Future Research

We conclude this paper by listing major open problems in Optical Music Recognition that significantly impede its progress and usefulness. While some of them are technical challenges, there are also many non-technical issues:

- *Legal aspects*: Written music is the intellectual property of the composer and its allowed uses are defined by the respective publisher. Recognizing and sharing music scores can be seen as copyright infringement, like digitizing books without permission. To avoid this dispute, many databases such as IMSLP only store music scores whose copyright protection has expired. So an OMR dataset is either limited to old scores or one enters a legal gray area if not paying close attention to the respective license of every piece stored therein.
- *Stable community*: For decades, OMR research was conducted by just a few individuals whose work was distributed and mostly uncoordinated. Most OMR researchers joined the field with minor contributions but left again soon afterward. Furthermore, due to a lack of dedicated venues, researchers rarely met in person [30]. This unstable setting and the fact that researchers were not paying sufficient attention to reproducibility led to the same problems being solved over and over again [115].
- *Lack of standards representations*: There exist no standard representation formats for OMR outputs, especially not for structured encoding, and virtually every system comes with its own internal representation and output format, even for intermediate steps. This causes incompatibilities between different systems and makes it very hard to replace subcomponents. Work on underlying formalisms for describing music notation can also potentially have a wide impact, especially if done in collaboration with the relevant communities (W3C Community Group on Music Notation, Music Encoding Initiative).
- *Evaluation*: Due to the lack of standards for outputting OMR results, evaluating them is currently in an equally unsatisfactory state. An ideal evaluation method would be rigorously described and verified, have a public implementation, give meaningful results, and not rely on a particular use-case, thus only intrinsically evaluating the system [81].

On the technical side, there are also many interesting avenues, where future research is needed, including:

- *Music Object Detection*: recent work has shown that the music object detection stage can be addressed in one step with deep neural networks. However, the accuracy is still far from optimal, which is especially detrimental to the following stages of the pipeline that are based on these results. In order to improve the detection performance, it might be interesting to develop models that are specific to the type of inputs that OMR works on: large images with a high quantity of densely packed objects of various sizes from a vast vocabulary.
- *Semantic reconstruction*: merely detecting the music objects in the document does not represent a complete music notation recognition system, and so the music object detection stage must be complemented with the semantic reconstruction. Traditionally, this stage is addressed by hand-crafted heuristics that either hardly generalize or do not cover the full spectrum of music notation. Machine learning-based semantic reconstruction represents an unexplored line of research that deserves further consideration.
- *Structured encoding research*: despite being the main motivation for OMR in many cases, there is a lack of scientific research and open systems that actually pursue the objective of reconstructing the full structure encoding of the input.
- *Full end-to-end systems*: end-to-end systems are accountable for major advances in machine learning tasks such as text recognition, speech recognition, or machine translation. The state of the art of these fields is based on recurrent neural networks. For design reasons, these networks currently deal only with one-dimensional output sequences. This fits the aforementioned tasks quite naturally since their outputs are mainly composed of word sequences. However, its application for music notation—except



for simple monophonic scores—is not so straightforward, and it is unknown how to formulate an end-to-end learning process for the recognition of fully-fledged music notation in documents.

- *Statistical modeling*: most machine learning algorithms are based on statistical models that are able to provide a probability distribution over the set of possible recognition hypotheses. When it comes to recognizing, we are typically interested in the best hypothesis—the one that is proposed as an answer—forgetting the probability given to such hypothesis by the model. However, it could be interesting to be able to exploit this uncertainty. For example, in the standard decomposition of stages in OMR systems, the semantic reconstruction stage could benefit from having a set of hypotheses about the objects detected in the previous stage, instead of single proposals. Then, the semantic reconstruction algorithm could establish relationships that are more logical *a priori*, although the objects involved have a lower probability according to the object detector. These types of approaches have not been deeply explored in the OMR field. Statistical modeling could also be useful so that the system provides a measure of its confidence about the output. Then, the end user might have a certain notion about the accuracy that has been obtained for the given input. This would be especially useful in an interactive system (see below).
- *Generalizing systems*: A pressing issue is generalizing from training datasets to various real-world collections because the costs for data acquisition are still significant and currently represent a bottleneck for applying state-of-the-art machine learning models in stakeholders' workflows. However, music notation follows the same underlying rules, regardless of graphical differences such as whether it is typeset or handwritten. Can one leverage a typeset sheet music dataset to train for handwritten notation? Given that typeset notation can be synthetically generated, this would open several opportunities to train handwritten systems without the effort of getting labeled data manually. Although it seems more difficult to transfer knowledge across different kinds of music notation, a system that recognizes some specific music notation could be somehow useful for the recognition of shared elements in other styles as well, e.g., across the various mensural notation systems.
- *Interactive systems*: Interactive systems are based on the idea of including users in the recognition process, given that they are necessary if there is no tolerance for errors—something that at the moment can only be ensured by human verification. This paradigm reformulates the objective of the system, which is no longer improving accuracy but reducing the effort—usually measured as time—that the users invest in aiding the machine to achieve that perfect result. This aid can be provided in many different ways: error corrections that then feed back into the system, or manually activating and deactivating constraints on the content to be recognized. However, since user effort is the most valuable resource, there is still a need to reformulate the problem based on this concept, which also includes aspects related to human-computer interfaces. The conventional interfaces of computers are designed for entering text (keyboard) or performing very specific actions (mouse); therefore, it would be interesting to study the use of more ergonomic interfaces to work with musical notation, such as an electronic pen or a MIDI piano, in the context of interactive OMR systems.

We hope that these lists demonstrate that OMR still provides many interesting challenges that await future research.

## REFERENCES

- [1] Sanu Pulimootil Achankunju. Music search engine from noisy OMR data. In *1st International Workshop on Reading Music Systems*, pages 23–24, Paris, France, 2018.
- [2] Julia Adamska, Mateusz Piecuch, Mateusz Podgórski, Piotr Walkiewicz, and Ewa Lukasik. Mobile system for optical music recognition and music sound generation. In *Computer Information Systems and Industrial Management*, pages 571–582, Cham, 2015. Springer International Publishing.
- [3] Francisco Álvaro, Joan-Andreu Sánchez, and José-Miguel Benedí. An integrated grammar-based approach for mathematical expression recognition. *Pattern Recognition*, 51:135–147, 2016.
- [4] Alfio Andronico and Alberto Ciampa. On automatic pattern recognition and acquisition of printed music. In *International Computer Music Conference*, Venice, Italy, 1982. Michigan Publishing.

- [5] Tetsuaki Baba, Yuya Kikukawa, Toshiki Yoshiike, Tatsuhiko Suzuki, Rika Shoji, Kumiko Kushiya, and Makoto Aoki. Gocen: A handwritten notational interface for musical performance and learning music. In *ACM SIGGRAPH 2012 Emerging Technologies*, pages 9–9, New York, USA, 2012. ACM.
- [6] David Bainbridge and Tim Bell. Dealing with superimposed objects in optical music recognition. In *6th International Conference on Image Processing and its Applications*, number 443, pages 756–760, 1997.
- [7] David Bainbridge and Tim Bell. The challenge of optical music recognition. *Computers and the Humanities*, 35(2):95–121, 2001.
- [8] David Bainbridge and Tim Bell. A music notation construction engine for optical music recognition. *Software: Practice and Experience*, 33(2):173–200, 2003.
- [9] David Bainbridge and Tim Bell. Identifying music documents in a collection of images. In *7th International Conference on Music Information Retrieval*, pages 47–52, Victoria, Canada, 2006.
- [10] Stefan Balke, Sanu Pulimootil Achankunju, and Meinard Müller. Matching musical themes based on noisy OCR and OMR input. In *International Conference on Acoustics, Speech and Signal Processing*, pages 703–707. Institute of Electrical and Electronics Engineers Inc., 2015.
- [11] Arnau Baró, Pau Riba, Jorge Calvo-Zaragoza, and Alicia Fornés. Optical music recognition by recurrent neural networks. In *14th International Conference on Document Analysis and Recognition*, pages 25–26, Kyoto, Japan, 2017. IEEE.
- [12] Arnau Baró, Pau Riba, and Alicia Fornés. Towards the recognition of compound music notes in handwritten music scores. In *15th International Conference on Frontiers in Handwriting Recognition*, pages 465–470. Institute of Electrical and Electronics Engineers Inc., 2016.
- [13] Arnau Baró, Pau Riba, and Alicia Fornés. A starting point for handwritten music recognition. In *1st International Workshop on Reading Music Systems*, pages 5–6, Paris, France, 2018.
- [14] Louis W. G. Barton. The NEUMES project: digital transcription of medieval chant manuscripts. In *2nd International Conference on Web Delivering of Music*, pages 211–218, 2002.
- [15] Stephan Baumann. A simplified attributed graph grammar for high-level music recognition. In *3rd International Conference on Document Analysis and Recognition*, pages 1080–1083. IEEE, 1995.
- [16] Stephan Baumann and Andreas Dengel. Transforming printed piano music into MIDI. In *Advances in Structural and Syntactic Pattern Recognition*, pages 363–372. World Scientific, 1992.
- [17] Mert Bay, Andreas F. Ehmman, and J. Stephen Downie. Evaluation of multiple-f0 estimation and tracking systems. In *10th International Society for Music Information Retrieval Conference*, pages 315–320, Kobe, Japan, 2009.
- [18] Pierfrancesco Bellini, Ivan Bruno, and Paolo Nesi. Optical music sheet segmentation. In *1st International Conference on WEB Delivering of Music*, pages 183–190. Institute of Electrical & Electronics Engineers (IEEE), 2001.
- [19] Pierfrancesco Bellini, Ivan Bruno, and Paolo Nesi. Assessing optical music recognition tools. *Computer Music Journal*, 31(1):68–93, 2007.
- [20] Margaret Bent and Andrew Wathey. Digital image archive of medieval music. <https://www.diamm.ac.uk>, 1998.
- [21] Hervé Bitteur. Audiveris. <https://github.com/audiveris>, 2004.
- [22] Dorothea Blostein and Henry S. Baird. A critical survey of music image analysis. In *Structured Document Image Analysis*, pages 405–434. Springer Berlin Heidelberg, 1992.
- [23] Dorothea Blostein and Nicholas Paul Carter. Recognition of music notation: Spr’90 working group report. In *Structured Document Image Analysis*, pages 573–574. Springer Berlin Heidelberg, 1992.
- [24] Dorothea Blostein and Lippold Haken. Justification of printed music. *Communications of the ACM*, 34(3):88–99, 1991.
- [25] John Ashley Burgoyne, Johanna Devaney, Laurent Pugin, and Ichiro Fujinaga. Enhanced bleedthrough correction for early music documents with recto-verso registration. In *9th International Conference on Music Information Retrieval*, pages 407–412, Philadelphia, PA, 2008.
- [26] John Ashley Burgoyne, Ichiro Fujinaga, and J. Stephen Downie. Music information retrieval. In *A New Companion to Digital Humanities*, pages 213–228. Wiley Blackwell, 2015.
- [27] Donald Byrd and Eric Isaacson. A music representation requirement specification for academia. Technical report, Indiana University, Bloomington, 2016.
- [28] Donald Byrd and Megan Schindele. Prospects for improving OMR with multiple recognizers. In *7th International Conference on Music Information Retrieval*, pages 41–46, 2006.
- [29] Donald Byrd and Jakob Grue Simonsen. Towards a standard testbed for optical music recognition: Definitions, metrics, and page images. *Journal of New Music Research*, 44(3):169–195, 2015.
- [30] Jorge Calvo-Zaragoza, Jan Hajič jr., and Alexander Pacha. Discussion group summary: Optical music recognition. In *Graphics Recognition, Current Trends and Evolutions*, Lecture Notes in Computer Science, pages 152–157. Springer International Publishing, 2018.
- [31] Jorge Calvo-Zaragoza and Jose Oncina. Recognition of pen-based music notation: The HOMUS dataset. In *22nd International Conference on Pattern Recognition*, pages 3038–3043. Institute of Electrical & Electronics Engineers (IEEE), 2014.
- [32] Jorge Calvo-Zaragoza and David Rizo. Camera-primus: Neural end-to-end optical music recognition on realistic monophonic scores. In *19th International Society for Music Information Retrieval Conference*, pages 248–255, Paris, France, 2018.
- [33] Jorge Calvo-Zaragoza and David Rizo. End-to-end neural optical music recognition of monophonic scores. *Applied Sciences*, 8(4), 2018.
- [34] Jorge Calvo-Zaragoza, Alejandro Toselli, and Enrique Vidal. Handwritten music recognition for mensural notation: Formulation, data and baseline results. In *14th International Conference on Document Analysis and Recognition*, pages 1081–1086, Kyoto, Japan, 2017.
- [35] Jorge Calvo-Zaragoza, Alejandro H. Toselli, and Enrique Vidal. Probabilistic music-symbol spotting in handwritten scores. In *16th International Conference on Frontiers in Handwriting Recognition*, pages 558–563, Niagara Falls, USA, 2018.
- [36] Carlos E. Cancino-Chacón, Maarten Grachten, Werner Goebel, and Gerhard Widmer. Computational models of expressive music performance: A comprehensive and critical review. *Frontiers in Digital Humanities*, 5:25, 2018.

