

# Connecting the Dots: Engaging Wider Forms of Openness for the Mutual Benefit of Musicians and Musicologists

MARK R. H. GOTHAM

*Technische Universität Dortmund, Germany*

**ABSTRACT:** While it is encouraging to see renewed attention to ‘openness’ in academia, that debate (and its interpretation of the F.A.I.R. principles) is often rather narrowly defined. This paper addresses openness in a broad sense, asking not so much whether a project *is open*, but *how open* and *to whom*. I illustrate these ideas through examples of my own ongoing projects which seek to make the most of a potential symbiosis between academic and wider musical communities. