The “OpenScore String Quartet” Corpus

Mark Gotham
Durham University
Durham, UK

Maureen Redbond
Independent
UK

Bruno Bower
Imperial College London
London, UK

Peter Jonas
MuseScore.com
UK

ABSTRACT

The “OpenScore String Quartet” Corpus is a new dataset of historic works for string quartets, encoded by a dedicated team of volunteers, and released freely for all use cases (CC0). In creating this corpus, we built on the experience amassed during the ‘OpenScore Lieder Corpus’ (Gotham and Jonas, MEC 2021), however, the quartets presented some additional challenges including the need for more significant editorial intervention. Here we report on the size and contents of the corpus (more than 100 full quartets by over 40 composers), we discuss the editorial-musicological aspects of producing modern playing scores from ambiguous or incomplete source material, and we suggest some prospective use cases for this dataset in music information retrieval (MIR).

CCS CONCEPTS

• Applied computing → Sound and music computing; **• Information systems → Music retrieval**; **• General and reference → Validation**; **Reliability**; **Reference works**; **• Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

Digital Music Library, Corpus study, Music Information Retrieval, Musical Scores, String Quartet, Strings

ACM Reference Format:
Mark Gotham, Maureen Redbond, Bruno Bower, and Peter Jonas. 2023. The “OpenScore String Quartet” Corpus. In *10th International Conference on Digital Libraries for Musicology (DLfM 2023), November 10, 2023, Milan, Italy.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3625135.3625155

1 INTRODUCTION AND HISTORY

‘OpenScore’ began with a 2017 initiative to encode sheet music chosen by sponsors of a crowdfunding campaign,[1] but the effort took off in earnest in 2018 with the launch of a (then ‘satellite’) project focussing specifically on 19th-century songs. The ‘OpenScore Lieder Corpus’ (as it became known) now includes over 1,300 songs in a range of languages on MuseScore.com and also exists in a GitHub.com mirror specifically for corpus study/MIR research [8].

The satellite did not seek crowd-*funding*,[2] but it did benefit from the *crowd-sourced contributions* of a gradually evolving team of transcribers, reviewers, and managers. This is axiomatic to ‘OpenScore X Corpus’ efforts: not only to *serve* a range of communities (musicians, academics, others), but also to *engage* those groups in the creation of the corpus [6].

We (another generation of that evolving team) now announce a new OpenScore corpus effort specifically for string quartets. This time, we are able to present at the same time both the public-facing collection on MuseScore.com and the dataset on GitHub for use by the MIR community. The latter is downloadable in bulk (unlike the collection on musescore.com, which it mirrors), and comes with curated, linked metadata as described below (§2).

In some respects, we have been able to build on lessons learned from the lieder; in other cases, the quartets presented new challenges. In particular, our ongoing commitment to producing a diverse corpus and to rediscovering forgotten works led in this case to an ever more disparate range of sources with very variable editorial quality. This, in turn, meant that it was not possible to avoid editorial intervention to the same degree as for the lieder, where only occasional interventions were required. Where possible, we still seek to keep such interventions to a minimum, but the range in quality of source material meant that more significant changes were required more often simply to make the quartet scores usable.

Here we report on the creation of the corpus of scores; document the processes, issues, and editorial decisions involved; and expand on some notable and illustrative cases. We are mindful that this undertaking is interdisciplinary in nature and that serving a range of interests across the musical and computational communities is complex. As discussed, we provide the musescore.com “version of record” to serve most musical use-cases, and for the GitHub “mirror” to serve as the structured datasets for computational uses.

We also intend for the documentation of our decision-making to serve a range of users, from satisfying professional editors’ standards, to providing some insight for those not familiar with this work just how complex it can be to work with older printed material and composer autographs, and how much extensive, subjective interpretation this necessarily involves. In short, we clarify why this kind of data can almost never be considered in terms of single, fixed “ground truth”.

2 CORPUS OVERVIEW

In this overview of the corpus, we seek to give a broad sense of its contents and to identify the attributes of relatively ‘typical’ entries, but also to draw special attention to some notable outliers which can be highly valuable for comparative MIR tasks like clustering by style.

[1] This was initiated by the then owners of MuseScore (Bonte, Froment and Schweer, before subsequent sale of the company), and run by Peter Jonas, who is the only point of continuity from that initial effort. There is no academic write up of that first effort, though the flagship website is still up at https://openscore.cc/.

[2] It has received small grants from academic funding bodies. Please see the acknowledgements section.

DLfM 2023, November 10, 2023, Milan, Italy
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0833-6/23/11.
https://doi.org/10.1145/3625135.3625155

2.1 What counts?

As with the lieder corpus (and indeed many corpora), we must first establish some ground rules for what ‘counts’. The term ‘String quartet’ refers both to the *ensemble* (two violins, one viola, one ’cello) and to a *piece* written for that group. This ensemble is said to have emerged in the mid-18th century (particularly in the works of Haydn) out of comparable baroque equivalents (which, however, typically included a continuo keyboard part). More specifically, while composers still write for ‘string quartet’ today, the term sometimes implicitly refers to a more exclusively classical-period mentality, including (for instance) the assumption that the work is in many (usually four) movements, as discussed further below.

This question over what ‘counts’ is required for our lieder and quartet corpora, and any like them for which clear limits of which are not given. This contrasts with many (perhaps most) corpus-building efforts in music which tend to focus on a clearly defined set for which the membership criteria are clear.

We see a value in our less-strictly defined corpora exactly because they are more varied. For instance, [10] shows one context in which such a corpus apparently leads to an improved performance on computational tasks: the data derived from the (highly varied) OpenScore lieder corpus performed better for the task at hand on the Beethoven sonatas than the equivalent data derived from the Beethoven sonatas themselves. At a minimum, these corpora clearly engage more positively with the current efforts to diversify the music we encounter (in performance, teaching, and research), by emphasising a wider range than the many corpora dedicated to one ‘great’ composer, to the exclusion of all others.

