Towards a Standard Testbed for Optical Music Recognition: Definitions, Metrics, and Page Images

Donald Byrd, Indiana University Bloomington
and
Jakob Grue Simonsen, University of Copenhagen

Early March 2013

Abstract

We believe that progress in Optical Music Recognition (OMR) has been held up for years by the absence of anything like the standard testbeds in use in other fields that face difficult evaluation problems. One example of such a field is text IR, where the TREC conference has annually-renewed IR tasks with accompanying data sets. In music informatics, MIREX, with its annual tests and meetings held during the ISMIR conference, is an almost exact analog to TREC; but MIREX has never had an OMR track or a collection of music such a track could employ. We describe why the absence of an OMR testbed is a problem and how this problem may be mitigated or solved outright. To aid in the establishment of a standard testbed, we devise performance metrics for OMR tools that take into account score complexity and graphic quality. Finally, we provide a small corpus of music for use as a miniature baseline for a proper OMR testbed.

The Problem and The Solution

What is the most accurate Optical Music Recognition (henceforth OMR) system available? A perfectly good answer is “No one knows, and there’s no practical way to find out”. But—considering the enormous variability of music notation—it is unreasonable to expect an answer to such a general question; a more helpful answer would be “It depends”. Consider this question instead: What’s the most accurate OMR system available for a *specific* kind of music and publication, in digitized page images with given quality? That certainly seems like a reasonable question, but the answer is still “No one knows, and there’s no practical way to find out”. In our view, the reason is that OMR evaluation is in the dark ages, and we believe it will remain there until a *standard testbed* exists for it.

Some eighteen years have passed since the groundbreaking study of OMR systems by Nick Carter and others at the CCARH (Selfridge-Field, Carter, et al, 1994), and it is not at all clear that much progress has been made on OMR systems since then. Perhaps it has been, but no one can be certain. It is well known in the document-recognition community that the difficulty of evaluating document-recognition systems (as opposed to the difficulty of creating them) varies enormously. The most familiar type of recognition system, Optical Character Recognition (OCR), is probably among the easiest to evaluate; the system to recognize conventional Western music notation is undoubtedly among the hardest. (See the section “Why OMR Evaluation is Difficult” below for discussion of the reasons.)

In a recent survey of the state-of-the-art of OMR, Rebelo et al (2012) listed evaluation as one of the four main open issues. But evaluation is notoriously difficult: Some years ago, Droettboom and Fujinaga (2004) wrote that “It could be argued that a complete and robust system to evaluate OMR output would almost as complex and error-prone as an OMR system itself.” But other disciplines—among them, text IR, speech recognition, and even music IR (for, e.g., audio chord estimation)—have faced very difficult evaluation problems and solved them. In every case we’re aware of, the solution has been the same: The interested community has established standard testbeds, with well-thought-out test collections and evaluation metrics; run regular "contests" for researchers in the field; and held a conference after each "contest" in which the participants report on and discuss what happened. Text IR may be the earliest example, with its TREC (Text REtrieval Conference) series. The U.S. National Institute of Standards and Technology established annual TREC conferences, with "tracks" for specific text-IR tasks, some 20 years ago. In music informatics, MIREX and its annual tests, held in conjunction with the annual ISMIR conferences, is a close analog to TREC; but MIREX has never had an OMR track, and it is not self-evident that the "well-thought-out test collections and evaluation metrics" materials such a track would need exist.

The case for a standard testbed for OMR has been made before. For example, from Byrd et al (2010), Sec. 4.2:

> "It is clear that the state of [OMR evaluation] is dismal, and one reason, we have argued, is that it is inherently very difficult. An appropriate evaluation metric is, to put it mildly, not straightforward... As a first step, the existence of a standard testbed—i.e., some reasonable (even if very imperfect) evaluation metrics, and a set of carefully selected music page images to apply them to—would at least allow some meaningful comparisons of OMR programs. But, to our knowledge, there are absolutely no standards of any kind in general use; that is the second reason current OMR evaluation is so unsatisfactory... If program A is tested with one collection of music and program B is tested with a different collection, and there is no basis for saying which music is more challenging, how can the results possibly be compared? What if one system is evaluated at a low level and the other at a high level? That alone might lead to a very different error rate. As things stand, there's no way to know if any OMR system, conventional or multiple-recognizer, really is more accurate than any other for any given collection of music, much less for music in general."

(We define the terms "low level" and "high level" under Why OMR Evaluation is Difficult, below.) If anything, the above quotation understates how bad the current situation is. The vast majority of OMR studies we have seen say little about how they chose the pages in their test collection; they appear to be what statisticians would call "convenience samples", chosen primarily because they were easily available. One exception is the "SFStudy" collection of 24 pages, so called because the pages were first used for a "Scoping and Feasibility Study" for the OMR component of the MeTAMuSE project (Byrd & Schindele 2006, 2007). The SFStudy collection has since been used by Bugge et al (2011) in their own OMR work. We discuss this collection in detail under A Small Corpus of Music Pages, below.

OMR and OMR Evaluation

If one wishes to understand OMR evaluation, understanding how OMR works is very helpful. Several good accounts have been published. See for example Bainbridge & Bell (2001), Jones et al. (2008), and Rebelo et al. (2012). Jones et al. (2008) is particularly interesting because its lead author is the creator of SharpEye, one of the best-known commercial OMR programs, and the article contains a detailed description of how SharpEye works. It also has some insightful discussion of OMR evaluation, as does the Rebelo et al. paper.

Is Evaluation Really a Serious Problem Yet?

It has been argued that OMR research/development is in such a primitive state that evaluation is not yet a serious problem. We disagree. That argument makes sense for a field in its infancy, where noone has tried to achieve the same thing in two different ways; but the OMR literature already contains many examples of things that people have tried to achieve in various ways. In addition, commercial OMR systems have now been available for perhaps 15 years. For most of that time, their vendors have made claims about their accuracy that are not just impossible to compare but so ill-defined as to be virtually meaningless. Thus, there has been no effective competition between OMR systems on the basis of accuracy, and far less incentive than there might be for vendors to improve their products and for researchers to come up with new ideas.

Approaches to OMR Evaluation

The obvious way to approach evaluation of any system is "end-to-end", that is, to measure the performance of the system as a whole. But with a problem as difficult as OMR, there is much to be said for comparing ways to solve subproblems: in the case of OMR, tasks like identifying and (for most systems) removing staff lines, segmenting the image into symbols, recognizing the individual symbols, and so on. The paper by Rebolo et al. (2012) lists available materials for evaluating a number of OMR subproblems.

We are personally interested in the end-to-end approach, and that is what the remainder of this paper is concerned with.

Why OMR Evaluation is Difficult

We have quoted approvingly the statement by Byrd et al (2010) that "An appropriate evaluation metric [for OMR systems] is, to put it mildly, not straightforward." Why is Western music notation so much harder for computers to handle than text or speech? One reason is that text and speech are both representations of natural language, and natural language is fundamentally one-dimensional; but Western music notation represents Western music, which—except for monophonic music like unaccompanied songs—is fundamentally two-dimensional (Chris Raphael, personal communication, February 2013). A second reason (in part a result of the first) is the enormously complex semantics of music notation and the high degree of context dependency in symbols that it entails (Byrd, 1984).

Byrd et al (2010) comment further that "evaluating OMR systems presents at least four major problems", namely:

1. The level at which errors should be counted.

2. Number of errors vs. effort to correct.

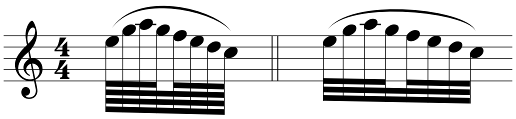
3. Relative importance of symbols.

4. Variability of complexity of the notation.

We now briefly discuss these issues. See Bainbridge & Bell (2001), Droettboom & Fujinaga (2004), and Byrd et al (2010) for more detail.

Problem 1. Level of Evaluation

A phenomenon that is discussed more or less directly by Reed (1995), Bainbridge & Bell (2001), Droettboom & Fujinaga (2004), and others, is: *document-recognition systems can be described at different levels, and they may behave very differently at each level.* In particular, they may be far more accurate at one level than at another. Almost any optical recognition system can be described at the pixel level—the lowest possible level—but such a description is not likely to be very informative. OCR systems are usually evaluated at the low level of characters, and for most text-recognition situations, that is satisfactory. If not, the word level—a higher level—nearly always suffices, and the relationship between the two levels is very straightforward. With music, however, things are much more complex. To clarify, consider Figure 1 (from Reed, 1995, p. 73).



(a) Original score (b) Reconstructed score

Figure 1

In this case, the *high-level symbols*, i.e., symbols with semantics, are the clef, time signature, notes (64ths in Figure 1a, 32nds in 1b), and slur. The clef and slur are each a single *low-level symbol*. But the time signature and notes are comprised of multiple low-level symbols: for the former, numbers; for the latter, noteheads and beams (and, in other cases, flags, accidentals, augmentation dots, etc.).

Reed points out that in this example (ignoring the clef and time signature), the only problem in the reconstructed score is a single missing beam, and if you count low-level symbols, 19 of 20 (95%) are correct. But if you count high-level symbols, only 1 of 9 (11%) is correct. This example shows how a mistake in one low-level symbol can cause numerous errors in high-level symbols, and therefore (in this case) in durations of many notes. But context dependence in music notation is not just a matter of low-level vs. high-level symbols. For example, getting the clef wrong is likely to cause *secondary errors* in note pitches for many following measures: perhaps dozens of notes, if not more.

Unfortunately, each level has its advantages (Droettboom & Fujinaga, 2004), so the best choice depends on the situation. Both of us have used high-level evaluation in our previous research simply because we were working on "multiple-recognizer OMR": the idea was to at "triangulate" in the output of several commercial OMR systems to come up with a reading more accurate than any of them. However, we had access only to these systems' output of finished scores, and sometimes only in CWMN form. Under the circumstances, high-level evaluation made more sense.

Problem 2. Number of errors vs. effort to correct

It is not clear whether an evaluation should consider the number of errors or the amount of work necessary to correct them. The latter is more relevant for many purposes, but it is very dependent on the tools available, e.g., for correcting the pitches of notes resulting from a wrong clef. As Ichiro Fujinaga has pointed out (personal communication, March 2007), it also depends greatly on the distribution and details of the errors: it is far easier to correct 100 consecutive eighth notes that should all be 16ths, than to correct 100 eighth notes whose proper durations vary sprinkled throughout a score. Another issue, closely related, is whether "secondary errors" clearly resulting from an error earlier in the OMR process should be counted, or only primary errors?

Problem 3. Relative importance of symbols

With media like text, at least for natural languages, it is reasonable to assume that all symbols and all mistakes in identifying them are equally important. With music, that is not even remotely the case. It seems clear that note durations and pitches are the most important things, but after that, nothing is obvious. Identifying the voice membership of notes (at least) is important in many cases but not all. How important are redundant or cautionary accidentals? Fingerings? Articulation marks?

Problem 4. Variability of complexity of the notation

The complexity of notation of some music inherently presents far greater challenges for OMR than does the notation of other music, independent of the printing quality and condition of the page image (Byrd 2010). Given the same quality of scanned images, a page of *Le Sacre du Printemps* or of the piano reduction of the Prelude to *Parsifal* is almost certain to have proportionally more recognition errors than a page of fiddle tunes or of a Bach suite for solo 'cello. Again, this factor hardly applies to text.

Pros and Cons of A Standard OMR Testbed

It is not hard to see some reason why a proper OMR testbed does not exist. Ideally any such testbed would include a large collection of test pages—say, at least a few hundred pages—and would be updated regularly in order to keep systems from being targeted at a known, static dataset, in a fashion similar to the current practice at TREC and MIREX. But resources are scarce: the OMR community is much smaller than the text IR, speech recognition, and musical audio communities, and there is no commercial organization representing the companies involved in OMR. It seems unlikely that an annually updated testbed with a lot of music, associated tasks, and impartial evaluation is forthcoming anytime soon.

On the other hand, as things stand, almost every new OMR project we have seen uses a different evaluation method and test collection, apparently put together from scratch. Besides the obvious duplication of effort, it seems unlikely that most of the resulting testing systems are well-planned—and, of course, comparing the performance of different systems is virtually impossible. Even if maintaining an ideal testbed is not viable, we believe that a publicly available testbed with a static collection is far preferable to current practice. Furthermore, even in the absence of active updates or curation, researchers can point out deficiencies in such a testbed which may be taken into account by future OMR projects utilizing it.

After all, once evaluation metrics (and definitions to build them) and a methodology for choosing test pages are agreed on, replacing the pages with new ones is a relatively simple matter. And if researchers or developers of OMR systems hyper-specialize their tools for the testbed's initial collection (the software equivalent of public-school teachers "teaching to the test"), a new set of pages will stop them dead.

The Way Forward

As stepping stones towards establishing a standard testbed, we offer:

- definitions of levels for notation complexity (to handle Problem 4) and for flaws in image quality;
- an evaluation metric in the form of OMR error types and guidelines for counting errors at a high level (to handle Problems 2 and 3); and
- a small corpus: 24 pages of music.

These are all the components a real testbed requires, though, as explained above, no static corpus can be completely satisfactory. However, we believe that the definitions and metrics will be of value independently from the corpus. On the other hand, the definitions and corpus will be valuable independently of the metrics: for example, with a different set of metrics to support low-level evaluation (Problem 1).

Jones et al. (2008) suggest an approach to evaluation relying on similar components.

Complexity of the Music

Byrd et al (2010) discusses our Problem 4, the fact that some music is simply much more difficult to recognize than other music, and gives examples of easy and very difficult music. Reducing the complexity of notated music to a one-dimensional scale in a meaningful way is no easy feat, but here is an attempt to make at least some distinctions that seem appropriate for OMR. In all cases, suffixing “x” to the level means artificial music (“composed” for testing); otherwise it’s real music.

Level 1: monophonic music with one staff per system; simple rhythm (no tuplets) and pitch notation (no octave signs); no lyrics, grace notes, cues, or clef changes. No beams or slurs.

Level 2: one voice per staff (but chords are allowed); not-too-complex rhythm (no tuplets) and pitch notation (no octave signs); no lyrics, cues, or clef changes; no cross-system anything (chords, beams, slurs, etc.); chords, beams or slurs in a single system are allowed. (Music as simple as Haydn’s Symphony No. 1, used in the landmark CCARH study (Selfridge-Field et al, 1994), exceeds these limits with numerous triplets, both marked and unmarked, and with two voices on a staff in many places.)

Level 3: Level 2 features plus triplets, both marked and unmarked, and two voices on a staff (separate notes, note vs. rest, or single notehead with 2 stems); also grace notes, slurs, measured tremolos, dynamics, repeat bars, one verse of lyrics, lyric extender lines. (These are the notational features Haydn Symphony No. 1 uses, plus more-or-less those in “typical” simple piano/vocal scores like simpler Schubert and Beatles songs.)

Level 4: anything else.

Image Quality Flaws

The quality rating scale is based on inspection by naked eye with normal (20/20) vision. It is assumed that very low resolution images will result in low ratings (see “jagged” in the descriptions below). Note that these ratings describe only the quality of the reproduction of the original page; they do not take into account the quality of the notation on the page.

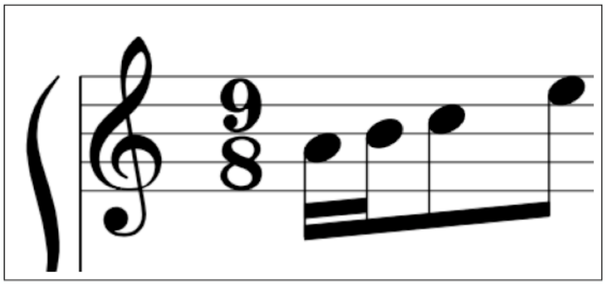
Level 1: Flawless quality. Expected from “born digital” (as opposed to scanned) images. Absolutely no visual noise, no pixels turned on outside music symbols, no pixels turned off inside symbols. Each symbol (including staff lines, barlines, etc.) appears totally contiguous with no jagged or fuzzy edges.

Level 2: Very high quality. Expected from the best scans. Small amounts of visual noise, no pixels turned off inside music symbols; a few pixels may possibly be turned on outside symbols. Symbols may appear jagged, but each symbol is totally contiguous.

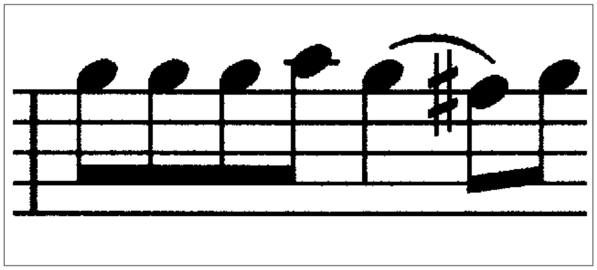
Level 3: Expected from some scans. Some amount of visual noise, pixels may be turned on outside music symbols, and pixels may be turned off inside music symbols if this does not render the symbols non-contiguous (pixels may be turned off in note heads, inside dynamics or accidentals, but no barlines, stems, articulations, etc. may be broken).

Level 4: Expected from some low-quality scans. Visual noise sufficient to blur separation of symbols (e.g., accidentals or flags being amalgamated into stave or barlines). Pixels being turned off may render symbols non-contiguous.

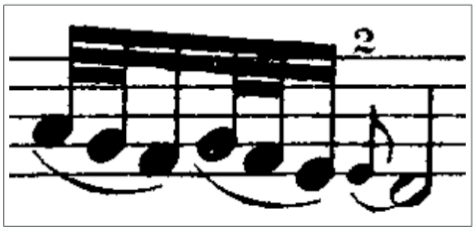
Level 5: Expected from low-quality scans of poorly-printed originals. Large amounts of visual noise. Pixels turned on or off make sight reading difficult, and detailed inspection is necessary to identify some symbols.



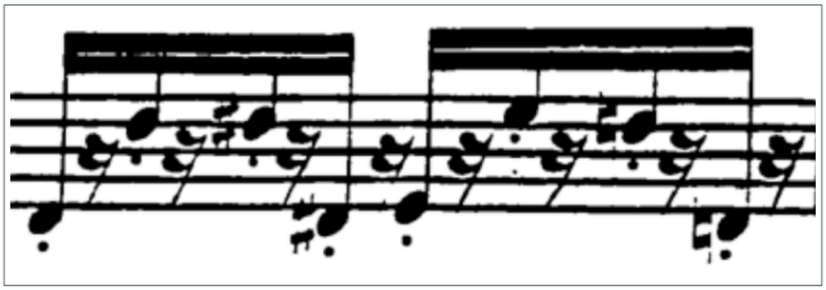
Level 1



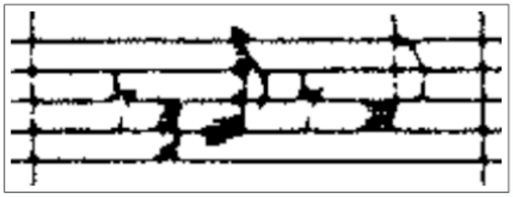
Level 2



Level 3



Level 4



Level 5

How to Count Errors

Types of Errors

Is it worth classifying errors into different types? Some researchers do, some don't, and, to our knowledge, there's no agreement on types among those that use them. The two of us have used very different approaches in our previous work. Simonsen and his colleagues (Bugge et al., 2011) used a relatively simple approach, reporting only total high-level errors. Byrd's MeTAMuSE research (Byrd & Schindele, 2006, 2007) had a much more complex method, also counting high-level errors, but distinguishing seven types. Unless otherwise indicated, the word "note" here means a normal note, not a grace note, cue note, or rest:

	Description of Error
1	Wrong pitch of note (even if due to extra or missing accidentals)
2	Wrong duration of note (even if due to extra or missing augmentation dots)
3	Misinterpretation (other symbol for note, note for other symbol, misspelled line of text, slur beginning/ending on wrong notes, note stem interpreted as barline, etc.)
4	Missing note (even if due to note stem interpreted as barline)
5	Missing symbol other than notes (and accidentals and augmentation dots)
6	Extra symbol (other than accidentals and augmentation dots)
7	Gross misinterpretation (e.g., entire staff missing); *or,* note interpreted as cue or grace note

Rules And Guidelines For Hand Counting Errors

The mere descriptions of error types above leave many questions unanswered. MeTAMuSE, the multiple-recognizer OMR project of one of us (Byrd), used the following rules and guidelines for its high-level evaluation (Byrd & Guerin, 2007). These rules are certainly not the last word, but we believe they are a reasonable starting point.

1. Count errors even in symbols that really carry no information, i.e., that should be redundant. For example, we've seen cases where—despite the fact that there are no key signature changes on the whole page—an OMR program gets the wrong signature in just one system in the middle of the page: that error should be counted.

2. Separate graphical symbols that don't correspond to anything in the original: count as "extra element", even if they aren't clearly recognizable as any musical symbol. This applies, for example, to little squiggles superimposed on beams that the OMR program might have intended to be slurs.

3. Dotted slurs in the original: ignore; also ignore anything that evidently results from them, e.g., staccato or augmentation dots. The rationale is the assumption that originals shouldn't have any dotted/dashed slurs. This is reasonable because recognizing curved dotted/dashed lines seems to be well beyond the state of the art of optical pattern recognition.

4. Missed (or added) fingerings: ignore rather than counting as missing symbols (or extra elements). The rationale is that—for most purposes—fingerings aren't very important; we could count them as, say, 1/20 as much as a missing note, but it's not worth the effort.

5. Missed (or added) accents, articulation marks, etc.: ignore rather than counting as missing symbols (or extra elements). The rationale is that—for most purposes—they aren't very important; we could count them as, say, 1/10 as much as a missing note, but it's not worth the effort.

6. Key signature only partly correct: count as misinterpreted symbol. For example, a key signature of three flats interpreted as one flat and two sharps is one misinterpreted symbol; likewise if it's interpreted as five flats and three sharps (unlikely though that is).

7. Text string only partly correct: count as misinterpreted symbol. This applies to cases of extra, missing, and wrong characters in the string. In multi-line blocks, consider each line as a separate string.

8. If a note's pitch is wrong for any reason, including missing or extra accidentals, count it as just pitch error. Do not count these situations as missing symbols or extra elements. *Exceptions*: (a) if the pitch is wrong because of a missing or extra octave sign, count the octave sign itself as a missing symbol or extra element; do not count the pitch as wrong for the affected notes. (b) If a missing or extra accidental results in several following notes having the wrong pitch, count only the first note as having wrong pitch. If a missing or extra accidental results in *no* notes having wrong pitch, ignore it; the rationale is this isn't important enough to bother with. (c) If a missing, extra or wrong clef results in several following notes having the wrong pitch, count only the clef as a misinterpreted symbol.

9. If a note's duration is wrong for any reason, including missing or extra augmentation dots, count it as just duration error. Do not count these situations as missing symbols or extra elements. *Exceptions*: if the duration is wrong because of a missing or extra tuplet sign, count the tuplet sign itself as a missing symbol or extra element; do not count the duration as wrong for the affected notes.

10. Missing extender (dashed line) following text: ignore. However, if pieces of the extender turn into accent marks, ornaments, etc., count them as misinterpreted symbol.

11. "Notes" are just that; treat rests and grace notes as "other symbols", not as notes.

12. Note with stem recognized as barline: count as both misinterpretation and missing note. There's certainly misinterpretation here, but missing a note is too important not to count it as such. Similarly, barline recognized as note with stem counts as both misinterpretation and extra note.

A Small Corpus of Music Pages

The music pages described in the table below make up the "SFStudy" test corpus used in Byrd & Schindele (2006, 2007). Corpus B of Bugge et al (2011) is a slight variation, omitting test page 23 but adding two earlier versions of page 24. (The inclusion of the earlier versions of page 24 was unintentional.) 300 dpi images of the Corpus B pages can be downloaded from http://code.google.com/p/omr-errorcorrection/ as corpusB.zip . The table itself is based on Table 1 of Byrd & Schindele (2007), adding image quality flaws according to the ratings of Bugge et al., but with higher numbers indicating more flaws (in Bugge et al., they indicate fewer flaws).

The 24 pages were chosen in a principled way, based on the factors described by Byrd (2013):

- the researchers' opinion of factors likely to influence the performance of OMR, plus
- a range of music that is of interest in its own right, mostly in well-known and respected editions, plus

- tests of individual music-notation symbols in very straightforward contexts (e.g., the OMR "Quick-Test" of Ng & Jones, 2003).

Test page no.	Cmplx. level	Image quality flaws	Title	Catalog or other no.	Publ. date	Display page no.	Edition or Source
1	1x	1	OMR Quick-Test	1	Cr.2003	*	IMN Web site
2	1x	1	OMR Quick-Test	2	Cr.2003	*	IMN Web site
3	1x	1	OMR Quick-Test	3	Cr.2003	*	IMN Web site
4	1x	1	Level1OMRTest1		Cr.2005	*	DAB using Ngale
5	1x	1	AltoClefAndTie		Cr.2005	*	MS using Finale
6	1x	1	CourtesyAccsAndKSCancels		Cr.2005	*	MS using Finale
7	1	3	Bach: Cello Suite no.1 in G, Prelude	BWV 1007	1950	4	Barenreiter/ Wenzinger
8	1	2	Bach: Cello Suite no.1 in G, Prelude	BWV 1007	1967	2	Breitkopf/ Klengel
9	1	3	Bach: Cello Suite no.3 in C, Prelude	BWV 1009	1950	16	Barenreiter/ Wenzinger
10	1	3	Bach: Cello Suite no.3 in C, Prelude	BWV 1009	1967	14	Breitkopf/ Klengel
11	1	2	Bach: Violin Partita no. 2 in d, Gigue	BWV 1004	1981	53	Schott/Szeryng
12	1	2	Telemann: Flute Fantasia no. 7 in D, Alla francese		1969	14	Schirmer/Moyse
13	1	2	Haydn: Qtet Op. 71 #3, Menuet, viola part	H.III:71	1978	7	Doblinger
14	1	2	Haydn: Qtet Op. 76 #5, I, cello part	H.III:79	1984	2	Doblinger
15	1	3	Beethoven: Trio, I, cello part	Op. 3 #1	1950-65	3	Peters/Herrmann
16	1	2	Schumann: Fantasiestücke, clarinet part	Op. 73	1986	3	Henle
17	1	3	Mozart: Quartet for Flute & Strings in D, I, flute	K. 285	1954	9	Peters
18	1	2	Mozart: Quartet for Flute & Strings in A, I, cello	K. 298	1954	10	Peters
19	1	3	Bach: Cello Suite no.1 in G, Prelude	BWV 1007	1879	59/pt	Bach Gesellschaft
20	1	3	Bach: Cello Suite no.3 in C, Prelude	BWV 1009	1879	68/pt	Bach Gesellschaft
21	2	2	Mozart: Piano Sonata no. 13 in Bb, I	K. 333	1915	177	Durand/ Saint-Saens
22	4	2	Ravel: Sonatine for Piano, I		1905	1	D. & F.
23	4	2	Ravel: Sonatine for Piano, I		1905	2	D. & F.
24	3x–4x	1	QuestionableSymbols		Cr. 2005	1	MS using Finale

Publication date: Cr. = creation date for unpublished items.

Display page no.: the printed page number in the original. "/pt" means only part of the page.

Cmplx. level: complexity level of the music, as defined above (higher number = more complex).

Image quality flaws: graphical quality of the bitmap reproduction of the original page, as defined above (higher number = more flawed).

A Resource for the Future: IMSLP

The advent of IMSLP, the International Music Score Library Project, changes the OMR landscape significantly. As their Web site (IMSLP, 2012) says, IMSLP's Petrucci Music Library is "a community-built library of public domain sheet music" containing an "extensive collection of original scores scanned to PDF". The collection is indeed extensive: at the moment, it contains over 220,000 scores of over 60,000 works. Thus, IMSLP has great promise as a source of material, both for OMR evaluation and for practical use in converting music to symbolic form via OMR.

Not surprisingly, the scans vary considerably in quality. IMLSP formerly displayed ratings of scan quality for their offerings, but, regrettably, it no longer does so. It does show ratings of one to five stars, apparently consensus ratings by users. But are these users rating the scan quality, the edition, or the music? We have not even been able to find any information as to what these ratings are *intended* to mean, though they may have a significant correlation with scan quality.

Clearly, it would be ideal if scan quality ratings were supplied by trusted users adhering to a set of fixed, clearly described principles, and preferably augmented with ratings describing the complexity of the scores. But even without such ratings, we believe that IMSLP is an extraordinarily valuable resource.

Conclusions

Some months ago, one of us received a message from the president of the company that develops one of the best-known commercial OMR systems, arguing that comments we had made in print about their company's product were unfair, and claiming that their system is the most accurate available and that its accuracy has improved tremendously over the years. This is obviously an important question for his company, and he may well be correct, but there's no way to know! This unfortunate situation need not and should not be allowed to persist. We believe that the advent of a real testbed—even one with a small, static collection of music pages which is updated infrequently—would substantially facilitate advances in the state of the art.

The obvious environment in which to implement an OMR testbed is MIREX and its annual tests, held in conjunction with ISMIR. While the "well-thought-out test collections and evaluation metrics" materials such a track needs do not really exist, we believe the materials and ideas this paper presents are a large step in that direction.

Acknowledgements

The definitions of quality levels and images illustrating them are reprinted with kind permission of our collaborators Brian Søborg Mathiasen, Esben Paul Bugge, and Kim Lundsteen Juncher.

Ichiro Fujinaga's comments on an early version of this paper were extremely helpful, as were numerous conversations with Chris Raphael over a period of years.

Most of Donald Byrd's work on OMR was supported by the Andrew W. Mellon Foundation.

References

Bainbridge, David, & Bell, Tim (2001). The Challenge of Optical Music Recognition. *Computers and the Humanities* 35(2), pp. 95–121.

Bugge, Esben Paul; Juncher, Kim Lundsteen; Mathiasen, Brian Søborg; & Simonsen, Jakob Grue (2011). Using Sequence Alignment and Voting to Improve Optical Music Recognition from Multiple Recognizers. *In Proceedings of the 12th International Society for Music Information Retrieval Conference* (ISMIR 2011), pp. 405–410.

Byrd, Donald (1984). *Music Notation by Computer* (doctoral dissertation, Computer Science Dept., Indiana University). Ann Arbor, Michigan: UMI ProQuest (order no. 8506091); also available from www.npcimaging.com. Retrieved (in scanned form) February 20, 2013, from the World Wide Web: http://www.informatics.indiana.edu/donbyrd/Papers/DonDissScanned.pdf

Byrd, Donald (2008). Music-Notation Complexity Levels for MeTAMuSE OMR Tests. MeTAMuSE/IU project working paper. Retrieved February 20, 2013, from the World Wide Web: http://www.informatics.indiana.edu/donbyrd/MROMR2010Pap/OMRNotationComplexityDefs.txt

Byrd, Donald (2013). Guidelines for Choosing OMR Test Pages (and factors affecting OMR accuracy). Revised from a MeTAMuSE/IU project working paper. Available at http://www.informatics.indiana.edu/donbyrd/OMRTestbed/Guidelines4OMRTestPages_2013.txt

Byrd, Donald, & Schindele, Megan (2006). Prospects for Improving OMR with Multiple Recognizers. In *Proceedings of the 7th International Conference on Music Information Retrieval* (ISMIR 2006), Victoria, Canada, pp. 41–46.

Byrd, Donald, & Schindele, Megan (2007). Prospects for Improving OMR with Multiple Recognizers, revised and expanded version. Retrieved February 20, 2013, from the World Wide Web: http://www.informatics.indiana.edu/donbyrd/MROMRPap

Byrd, Donald, Guerin, William, Schindele, Megan, & Knopke, Ian (2010). OMR Evaluation and Prospects for Improved OMR via Multiple Recognizers. Retrieved February 20, 2013, from the World Wide Web: http://www.informatics.indiana.edu/donbyrd/MROMR2010Pap/OMREvaluation+Prospects4MROMR.doc

Droettboom, Michael, & Fujinaga, Ichiro (2004). Micro-level groundtruthing environment for OMR. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004),* Barcelona, Spain, pp. 497–500.

Jones, G., Ong, B., Bruno, I., & Ng, K. (2008). Optical music imaging: music document digitisation, recognition, evaluation, and restoration. In: *Interactive Multimedia Music Technologies*, pp. 50–79. Hershey: IGI Global.

Ng, Kia C., & Jones, A. (2003). A Quick-Test for Optical Music Recognition Systems. *2nd MUSICNETWORK Open Workshop,* Workshop on Optical Music Recognition System, Leeds, September 2003.

NIST (2012). Text REtrieval Conference (TREC). Retrieved February 20, 2013, from the World Wide Web: http://trec.nist.gov/

Rebelo, Ana; Fujinaga, Ichiro; Paszkiewicz, Filipe; Marcal, Andre R. S.; Guedes, Carlos; & Cardoso, Jaime S. (2012). Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval* (March 2012), pp. 1–18.

Reed, K. Todd (1995). *Optical Music Recognition.* M. Sc. thesis, Dept. of Computer Science, University of Calgary.

Selfridge-Field, Eleanor, Carter, Nicholas, and others (1994). Optical Recognition: A Survey of Current Work; An Interactive System; Recognition Problems; The Issue of Practicality. In Hewlett, W., & Selfridge-Field, E. (Eds.), *Computing in Musicology,* vol. 9, pp. 107–166.