- [37] capella-software AG. Capella scan. <https://www.capella-software.com>, 1996.
- [38] Nicholas Paul Carter. A new edition of walton's façade using automatic score recognition. In *Advances in Structural and Syntactic Pattern Recognition*, pages 352–362. World Scientific, 1992.
- [39] Gen-Fang Chen and Jia-Shing Sheu. An optical music recognition system for traditional chinese kunqu opera scores written in gong-che notation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):7, 2014.
- [40] Liang Chen and Kun Duan. Midi-assisted egocentric optical music recognition. In *Winter Conference on Applications of Computer Vision*. Institute of Electrical and Electronics Engineers Inc., 2016.
- [41] Liang Chen, Rong Jin, and Christopher Raphael. Renotation from optical music recognition. In *Mathematics and Computation in Music*, pages 16–26, Cham, 2015. Springer International Publishing.
- [42] Liang Chen, Rong Jin, and Christopher Raphael. Human-guided recognition of music score images. In *4th International Workshop on Digital Libraries for Musicology*. ACM Press, 2017.
- [43] Liang Chen and Christopher Raphael. Optical music recognition and human-in-the-loop computation. In *1st International Workshop on Reading Music Systems*, pages 11–12, Paris, France, 2018.
- [44] Atul K. Chhabra. Graphic symbol recognition: An overview. In *Graphics Recognition Algorithms and Systems*, pages 68–79, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.
- [45] Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778, 2018.
- [46] Kwon-Young Choi, Bertrand Coüasnon, Yann Ricquebourg, and Richard Zanibbi. Bootstrapping samples of accidentals in dense piano scores for CNN-based detection. In *14th International Conference on Document Analysis and Recognition*, Kyoto, Japan, 2017. IAPR TC10 (Technical Committee on Graphics Recognition), IEEE Computer Society.
- [47] G. Sayeed Choudhury, M. Droetboom, Tim DiLauro, Ichiro Fujinaga, and Brian Harrington. Optical music recognition system within a large-scale digitization project. In *1st International Symposium on Music Information Retrieval*, 2000.
- [48] Arindam Chowdhury and Lovekesh Vig. An efficient end-to-end neural model for handwritten text recognition. In *29th British Machine Vision Conference*, 2018.
- [49] Bertrand Coüasnon and Jean Camillerapp. Using grammars to segment and recognize music scores. In *International Association for Pattern Recognition Workshop on Document Analysis Systems*, pages 15–27, Kaiserslautern, Germany, 1994.
- [50] Tim Crawford, Golnaz Badkobeh, and David Lewis. Searching page-images of early music scanned with omr: A scalable solution using minimal absent words. In *19th International Society for Music Information Retrieval Conference*, pages 233–239, Paris, France, 2018.
- [51] Christoph Dalitz, Georgios K. Michalakakis, and Christine Pranzas. Optical recognition of psaltic byzantine chant notation. *International Journal of Document Analysis and Recognition*, 11(3):143–158, 2008.
- [52] Jürgen Diet. Innovative mir applications at the bayerische staatsbibliothek. In *5th International Conference on Digital Libraries for Musicology*, Paris, France, 2018.
- [53] Ing-Jr Ding, Chih-Ta Yen, Che-Wei Chang, and He-Zhong Lin. Optical music recognition of the singer using formant frequency estimation of vocal fold vibration and lip motion with interpolated gmm classifiers. *Journal of Vibroengineering*, 16(5):2572–2581, 2014.
- [54] Matthias Dorfer, Jan Hajič jr., Andreas Arzt, Harald Frostel, and Gerhard Widmer. Learning audio–sheet music correspondences for cross-modal retrieval and piece identification. *Transactions of the International Society for Music Information Retrieval*, 1(1):22–33, 2018.
- [55] Matthew J. Dovey. Overview of the OMRAS project: Online music retrieval and searching. *Journal of the American Society for Information Science and Technology*, 55(12):1100–1107, 2004.
- [56] Hoda M. Fahmy and Dorothea Blostein. A graph grammar programming style for recognition of music notation. *Machine Vision and Applications*, 6(2):83–99, 1993.
- [57] Jonathan Feist. *Berklee Contemporary Music Notation*. Berklee Press, 2017.
- [58] Alicia Fornés and Lamiroy Bart, editors. *Graphics Recognition, Current Trends and Evolutions*, volume 11009 of *Lecture Notes in Computer Science*. Springer International Publishing, 2018.
- [59] Alicia Fornés, Anjan Dutta, Albert Gordo, and Josep Lladós. The ICDAR 2011 music scores competition: Staff removal and writer identification. In *International Conference on Document Analysis and Recognition*, pages 1511–1515, 2011.
- [60] Alicia Fornés, Anjan Dutta, Albert Gordo, and Josep Lladós. CVC-MUSCIMA: A ground-truth of handwritten music score images for writer identification and staff removal. *International Journal on Document Analysis and Recognition*, 15(3):243–251, 2012.
- [61] Alicia Fornés, Josep Lladós, and Gemma Sánchez. Primitive segmentation in old handwritten music scores. In *Graphics Recognition. Ten Years Review and Future Perspectives*, pages 279–290, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
- [62] Alicia Fornés, Josep Lladós, and Gemma Sánchez. Old handwritten musical symbol classification by a dynamic time warping based method. In *Graphics Recognition. Recent Advances and New Opportunities*, pages 51–60, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.
- [63] Alicia Fornés, Josep Lladós, Gemma Sánchez, and Horst Bunke. Writer identification in old handwritten music scores. In *8th International Workshop on Document Analysis Systems*, pages 347–353, Nara, Japan, 2008.
- [64] Alicia Fornés, Josep Lladós, Gemma Sánchez, and Horst Bunke. On the use of textural features for writer identification in old handwritten music scores. *10th International Conference on Document Analysis and Recognition*, pages 996–1000, 2009.
- [65] Stavroula-Evita Fotinea, George Giakoupis, Aggelos Livens, Stylianos Bakamidis, and George Carayannis. An optical notation recognition system for printed music based on template matching and high level reasoning. In *RIAO '00 Content-Based Multimedia Information Access*, pages 1006–1014, Paris, France, 2000. Le centre de hautes etudes internationales d'informatique documentaire.