2.2 Composers, Works, Length, Sets

As with the lieder, we have sought here to represent a wide range of composers and works, including lesser known and never-published pieces alongside the famous sets by Haydn, Mozart, and others. In proportionately more cases than for the lieder, this involved a modest amount of research, tracking down hard-to-access publications and working in tandem with IMSLP to develop their holdings alongside ours (and providing direct links to the corresponding IMSLP edition for every quartet). In total, the corpus now comprises over:

- 350 movements, across
- 100 multi-movement quartets (a.k.a. ‘sets’), from a total of
- 40 composers.

These headline figures are inevitably lower than the equivalent metrics from the lieder corpus given the more substantial nature of the works. While some song cycles are of a comparable length to whole string quartets, individual songs are typically much shorter than quartet movements.

Most of the quartets in this corpus are multi-movement works, though we do also include some of the notable single-movement works for string quartet including:

- Beethoven’s *Grosse Fuge*,
- Schubert’s *Quartetsatz*, and
- Wolf’s *Italian Serenade*.

The multi-movement works tend to follow the “sonata” pattern that was typical at the time for string quartets (as well as sonatas, symphonies, and more). This semi-standardised practice governed aspects of the structure, tempo, and tone in each of those (usually four) movements according to certain prototypes (e.g., a slow second movement). That said, there is naturally some variety with more or fewer movements. Notable here is the Joseph Boulogne (Chevalier de Saint-Georges) Op.1 collection with 2 movements in each of the 6 quartets.

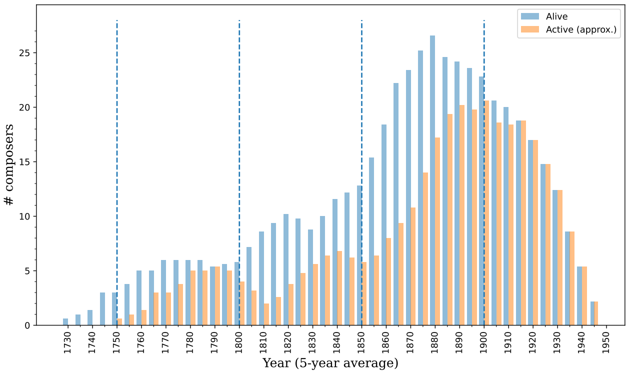


Figure 1: The dates of composers both in terms of the full lifespan, and with an approximate value for ‘active’ years, excluding the first 20 of life.

2.3 Time & Geography

Temporally, the works span c.150 years from the ostensible origins of ‘the string quartet’ to the current limits of public domain. The earliest set included is Haydn’s genre-defined Op.1 (1762) and the latest is Janáček’s second (String Quartet No.2 “Intimate Letters”, 1928).[3] See figure 1 for a broad overview.

Composer nationality is complicated, given that many individuals will have a mixture of nationalities in their heritage at birth and then proceed to settle in different places.[4] Nevertheless, and without indulging too much of a digression, we consider it worth recording this data in some form as part of keeping track of the corpus’ range. Our data records 13 distinct nationalities currently included, counting Austria and Germany separately. All but one of these nationalities is European: the corpus includes the Quartet in B minor (first publication 1896) by Venezuelan composer María Teresa Gertrudis de Jesús Carreño García (1853–1917).

3 VERSION OF RECORD & CORPUS MIRROR; DATASET, DIRECTORY, AND CODE BASE

As for the lieder, the MuseScore.com resource stands as our public-facing version of record, and this is complemented by a ‘mirror’ on GitHub.com. Updates to GitHub run occasionally so they are sometimes “behind” the version of record but are a complete and up-to-date mirror of the source at the time of each update. Apart from

[3] This is a few years later than Alban Berg’s Op.3 which incidentally was probably the most ‘dense’ and ‘difficult’ to transcribe.

[4] For example, the nationality given for Joseph Bologne (Chevalier de Saint-Georges) is ‘French’. Bologne moved to France aged 7 and remained there for most of his life, though he was born in Guadeloupe.

ease of download at scale, the corpus mirror offers a slightly expanded set of files and metadata in a carefully constructed directory structure specifically to assist MIR research.

3.1 Scores

The corpus centres on a sub-folder called 'scores' with files arranged in the structure: `<composer>/<quartet>/<score>`. Specifically,

- `<composer>` gives the composer's name in the form `<Last,_First_Second>`.
- `<quartet>` provides the opus number, name, or other identifier for the multi-movement work.

Score files within each quartet directory are named in the format `<id>.<format>`. The format is `mscx`: an uncompressed MuseScore file. The `<id>` takes the form of `sq` and a unique number assigned to each quartet. That `<id>` also provides a direct link to the public-facing version of the score as hosted on MuseScore.com by appending it to any of:

(1) https://musescore.com/openscore-string-quartets/scores/
(2) https://musescore.com/user/37221589/scores/
(3) https://musescore.com/score/ (note singular *score*).

For example, for the score with id 7284122 (Haydn's "Lark" Quartet), the following currently work:

(1) https://musescore.com/openscore-string-quartets/scores/7284122
(2) https://musescore.com/user/37221589/scores/7284122
(3) https://musescore.com/score/7284122

These URLs all redirect to the same page, on which the score is displayed. We list these multiple options in case the MuseScore.com URL scheme changes in the future (which would be beyond our control). The GitHub mirror, by contrast, *is* under our control, and we do not anticipate making significant changes to its current structure. We may correct directory names if any mistakes are identified; even in that case, the `<id>` will remain intact.

MuseScore files are convertible to other formats (MusicXML, MIDI, PDF) individually or in bulk using a plugin in the app or via the command line. We provide a contents file in the format required for the command line option and also a script for updating this file. We decided against including multiple formats in the repository given this ease of conversion and a desire to keep the overall repository to a manageable size.

We consider the scores to cover the core provision and we elect to avoid overloading this repository with anything else. Some users may wish to consult another repository, the 'When in Rome' corpus,[5] which includes another mirror of the score collection, in the context of providing harmonic analyses for many of the quartets (alongside other corpora). The set of manual analyses includes both new work and conversions, for instance, of [13]'s MTG dataset for analyses of the Haydn Op.20 set.

3.2 Data

As with the lieder corpus, the metadata reported here is stored in a dedicated folder (`data/`), separate from the corpus itself (`scores/`), and encoded in both TSV format and StrictYAML:[6] a simplified variant of YAML that preserves the order of keys within object mappings.

There are separate files for `composers`, `corpus`, `scores`, and `sets`, organised as consistently as possible with the lieder (as described in both README's and in [8]). This provides between-corpus consistency. We currently continue to include both `scores` and `sets` despite the fact that the quartet encodings are not separated by movements. This is partly for the consistency and partly in case there is a call to include the separate parts files (violin 1, ...) which are currently hosted on musescore.com, but not on the mirror as they are of no clear use for computational work. In the case of transcription from parts with separate IMSLP edition numbers (see §4.3), the `scores` file records a single IMLSP ID for consistency of data type (always a string, never a list), and full information is available in the coordination spreadsheet.

Once again, we emphasise the importance of the linked open data (LOD) provided by the external links (to IMSLP, Wikipedia, and Wikidata). Comments made in [6] still hold, and see [15] for a recent report on the use of Wikidata in Digital Humanities projects.

3.3 Code: Movements, Texture, Syntax

Apart from the corpus management/metadata code, we also provide a small amount of additional functionality that has proven useful for this corpus building effort and may serve much wider use cases.

3.3.1 *Movement Splitting.*

We decided to encode entire, multi-movement quartets in one file, rather than separating by movement. This greatly assists with the preparation of parts, which are the primary version of this material that musicians will play from.[7] There are detractions, however: it is a point of divergence from the lieder, and inconvenient for some other uses like in MIR research.

In short, we put musicians first in this case, though we also compensate by providing code for splitting the whole quartet into separate movements automatically. As this involves going via MusicXML, some of the style is lost, but we anticipate that MIR tasks for which separating the movement is essential will find the style information less (or un-)important.

3.3.2 *Annotation Extraction and Checks.*

Although spreadsheet entry for the editorial notes was mostly manual, we also provide a script for automatically extracting that data: the comment, instrument, and bar number. This could easily be extended in future work, for instance with a tag to indicate that a comment applies to all parts rather than only the one to which it applies, and perhaps for assigning IDs so comments can be explicitly linked. Additional code also checks the spreadsheet for consistency of string formatting (as discussed in §4.2).

3.3.3 *Survey of Textural 'Homorhythmicity'.*

One asset of the string quartet repertoire for computational tasks is textural clarity. Small challenges notwithstanding, it is easier to describe the textural role of all instruments in a string quartet than to do the same for piano textures, especially automatically.[8] Figure 2 provides a sense of how this summarises texture across a work of any length, and the GitHub repository includes the code used to generate it, complete with full documentation.

[5] https://GitHub.com/MarkGotham/When-in-Rome.
[6] https://GitHub.com/crdoconnor/strictyaml; https://pypi.org/project/strictyaml/.
[7] E.g., note IMSLP's policy against separate files for individual part-movements.
[8] See §5 and [4] on the specific challenges of *piano* textures.

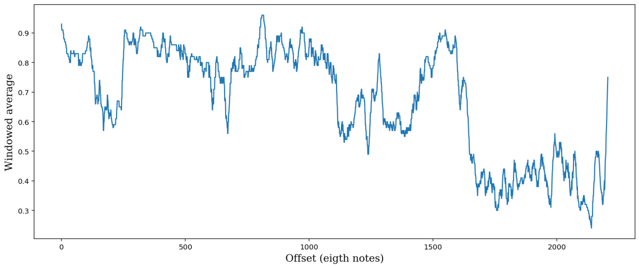


Figure 2: Relative homorhythmicity (y) against time (x), showing lower values for a fugal passage towards the end.

4 EDITION

As mentioned, the quartets have presented us with additional challenges in terms of original source material. Professionally published editions of the quartet in full score format exist in some cases. Others exist only in unpublished (but nonetheless perfectly clear) manuscript copies. More problematic still were works that survive in incomplete or other atypical formats including the following:

1. Parts alone (see §4.3);
2. Arrangement for piano duet (a special case, see §4.4)
3. Incomplete manuscript (see §4.5).

We discuss examples of each of these in the given sections below, after first dealing with preliminary comments on the general editorial approach taken overall.

4.1 General Editorial "Policy"

It is worth noting that human errors requiring correction appear in every type of source material. For the printed materials, these might originate either from the manuscript on which they were based or the typesetter who prepared it for printing. In the manuscript sources, these emerge in much the same way as typos do in written texts, due to quick working or a reliance on conventions and an expectation that someone else will fill in the blanks later. These issues are more common in lesser-known works as they typically survive in fewer and/or lower quality sources (i.e., lacking professional editorial work and extensive dissemination). We discuss several of these source and issues in the subsections below.

Our overall editorial approach focuses primarily on the performers who will engage with this music, ensuring that the scores and parts are as clean and ready-to-use as possible (see [11]). Errors and inconsistencies take precious rehearsal time to address and might dissuade performers from using other scores from the corpus. Composers with limited reputations are especially likely to suffer under these circumstances, with the gravitation towards better-known composers (potentially compounded by unconscious bias) putting performers off engaging with their music much more quickly when they encounter obstacles. Special care must be taken with these scores and parts (doubly so given the typically higher rate of errors noted above), to ensure that they make the best possible first impression on those who might bring them to the concert platform.

We correct for two broadly separable kinds of issues: *errors* and *inconsistencies*. We understand *errors* to encompass anything which can be unequivocally identified as incorrect. Examples include:

- *incorrect* rhythms, e.g., the duration of notes within a bar does not add up to the correct duration for the bar overall.
- *missing* elements, e.g., an instruction to start a *pizzicato* (plucked) lacks the paired marking for the return to use of the bow ("arco").

By contrast, we understand *inconsistencies* to be a more subjective matter. As an example, the notation may give a clear indication of what is expected in a particular passage, but there might be conflicting information in an equivalent passage elsewhere. Sometimes there are musical reasons to justify these discrepancies, but often it is more likely that the two passages ought to match each other. At this point a decision must be made as to which version is correct, possibly incorporating features from both versions. (Examples follow below). Matters of consistency include ensuring that all four parts have dynamics and that these dynamics align (e.g., all moving from ***p*** to ***f*** in the same place) unless there is a compelling reason not to. Similarly, articulations usually ought to align in this way, as should reprises of the same material at a later point in the piece.[9]

These *inconsistent* aspects could potentially be left as they appear in the manuscript since they are not self-evidently *incorrect* but changing them produces scores that present-day performers will more readily accept and be able to use. If we have done our work effectively, they might even be closer to what the composer intended.

4.2 Documenting Editorial Changes

In making these editorial interventions, we distinguish between works which require only occasional, minor edits, and those that need many and/or substantial revisions. For minor interventions (scores with a couple of rhythms and dynamic edits) we simply make the change and make the score available as it is. These scores can be transcribed and reviewed by anyone on the team. By contrast, we consider movements requiring many and/or substantial or complex interventions to require documentation. Those cases were assigned to a smaller set of contributors trained to document the changes with comprehensive editorial notes. These notes were intended to help both transcriber and reviewer to keep track during the review process, and also to assist future users interested in the decisions made. As these scores are in an editable format, any user could go on to make further changes, effectively creating their own edition.

These editorial notes are now publicly available.[10] The notes are set out with each piece on a separate tab. Each such work-tab began life as a copy of the initial template (also provided in the first tab). This ensures that all contributors present their work in the same format. Further consistency of notation includes consistent (non-case-sensitive) tags for the instruments: 'v1' for violin I, 'v2' for violin II, 'va' for viola, 'vc' for cello, any combination thereof (e.g., 'va,vc'), and 'all' for all parts. We ask contributors to adopt this practice, we provide examples of existing work, and we check for consistency both automatically and manually during review (see the 'Code' section, §3.3). This makes it possible to search for topics of interest across the notes.

[9] To be clear we are talking about within-score consistency here. Any attempt on our part towards corpus-wide consistency (i.e., a kind of house style) is entirely separate and obviously not the responsibility of the composers or editors of the source material.
[10] https://bit.ly/StringQuartetsEdits

Aside from the basic details concerning *what* has been changed between source and edition, we obviously note *where* this change occurred (which instrumental part, bar, and beat), and also leave room for a brief rationale behind the editorial intervention in prose – the *why*.

For the avoidance of doubt, we note the *where* using the total bar “count” in which every successive entry has a number (including anacrusis, 1st/2nd time bars are more). This is easy for MuseScore users as it is shown in the “toolbar” at the bottom of the screen. It is also useful for clear comparison as other, more musical counts of the bars are liable to change during review. For instance, for the ‘bar number’ shown on the printed score we follow the usual editorial practice of numbering incomplete first (anacrustic) bars 0, but MuseScore does not do this by default. For this reason among many others, the score’s bar numbering may need to change at a late stage in the review process, meaning that those numbers for (off-score) comments also need to be updated. This apparently small issue can create significant issues as discussed in another contribution to this year’s DLfM: [7].

We consider the inclusion of the free-prose *why* explanation to be potentially significant in terms of the computation of humanistic data. We recognise the benefits of concision and type-specified, computer retrievable standards for the entries. For instance, bar count is always an `int`, beat is either a `float` or an `int`. That said, sometimes the most responsible course is to support full prose explanation that cannot be further reduced to succinct typing.

All contributors can see the work that has been done by others, this provides each with insight into the kinds of changes that can be made and how they can be justified. One volunteer spoke of the valuable learning experience of watching a reviewer document further changes to their first effort in real-time. As discussed in [9], this kind of ‘added value’ for contributors is a significant part of the motivation for OpenScore projects.

4.3 Parts Alone: Bologne’s Op.1 Set

String quartets typically perform from parts (one per player). If they have a score, it is usually simply on hand for reference in rehearsal.[11] Not infrequently, this leads to historical sources surviving in either score or parts but not both, particularly for composers who did not enjoy a great deal of editorial-scholarly attention during and after their lifetime.

Works surviving only in parts format often present several additional challenges that would be less likely to emerge in a score setting, even when they are professionally published. Most such cases in the present corpus exhibit sufficient discrepancy to warrant editorial intervention and inclusion in the editorial document.[12]

At the more straightforward end of the spectrum, our editorial interventions included simple, obvious matters like making sure hairpin start/end positions align, matching articulations across phrases and parts, and calibrating the use of dynamics (removing some that are redundant, and adding some where needed, e.g., at the start of a new phrase after a rest).

The quartets by Joseph Bologne, Chevalier de Saint-Georges required particularly active and creative interventions across most musical parameters. Although the parts were printed, their publication date of c.1773 puts them at a relatively early phase of music publishing with less of the standardised and professionalised practices that would emerge later.[13] The notes and phrases broadly line up (though even the repeat structure is not initially obvious) but the general discrepancies between the four parts are considerable. It seems likely that different people were involved in the typesetting process. This was a common time-saving measure of that less-mechanised era, since using more people was the only way to speed up production. These issues are especially prominent for matters where there is no consensus even among scholars of 18th-century performance practice. For example, both dots and strokes are used in the Bologne parts, apparently to indicate staccato. It is unknown whether these encode different meanings, or are interchangeable.[14] Even some unison passages contained different articulations across the four parts.

It is highly likely that these should be converted to matching articulation (one type in all parts), but there is no clear way of determining which is preferable. The only way to proceed with editing was to rely on a more personal sense of what makes the most musical sense, and to use this consistently across not only all four parts, but also all instances of the figure. Present-day performers would otherwise have regarded such passages as nonsensical.

The sources were no more internally consistent within individual parts: different articulations and dynamics appear even where every other musical parameter indicates an exact repeat of an earlier passage. Some of these changes could be justified for the sake of variety: it may be that Saint-Georges wanted the music to sound different on a subsequent presentation of the material. Once again, it is impossible to firmly distinguish these cases from those resulting from the transcription errors clearly evident elsewhere.

Cases of incorrect notes proved especially tricky to solve: although unambiguously ‘errors’ insofar as they clashed with the surrounding harmony, choosing the correct alternative sometimes relied on sequences of creative logic that overlaps with the act of composition itself; in such cases, are we editing or rewriting? As an example, consider figure 3 and the two versions of Bologne’s Quartet Op.1, No.1, movement 2.[15] In the first version, the A in the 2nd violin at the cadence (downbeat of the second bar shown) is a clear clash with the G in every other part. Simply changing to a G, though, produces a new problem in the form of a slur over repeated notes. Short slurs in string writing typically indicate bow direction change (up/down), and it is only possible to play a repeated note

[11] Conventions vary by ensemble. Today, each singer in a choir typically performs from a full score (all parts shown) while orchestral musicians play from only their own part (the conductor is the only one to read from a full score). There are always exceptions to this, and trends change with time. For example, in early choral settings (c.1600 and earlier), it was common to perform from ‘part books’. Returning to our medium and present day, quartets specialising in contemporary music now do often work from scores. Detractions to this (such as having more pages to turn) are also being mitigated by modern technology (e.g., with scores on electronic devices and page turns controlled by foot pedals).

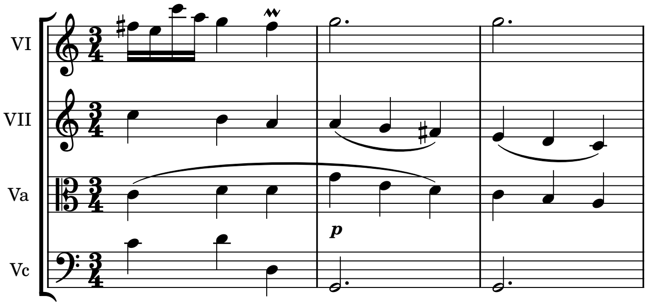
[12] Bedřich Smetana’s String Quartet No.1 provides a notable exception which was transcribed from parts and presented no significant editorial challenges.

[13] For a brief, image-rich, online introduction to some of the practices in music printing history, see https://musicprintinghistory.org/.

[14] For more specifically on the subject of dots *vs* strokes see [2] and [3, pp.200-223].

[15] For those comparing with the editing records, please note that the measure numbering there follows MuseScore’s internal convention (shown in the bottom of all score-editing windows). In this case, the measure numbers are 139–142 (the count does not start again for the 2nd movement). We elected to use this number in the editing record to prevent confusion, though clearly we adjusted measure numbers for the print-ready versions of the performing editions themselves. For more on this, see [7].

Original from parts, before edition:



Final, after edition:

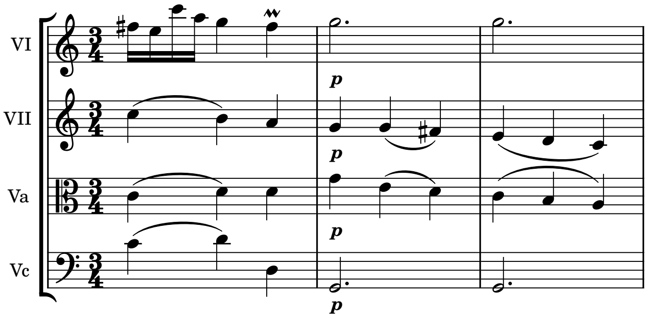


Figure 3: Extract from Joseph Bologne (Chevalier de Saint-Georges) String Quartet Op.1, No.1, movement 2, before and after edition.

with the same bow direction using special techniques that need to be specifically indicated and are rare in this style/era. Changing the slur so that it begins on the second crotchet of the bar then implies that the viola ought to do the same, since the two parts are playing the same material. That again requires a change from what is printed, though the change is arguably warranted given the repeated note slur there too (va, first bar shown), and the extent of inconsistency across parts here. We are content that the editorial result improves this situation by removing errors and discrepancies, though we note that this result is a substantial change from the source. Even though each step made logical sense from an editorial perspective, the result can only be justified on more subjective grounds, namely that the new version is musically satisfying and will be more acceptable for present-day performers.[16]

4.4 Piano duet: Mayer, Quartet in D major

Emilie Mayer's quartet in D major is an exceptional case where the original seems to be lost and we have source material not in score, nor (string quartet) parts, but in an arrangement of the work for piano duet by the composer.[17] The piano duet medium involves two players at one instrument. It is sometimes called 'piano four hands' or similar for this reason. The layout is (both typically and in this case) on paired pages in landscape format, which lies open on the piano and with the two (bass a.k.a *secondo* and treble a.k.a *primo*) players reading from left and right pages respectively.

Reconstructing a string quartet version from this piano duet could have potentially forced us to stray beyond edition into yet more creative activities such as 'arrangement': the conversion of music from one medium into another (or rather in this case 're-arrangement' or 'de-arrangement' back to a putative original). As with the more detailed case of the Saint-Georges editorial interventions above, this would be another leap away from transcription and would potentially open further historical-musicological questions. And while we might strive to realise the original composition as intended, we may well have to accept that this task takes us further into the realm of 'educated guesswork'. At the same time, this may be a necessary task if we are to reacquaint the world with these works.

In this case, (re)arrangement turned out to be mostly unnecessary, since the piano duet version appears to be more of a transcription than a reimagining of the music for the new piano duet format.[18] For the most part, there is one line of music for each hand, suggesting that this is a direct transcription of the string parts:

- Violin 1 = *Primo* right hand (upper staff)
- Violin 2 = *Primo* left hand (lower staff)
- Viola = *Secondo* right hand (upper staff)
- Cello = *Secondo* left hand (lower staff)

There were three main indicators that pointed towards a simple transcription of the original four parts. First is the fact that most of the music is in four distinct lines, with almost no chords, (one for each of the *Primo*/*Secondo* and Left/Right hand pairs). It's not a problem to imagine this as piano music, but even as set out in this keyboard format, it is more idiomatic to four string lines.

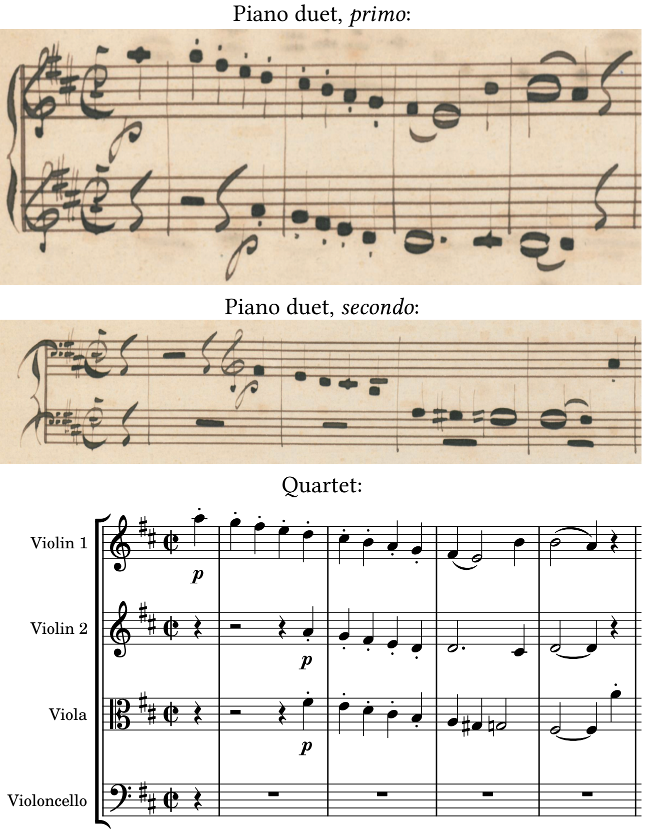
The second indicator is the overall pitch register of the *secondo* RH. While this part of a piano duet is probably the most liable to roving across treble and bass clef ranges (as in fig.4, middle), this part in the Mayer sits in a consistently higher register than normal, spending much of its time in the treble clef. In short the register is consistent with an original line in the viola's main range (i.e., in alto clef, between treble and bass). It suggests that Mayer left her viola part in its existing register for the transcription rather than adapting it to one more typical of a *secondo* RH register.

The third and final main indication of simple transcription comes in the third and fourth movements, where the *primo* LH drops out for some passages that are otherwise in unison across parts. In a simple transcription, this would be the logical consequence of the two violins playing the same notes (doubling at the unison): since the two hands of the *primo* player cannot both play the same keys, one of the hands drops out. A more pianistic solution was certainly available (e.g., putting the RH up the octave to allow space for the

[16] [14] discusses the role of "intuition" in editing music, albeit for a different era.

[17] The score describes this work as *composée est arrangée pour le piano à quatre mains par Emilie Mayer.* Here is the ISMLP copy of that source: https://imslp.org/wiki/String_Quartet_in_D_major_(Mayer,_Emilie)#IMSLP748464

[18] This is presumably for practicality of performance: Pianos were common in the homes of those with means enough to (be musically minded and) afford them; string quartets, while still somewhat "'domestic", were less accessible.


Piano duet, *primo*:

Piano duet, *secondo*:

Quartet:

Figure 4: Extract from Emilie Mayer’s String Quartet in D major in the piano duet version (IMSLP#748464) and the corresponding extract in our transcription.

LH to play the lower), but Mayer appears to have chosen to remain closer to what was probably in the original quartet.

Although the source overall certainly represents a relatively direct transcription, editorial interventions were still required to make the material idiomatic for a 19th-century string quartet. There were some moments where Mayer had no choice but to adapt the music for the new instrumental format. This includes the unison passages cited above: their role as indicators of a simple transcription also marked them out as needing a return to the probable original, with the *primo* RH music applied not only to the 1st violin music, but also copied into a doubling 2nd violin line. Moreover, while Mayer did preserve the original registers for most of the piece, there are some moments that appear to have been changed to fit the new format. This includes some:

- 7th leaps in the second movement, apparently introduced to avoid clashes between the hands.
- very high passages for the *primo* RH in the first and last movements that move the material out of the way of the other hands, but would be less idiomatic on a violin.

This is all in addition to the kinds of editorial interventions that were necessary for a manuscript source rather than a published edition, which will be discussed in more detail below (§4.5).

Finally, it is worth mentioning an exceptional case drawn from the Mayer, namely an apparent discrepancy in the *number of bars* between corresponding *secondo* and *primo* pages in the last movement. As it turned out, this was simply due to a 4-bar repeat that was marked as such in the *primo* part (with *bis* in Mayer’s hand) and written-out in full in the *secondo*, giving it four more bars. Every (joint) page turn in a piano duet provides a checkpoint for coordination. Piano duets are almost unique in offering this.[19]

4.5 Incomplete Manuscript: Röntgen-Maier Quartet in A major

Swedish violinist and composer Amanda Röntgen-Maier (1853–1894)’s quartet in A major posed the most difficulties of the works taken on so far. Again, this is not a comment on the work itself, but simply on the state of the surviving material which required very significant detective work to make any sense of, let alone to produce a performable edition. Of the four movements, we have been able to produce:

(1) An incomplete movement
(2) A complete edition with the usual level of confidence.
(3) A complete edition (as for movement 2).
(4) An incomplete movement, requiring some deduction to ascertain that material is indeed missing.

The IMSLP scan appears to bring together two sources,[20] showing that the different sections of the quartet were written on different kinds of manuscript paper. The first movement is written on large, landscape pages of 14 staves each. At the end of the 12th page of the scan, the file abruptly switches to a new kind of manuscript paper: portrait pages of 12 staves each, and showing the start of the second movement. The rest of the pages follow in this new format, containing the third and fourth movements as well. There is nothing in the rest of the file to indicate the remainder of the first movement (the material had reached part-way through the recapitulation at the point where it cuts off), so the owner of the manuscript must not have been able to locate the rest. We therefore deduce that this first movement is incomplete, at least for now as represented in this source.

Although the rest of the file consists of paper in a consistent format, and overall seems to contain three movements, the ordering of the material is very unusual. After the end of the second movement on page 20 of the scan, what follows is not the third movement, but rather 12 pages of material in 2/4 time that clearly represent part of the last movement. The third movement (in 3/4 time) then begins on page 33 of the scan and continues unbroken through to the start of the final movement on page 43.

From a purely editorial perspective, this material represented much the same issues as other manuscript sources cited above. If anything, the second and third movements were considerably cleaner and more consistent in their notation than some other manuscripts we worked with, requiring only c.40 interventions between them. The handwriting in the first movement is more hurried, and correspondingly a greater number of interventions were required here to account for the increase in human error.

[19] For more on source alignment, see [7].
[20] https://imslp.org/wiki/String_Quartet_in_A_major_(Maier%2C_Amanda)#IMSLP562387

What we could make of the last movement also needed extensive changes for consistency, but again this was in line with the usual requirements for manuscripts.

However, the unusual ordering of the second part of the scan masked a deeper problem, the resolution of which required meticulous investigative work to understand the nature of the source itself. First, it transpired that the second part of the manuscript was constructed from a sequence of four-page folios (two sheets folded in half each). Secondly, the page numbers were no guide to the ordering of the music, since they ran sequentially through each movement even through those passages that were clearly out of place, suggesting that they were added only after the folios were assembled in that wrong order. Finally, even sections that seemed to represent continuous spans of music in the fourth movement were in fact disrupted by further mis-ordering of the folios.

With this information, it became possible to determine that the twelve pages of the fourth movement that appeared out of place between movements two and three in fact consisted of two disconnected fragments of the fourth movement, one spanning a single folio and the other spanning two. Similar dissections were necessary for the apparently continuous section of the fourth movement from page 43 to the end of the scan, which appears in fact to contain four disconnected fragments of music.

Having identified these six fragments across the whole file, one might wonder if it is possible to link them up in a "corrected" order. This requires drawing on aspects of phrasing, material, formal structure, and harmony to indicate which sections belonged together. Some sections matched perfectly in musical terms, clearly indicating that certain folios belonged together, thus making it possible to re-assemble longer spans of music this way. However, this work eventually led to the conclusion that there must be further material missing from the booklet, since it was not possible to link all six sections continuously. The last fragment in particular starts with a tied low E in the cello and contains an end repeat marking. None of the other fragments contain the corresponding features. Those missing pieces of information must lie in material that is not present here, and which may well be lost altogether.

For all the work involved in coming to this conclusion, the resulting transcription posed a challenge for how to integrate it into the corpus. We hope that performers will form a substantial user-base for these scores and parts, and as noted above, any obstacles they encounter would be detrimental to the reputation of already-marginalised composers. We ultimately decided to upload the second and third movements alone. Although this means the official set will remain incomplete, these are the only movements that are unambiguously performable as they stand. The complete transcription, including the incomplete movements, and individual fragments, has been uploaded to the GitHub mirror so that those who are interested can see what the fullest available version contains and can access it in a free and open-source environment. Links to the GitHub score, as well as to this report, will be provided in the score description on MuseScore.com so that those who come across the piece can follow up on our work and perhaps advance it, particularly if new material emerges.

This case of the Maier A major quartet crucially demonstrates the limits of our interventions. It would not be impossible to compose completing sections for the first movement and the finale, as indeed others have: there exists a 2020 recording of the piece (as completed by B. Tommy Anderson for dB Productions Sweden). However, such manifestly creative work is beyond our remit. We can address details, correct errors, and iron out inconsistencies, but we draw a line at the point where actual composition is required. Moreover, the completed version is clearly in-copyright and thus not available to this project (we instruct contributors not to consult that material).

5 CONCLUSION; OUTLOOK FOR RESEARCH

We consider this dataset to be an important resource for musicians, music-lovers, and musicologists alike. First and foremost, many of the quartets presented here have been un- or barely- available until this intervention. Professional and amateur quartets are already involved in performing these 'forgotten' works. For example, the modern premiere of the Mayer will take place in December 2023.[21]

Among the many research possibilities, we see particularly strong potential for studies of:

- **Musical parameters**, such as
  - *texture*, expanding on the basic script provided here and building on the work of Florence Léve and colleagues (on texture in general [5], as well as in the more specific cases of music for orchestra [12], and piano [4]).
  - *form*, addressing long works in a range of formats, including the typical sonata movement types, and for works of clearly manageable scale (i.e., four distinct instrumental parts rather than the 20+ of an orchestral work).
  - *harmony, counterpoint, and voicing*, benefiting from the clear separation of voices, for which see also [1]'s MCMA and the 'EFER' corpus (forthcoming).
- **automatic composition** and related tasks, partly through the domain-expertise studies described above.
- **OMR**, using this corpus' clear and explicitly identified IMSLP source editions to test, develop, and evaluate OMR. For instance, it would be interesting to convert each edition of the work in question, test the similarity of those results against our transcription, and see how often the OMR technology (and similarity metrics) are robust enough to correctly identify the source used.

We continue to develop this corpus and welcome contributions. Naturally, we also welcome suggestions and PR requests from the MIR community for improvements to the provision as it stands and are open to suggestions for collaborations in developing and/or using the corpus. We note that MuseScore 4.2 will include MEI import and export,[22] and could imagine this serving for greater coordination across the community. OpenScore X (string quartets and lieder) could provide a testing ground.

We also hope that the work outlined above will stand as an important reminder of the ongoing need for human decision-making in preparing musical scores. Many of the interventions outlined above relied on extensive lived experience of music and notation conventions, as well as publishing and printing practices that would have been impossible to replace with existing technology alone.

[21] https://benslowmusic.org/index.asp?PageID=3273

[22] See https://github.com/musescore/MuseScore/pull/18705 and subsequent PRs here: https://github.com/musescore/MuseScore/pulls?q=is%3Apr+mei

ACKNOWLEDGMENTS

We wish to thank all who have contributed to this collection. Thank you to all the transcribers (individually credited on the musescore.com webpage for each quartet) and the core team of reviewers (credited in addition to the transcribers on our coordination spreadsheet). While contributors have mostly given their time and efforts as volunteers, we also very gratefully acknowledge in-kind support from MuseScore.com (pro accounts membership to contributors) as well as funding from the University of Cambridge (to initiate the lieder corpus effort) and the Central New York Humanities Corridor via the “Music Theory Examples by Women” working group founded by author MG in 2019 in support of these goals.

REFERENCES

[1] Anna Aljanaki, Stefano Kalonaris, Gianluca Micchi, and Eric Nichols. 2021. MCMA: A Symbolic Multitrack Contrapuntal Music Archive. *Empirical Musicology Review* 16, 1 (2021), 99–105. Publisher: The Ohio State University Library.

[2] Clive Brown. 1993. Dots and Strokes in Late 18th- and 19th-Century Music. *Early Music* 21, 4 (1993), 593–610. http://www.jstor.org/stable/3128368

[3] Clive Brown. 1999. *Classical and Romantic Performing Practice 1750-1900.* Oxford University Press.

[4] Louis Couturier, Louis Bigo, and Florence Levé. 2023. Comparing Texture in Piano Scores. In *International Society for Music Information Retrieval (ISMIR 2023)*. Milan, Italy.

[5] Mathieu Giraud, Florence Levé, Florent Mercier, Marc Rigaudière, and Donatien Thorez. 2014. Towards Modeling Texture in Symbolic Data. In *International Society for Music Information Retrieval Conference.*

[6] Mark Gotham. 2021. Connecting the Dots: Recognizing and Implementing More Kinds of “Open Science” to Connect Musicians and Musicologists. *Empirical Musicology Review* 16 (2021). https://doi.org/10.18061/emr.v16i1.7644

[7] Mark Gotham, Johannes Hentschel, Louis Couturier, Nathan Dykeaylen, Martin Rohrmeier, and Mathieu Giraud. 2023. The ‘Measure Map’: an inter-operable standard for aligning symbolic music. In *International Conference on Digital Libraries for Musicology (DLfM 2023).* Milano, Italy.

[8] Mark Gotham and Peter Jonas. 2021. The OpenScore Lieder Corpus. In *Music Encoding Conference (MEC ’21).* MEC. event-place: Alicante (Virtual Conference).

[9] Mark Gotham, Peter Jonas, Bruno Bower, William Bosworth, Daniel Rootham, and Leigh VanHandel. 2018. Scores of Scores: An OpenScore Project to Encode and Share Sheet Music. In *Proceedings of the 5th International Conference on Digital Libraries for Musicology (DLfM ’18).* ACM, New York, NY, USA, 87–95. https://doi.org/10.1145/3273024.3273026 event-place: Paris, France.

[10] Mark R. H. Gotham, Rainer Kleinertz, Christof Weiss, Meinard Müller, and Stephanie Klauk. 2021. What if the ’When’ Implies the ’What’?: Human harmonic analysis datasets clarify the relative role of the separate steps in automatic tonal analysis. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference.* ISMIR, Online, 229–236. https://doi.org/10.5281/zenodo.5676067

[11] James Norman Grier. 1996. *The Critical Editing of Music: History, Method, and Practice.* Cambridge University Press.

[12] Dinh-Viet-Toan Le, Mathieu Giraud, Florence Levé, and Francesco Maccarini. 2022. A Corpus Describing Orchestral Texture in First Movements of Classical and Early-Romantic Symphonies. In *Proceedings of the 9th International Conference on Digital Libraries for Musicology (DLfM ’22).* Association for Computing Machinery, New York, NY, USA, 27–35. https://doi.org/10.1145/3543882.3543884

[13] Néstor Nápoles López. 2017. Joseph Haydn - String Quartets Op.20 - Harmonic Analysis Annotations Dataset. https://doi.org/10.5281/zenodo.1095630

[14] Alon Schab. 2022. *A Performer’s Guide to Transcribing, Editing, and Arranging Early Music.* Oxford, New York.

[15] Fudie Zhao. 2022. A systematic review of Wikidata in Digital Humanities projects. *Digital Scholarship in the Humanities* 38, 2 (12 2022), 852–874. https://doi.org/10.1093/llc/fqac083

A ONLINE RESOURCES

- Public-facing collection on MuseScore.com: https://musescore.com/openscore-string-quartets
- Dataset mirror on GitHub for use by the MIR community: https://GitHub.com/OpenScore/StringQuartets/
- Coordination spreadsheet: https://bit.ly/StringQuartetsSheet
- Editorial notes: https://bit.ly/StringQuartetsEdits