- [66] Christian Fremerey, Meinard Müller, Frank Kurth, and Michael Clausen. Automatic mapping of scanned sheet music to audio recordings. In *9th International Conference on Music Information Retrieval*, pages 413–418, 2008.
- [67] Ichiro Fujinaga. Optical music recognition using projections. Master’s thesis, McGill University, 1988.
- [68] Ichiro Fujinaga and Andrew Hankinson. Simssa: Single interface for music score searching and analysis. *Journal of the Japanese Society for Sonic Arts*, 6(3):25–30, 2014.
- [69] Ichiro Fujinaga, Andrew Hankinson, and Julie E. Cumming. Introduction to SIMSSA (single interface for music score searching and analysis). In *1st International Workshop on Digital Libraries for Musicology*, pages 1–3. ACM, 2014.
- [70] Antonio-Javier Gallego and Jorge Calvo-Zaragoza. Staff-line removal with selectional auto-encoders. *Expert Systems with Applications*, 89:138–148, 2017.
- [71] Gear Up AB. iseenotes. <http://www.iseenotes.com>, 2017.
- [72] Susan E. George. Online pen-based recognition of music notation with artificial neural networks. *Computer Music Journal*, 27(2):70–79, 2003.
- [73] Susan E. George. Wavelets for dealing with super-imposed objects in recognition of music notation. In *Visual Perception of Music Notation: On-Line and Off Line Recognition*, pages 78–107. IRM Press, Hershey, PA, 2004.
- [74] Angelos P. Giotis, Giorgos Sfikas, Basilis Gatos, and Christophoros Nikou. A survey of document image word spotting techniques. *Pattern Recognition*, 68:310–332, 2017.
- [75] Michael Good. MusicXML: An internet-friendly format for sheet music. Technical report, Recordare LLC, 2001.
- [76] Michael Good and Geri Actor. Using MusicXML for file interchange. In *Third International Conference on WEB Delivering of Music*, page 153, 2003.
- [77] Albert Gordo, Alicia Fornés, and Ernest Valveny. Writer identification in handwritten musical scores with bags of notes. *Pattern Recognition*, 46(5):1337–1345, 2013.
- [78] Mark Gotham, Peter Jonas, Bruno Bower, William Bosworth, Daniel Rootham, and Leigh VanHandel. Scores of scores: An openscore project to encode and share sheet music. In *5th International Conference on Digital Libraries for Musicology*, pages 87–95, Paris, France, 2018. ACM.
- [79] Elaine Gould. *Behind Bars*. Faber Music, 2011.
- [80] Gianmarco Gozzi. OMRJX: A framework for piano scores optical music recognition. Master’s thesis, Politecnico di Milano, 2010.
- [81] Jan Hajič jr. A case for intrinsic evaluation of optical music recognition. In *1st International Workshop on Reading Music Systems*, pages 15–16, Paris, France, 2018.
- [82] Jan Hajič jr., Matthias Dorfer, Gerhard Widmer, and Pavel Pecina. Towards full-pipeline handwritten OMR with musical symbol detection by u-nets. In *19th International Society for Music Information Retrieval Conference*, pages 225–232, Paris, France, 2018.
- [83] Jan Hajič jr., Marta Kolárová, Alexander Pacha, and Jorge Calvo-Zaragoza. How current optical music recognition systems are becoming useful for digital libraries. In *5th International Conference on Digital Libraries for Musicology*, pages 57–61, Paris, France, 2018. ACM.
- [84] Jan Hajič jr., Jiří Novotný, Pavel Pecina, and Jaroslav Pokorný. Further steps towards a standard testbed for optical music recognition. In *17th International Society for Music Information Retrieval Conference*, pages 157–163, New York, USA, 2016. New York University, New York University.
- [85] Jan Hajič jr. and Pavel Pecina. Groundtruthing (not only) music notation with MUSICMarker: A practical overview. In *14th International Conference on Document Analysis and Recognition*, pages 47–48, Kyoto, Japan, 2017.
- [86] Jan Hajič jr. and Pavel Pecina. The MUSCIMA++ dataset for handwritten optical music recognition. In *14th International Conference on Document Analysis and Recognition*, pages 39–46, Kyoto, Japan, 2017.
- [87] Donna Harman. *Information Retrieval Evaluation*. Morgan & Claypool Publishers, 1st edition, 2011.
- [88] Kate Helsen, Jennifer Bain, Ichiro Fujinaga, Andrew Hankinson, and Debra Lacoste. Optical music recognition and manuscript chant sources. *Early Music*, 42(4):555–558, 2014.
- [89] George Heussenstamm. *The Norton Manual of Music Notation*. W. W. Norton & Company, 1987.
- [90] Władysław Homenda. Automatic recognition of printed music and its conversion into playable music data. *Control and Cybernetics*, 25(2):353–367, 1996.
- [91] Yu-Hui Huang, Xuanli Chen, Serafina Beck, David Burn, and Luc Van Gool. Automatic handwritten mensural notation interpreter: From manuscript to MIDI performance. In *16th International Society for Music Information Retrieval Conference*, pages 79–85, Málaga, Spain, 2015.
- [92] José Manuel Iñesta, Pedro J. Ponce de León, David Rizo, José Oncina, Luisa Micó, Juan Ramón Rico-Juan, Carlos Pérez-Sancho, and Antonio Pertusa. Hispamus: Handwritten spanish music heritage preservation by automatic transcription. In *1st International Workshop on Reading Music Systems*, pages 17–18, Paris, France, 2018.
- [93] Linn Saxrud Johansen. Optical music recognition. Master’s thesis, University of Oslo, 2009.
- [94] Graham Jones, Bee Ong, Ivan Bruno, and Kia Ng. Optical music imaging: Music document digitisation, recognition, evaluation, and restoration. In *Interactive multimedia music technologies*, pages 50–79. IGI Global, 2008.
- [95] Nicholas Journet, Muriel Visani, Boris Mansencal, Kieu Van-Cuong, and Antoine Billy. DocCreator: A new software for creating synthetic ground-truthed document images. *Journal of Imaging*, 3(4):62, 2017.
- [96] Michael Kassler. Optical character-recognition of printed music : A review of two dissertations. automatic recognition of sheet music by dennis howard pruslin ; computer pattern recognition of standard engraved music notation by david stewart prerai. *Perspectives of New Music*, 11(1):250–254, 1972.
- [97] Klaus Keil and Jennifer A. Ward. Applications of RISM data in digital libraries and digital musicology. *International Journal on Digital Libraries*, 2017.
- [98] Daniel Lopresti and George Nagy. Issues in ground-truthing graphic documents. In *Graphics Recognition Algorithms and Applications*, pages 46–67. Springer Berlin Heidelberg, Ontario, Canada, 2002.

- [99] Nawapon Luangnapa, Thongchai Silpavarangkura, Chakarida Nukoolkit, and Pornchai Mongkolnam. Optical music recognition on android platform. In *International Conference on Advances in Information Technology*, pages 106–115. Springer, 2012.
- [100] Rakesh Malik, Partha Pratim Roy, Umapada Pal, and Fumitaka Kimura. Handwritten musical document retrieval using music-score spotting. In *12th International Conference on Document Analysis and Recognition*, pages 832–836, 2013.
- [101] Chirstopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [102] T. Matsushima, I. Sonomoto, T. Harada, K. Kanamori, and S. Ohteru. Automated high speed recognition of printed music (WABOT-2 vision system). In *International Conference on Advanced Robotics*, pages 477–482, 1985.
- [103] Johann Mattheson. *Der vollkommene Capellmeister*. Herold, Christian, Hamburg, 1739.
- [104] Apurva A. Mehta and Malay S. Bhatt. Optical music notes recognition for printed piano music score sheet. In *International Conference on Computer Communication and Informatics*, Coimbatore, India, 2015.
- [105] Hidetoshi Miyao and Robert Martin Haralick. Format of ground truth data used in the evaluation of the results of an optical music recognition system. In *4th International Workshop on Document Analysis Systems*, pages 497–506, Brasil, 2000.
- [106] Musitek. Smartscore x2. <http://www.musitek.com/smartscore-pro.html>, 2017.
- [107] Neuratron. Notateme. <http://www.neuratron.com/notateme.html>, 2015.
- [108] Neuratron. Photoscore 2018. <http://www.neuratron.com/photoscore.htm>, 2018.
- [109] Kia Ng, Alex McLean, and Alan Marsden. Big data optical music recognition with multi images and multi recognisers. In *EVA London 2014 on Electronic Visualisation and the Arts*, pages 215–218. BCS, 2014.
- [110] Tam Nguyen and Guesang Lee. A lightweight and effective music score recognition on mobile phones. *Journal of Information Processing Systems*, 11(3):438–449, 2015.
- [111] Jiri Novotný and Jaroslav Pokorný. Introduction to optical music recognition: Overview and practical challenges. In *Annual International Workshop on DAtabases, TExts, Specifications and Objects*, pages 65–76. CEUR-WS, 2015.
- [112] Organum. Playscore. <http://www.playscore.co>, 2016.
- [113] Rafael Ornes. Choral public domain library. <http://cpdl.org>, 1998.
- [114] Tuula Pääkkönen, Jukka Kervinen, and Kimmo Kettunen. Digitisation and digital library presentation system – sheet music to the mix. In *1st International Workshop on Reading Music Systems*, pages 21–22, Paris, France, 2018.
- [115] Alexander Pacha. Advancing omr as a community: Best practices for reproducible research. In *1st International Workshop on Reading Music Systems*, pages 19–20, Paris, France, 2018.
- [116] Alexander Pacha and Jorge Calvo-Zaragoza. Optical music recognition in mensural notation with region-based convolutional neural networks. In *19th International Society for Music Information Retrieval Conference*, pages 240–247, Paris, France, 2018.
- [117] Alexander Pacha, Kwon-Young Choi, Bertrand Couasnon, Yann Ricquebourg, Richard Zanibbi, and Horst Eidenberger. Handwritten music object detection: Open issues and baseline results. In *13th International Workshop on Document Analysis Systems*, pages 163–168, 2018.
- [118] Alexander Pacha and Horst Eidenberger. Towards a universal music symbol classifier. In *14th International Conference on Document Analysis and Recognition*, pages 35–36, Kyoto, Japan, 2017. IAPR TC10 (Technical Committee on Graphics Recognition), IEEE Computer Society.
- [119] Alexander Pacha, Jan Hajič jr., and Jorge Calvo-Zaragoza. A baseline for general music object detection with deep learning. *Applied Sciences*, 8(9):1488–1508, 2018.
- [120] Victor Padilla, Alan Marsden, Alex McLean, and Kia Ng. Improving omr for digital music libraries with multiple recognisers and multiple sources. In *1st International Workshop on Digital Libraries for Musicology*, pages 1–8, London, United Kingdom, 2014. ACM.
- [121] Emilia Parada-Cabaleiro, Anton Batliner, Alice Baird, and Björn Schuller. The seils dataset: Symbolically encoded scores in modern-early notation for computational musicology. In *18th International Society for Music Information Retrieval Conference*, Suzhou, China, 2017.
- [122] Viet-Khoi Pham, Hai-Dang Nguyen, and Minh-Triet Tran. Virtual music teacher for new music learners with optical music recognition. In *International Conference on Learning and Collaboration Technologies*, pages 415–426. Springer, 2015.
- [123] Réjean Plamondon and Sargur N. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):63–84, 2000.
- [124] David S. Prerau. Computer pattern recognition of printed music. In *Fall Joint Computer Conference*, pages 153–162, 1971.
- [125] Gérard Presgurvic. *Songbook romeo & julia*, 2005.
- [126] Project Petrucci LLC. International music score library project. <http://imslp.org>, 2006.
- [127] Dennis Howard Pruslin. *Automatic Recognition of Sheet Music*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 1966.
- [128] Laurent Pugin. Optical music recognition of early typographic prints using hidden Markov models. In *7th International Conference on Music Information Retrieval*, pages 53–56, Victoria, Canada, 2006.
- [129] Laurent Pugin, John Ashley Burgoyne, and Ichiro Fujinaga. Reducing costs for digitising early music with dynamic adaptation. In *Research and Advanced Technology for Digital Libraries*, pages 471–474, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.
- [130] Laurent Pugin and Tim Crawford. Evaluating OMR on the early music online collection. In *14th International Society for Music Information Retrieval Conference*, pages 439–444, Curitiba, Brazil, 2013.
- [131] Gene Ragan. Kompapp. <http://kompapp.com>, 2017.
- [132] Sheikh Faisal Rashid, Abdullah Akmal, Muhammad Adnan, Ali Adnan Aslam, and Andreas Dengel. Table recognition in heterogeneous documents using machine learning. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 777–782, 2017.
- [133] Ana Rebelo, Ichiro Fujinaga, Filipe Paszkiewicz, Andre R.S. Marcal, Carlos Guedes, and Jamie dos Santos Cardoso. Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval*, 1(3):173–190, 2012.

- [134] Pau Riba, Alicia Fornés, and Josep Lladós. Towards the alignment of handwritten music scores. In *Graphic Recognition. Current Trends and Challenges*, Lecture Notes in Computer Science, pages 103–116. Springer Verlag, 2017.
- [135] Adrià Rico Blanes and Alicia Fornés Bisquerra. Camera-based optical music recognition using a convolutional neural network. In *14th International Conference on Document Analysis and Recognition*, pages 27–28, Kyoto, Japan, 2017. IEEE.
- [136] David Rizo, Jorge Calvo-Zaragoza, and José M. Iñesta. Muret: A music recognition, encoding, and transcription tool. In *5th International Conference on Digital Libraries for Musicology*, pages 52–56, Paris, France, 2018. ACM.
- [137] Heinz Roggenkemper and Ryan Roggenkemper. How can machine learning make optical music recognition more relevant for practicing musicians? In *1st International Workshop on Reading Music Systems*, pages 25–26, Paris, France, 2018.
- [138] Perry Roland. The music encoding initiative (MEI). In *1st International Conference on Musical Applications Using XML*, pages 55–59, 2002.
- [139] Florence Rossant and Isabelle Bloch. A fuzzy model for optical recognition of musical scores. *Fuzzy Sets and Systems*, 141(2):165–201, 2004.
- [140] Partha Pratim Roy, Ayan Kumar Bhunia, and Umapada Pal. HMM-based writer identification in music score documents without staff-line removal. *Expert Systems with Applications*, 89:222–240, 2017.
- [141] Sächsische Landesbibliothek. Staats- und universitätsbibliothek dresden. <https://www.slub-dresden.de>, 2007.
- [142] Charalampos Saitis, Andrew Hankinson, and Ichiro Fujinaga. Correcting large-scale OMR data with crowdsourcing. In *1st International Workshop on Digital Libraries for Musicology*, pages 1–3. ACM, 2014.
- [143] Zeyad Saleh, Ke Zhang, Jorge Calvo-Zaragoza, Gabriel Vigliensoni, and Ichiro Fujinaga. Pixel.js: Web-based pixel classification correction platform for ground truth creation. In *14th International Conference on Document Analysis and Recognition*, pages 39–40, Kyoto, Japan, 2017.
- [144] Marc Schobrun. *The Everything Reading Music Book: A Step-By-Step Introduction to Understanding Music Notation And Theory*. Everything series. Adams Media, 2005.
- [145] Eleanor Selfridge-Field. *Beyond MIDI: The Handbook of Musical Codes*. MIT Press, Cambridge, MA, USA, 1997.
- [146] Asif Shahab, Faisal Shafait, Thomas Kieninger, and Andreas Dengel. An open approach towards the benchmarking of table structure recognition systems. In *9th International Workshop on Document Analysis Systems*, pages 113–120, Boston, Massachusetts, USA, 2010. ACM.
- [147] Muhammad Sharif, Quratul-Ain Arshad, Mudassar Raza, and Wazir Zada Khan. [COMSCAN]: An optical music recognition system. In *7th International Conference on Frontiers of Information Technology*, page 34. ACM, 2009.
- [148] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2017.
- [149] Mahmood Sotoodeh, Farshad Tajeripour, Sadeh Teimori, and Kirk Jorgensen. A music symbols recognition method using pattern matching along with integrated projection and morphological operation techniques. *Multimedia Tools and Applications*, 77(13):16833–16866, 2018.
- [150] Daniel Spreadbury and Robert Piéchaud. Standard music font layout (SMuFL). In *First International Conference on Technologies for Music Notation and Representation - TENOR2015*, pages 146–153, Paris, France, 2015. Institut de Recherche en Musicologie.
- [151] StaffPad Ltd. Staffpad. <http://www.staffpad.net>, 2017.
- [152] Gabriel Taubman. Musichand : A handwritten music recognition system. Technical report, Brown University, 2005.
- [153] Jessica Thompson, Andrew Hankinson, and Ichiro Fujinaga. Searching the liber usualis: Using CouchDB and elasticsearch to query graphical music documents. In *12th International Society for Music Information Retrieval Conference*, 2011.
- [154] Lukas Tuggener, Ismail Elezi, Jürgen Schmidhuber, Marcello Pelillo, and Thilo Stadelmann. Deepscores - a dataset for segmentation, detection and classification of tiny objects. In *24th International Conference on Pattern Recognition*, Beijing, China, 2018.
- [155] Julián Urbano. MIREX 2013 symbolic melodic similarity: A geometric model supported with hybrid sequence alignment. Technical report, Music Information Retrieval Evaluation eXchange, 2013.
- [156] Julián Urbano, Juan Lloréns, Jorge Morato, and Sonia Sánchez-Cuadrado. MIREX 2010 symbolic melodic similarity: Local alignment with geometric representations. Technical report, Music Information Retrieval Evaluation eXchange, 2010.
- [157] Julián Urbano, Juan Lloréns, Jorge Morato, and Sonia Sánchez-Cuadrado. MIREX 2011 symbolic melodic similarity: Sequence alignment with geometric representations. Technical report, Music Information Retrieval Evaluation eXchange, 2011.
- [158] Julián Urbano, Juan Lloréns, Jorge Morato, and Sonia Sánchez-Cuadrado. MIREX 2012 symbolic melodic similarity: Hybrid sequence alignment with geometric representations. Technical report, Music Information Retrieval Evaluation eXchange, 2012.
- [159] Eelco van der Wel and Karen Ullrich. Optical music recognition with convolutional sequence-to-sequence models. In *18th International Society for Music Information Retrieval Conference*, Suzhou, China, 2017.
- [160] Gabriel Vigliensoni, John Ashley Burgoyne, Andrew Hankinson, and Ichiro Fujinaga. Automatic pitch detection in printed square notation. In *12th International Society for Music Information Retrieval Conference*, pages 423–428, Miami, Florida, 2011. University of Miami.
- [161] Gabriel Vigliensoni, Jorge Calvo-Zaragoza, and Ichiro Fujinaga. Developing an environment for teaching computers to read music. In *1st International Workshop on Reading Music Systems*, pages 27–28, Paris, France, 2018.
- [162] Quang Nhat Vo, Guee Sang Lee, Soo Hyung Kim, and Hyung Jeong Yang. Recognition of music scores with non-linear distortions in mobile devices. *Multimedia Tools and Applications*, 77(12):15951–15969, 2018.
- [163] Matthias Wallner. A system for optical music recognition and audio synthesis. Master’s thesis, TU Wien, 2014.
- [164] Gus G. Xia and Roger B. Dannenberg. Improvised duet interaction: Learning improvisation techniques for automatic accompaniment. In *New Interfaces for Musical Expression*, Aalborg University Copenhagen, Denmark, 2017.
- [165] Jianshu Zhang, Jun Du, Shiliang Zhang, Dan Liu, Yulong Hu, Jinshui Hu, Si Wei, and Lirong Dai. Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. *Pattern Recognition*, 71:196–206, 2017.

## APPENDIX A: OMR BIBLIOGRAPHY

Along with this paper, we are also publishing the most comprehensive and complete bibliography on OMR that we were able to compile at <https://omr-research.github.io/>. It is a curated list of verified publications in an open-source Github repository (<https://github.com/OMR-Research/omr-research.github.io>) that is open for submissions both via pull requests and via templated issues. The website is automatically generated from the underlying BibTeX files using the BibTex2HTML library, available at <https://www.lri.fr/~filliatr/bibtex2html/>.

The repository contains three distinct bibliographic files that are rendered into separate pages:

- 1) *OMR Research Bibliography*: A collection of scientific and technical publications, whose bibliographical metadata were manually verified for correctness from a trustworthy source (see below). Most of these entries have either a Digital Object Identifier (DOI) or a link to the website, where the publication can be found.
- 2) *OMR Related Bibliography*: A collection of scientific and technical publications, whose bibliographical metadata were manually verified for correctness from a trustworthy source but are not primarily directed towards OMR, such as musicological research or general computer vision papers.
- 3) *Unverified OMR Bibliography*: A collection of scientific and technical publications, that are related to Optical Music Recognition, but they could not be verified from a trustworthy source and might contain incorrect information. Many publications from this collection were authored before 1990 and are often not indexed by the search engines, or the respective proceedings could no longer be accessed and verified by us.

### *Acquisition and Verification Process*

The bibliography was acquired and merged from multiple sources, such as the public and private collections from multiple researchers that have historically grown, including a recent one by Andrew Hankinson, who provided us with an extensive BibTeX library. Additionally, we have a Google Scholar Alert on the survey published by Rebelo et al. [133], as it currently represents the latest state-of-the-art review and is cited by almost every new publication. To verify the information of each entry in the bibliography, we proceeded with the following steps:

- 1) Search on Google Scholar for the title of the work, if necessary with the authors last name and the year of publication.
- 2) Find a trustworthy source such as the original publisher, the authors' website, the website of the venue (that lists the article in the program) or indexing services including IEEE Xplore Digital Library, ACM Digital Library, Springer Link, Elsevier ScienceDirect, arXiv.org, dblp.org or ResearchGate. Information from the last three services are used with caution and if possible backed up with information from other sources.
- 3) Manually verify the correctness of the metadata by inspecting and correct it by obtaining the necessary information from another source, e.g., the conference website or the information state in the document. Suspicious information could be if the author's name is missing letters because of special characters or if the year of publication is before that of cited references.

Once we verified the entry, we add it to the respective bibliography with JabRef (<http://www.jabref.org/>) and link the original PDF file or at least the DOI. Articles that were only found as PDF without the associated venue of publication were classified as technical reports. Bachelor theses and online sources such as websites of commercial applications were classified as 'Misc' because of the lack of an appropriate category in BibTeX.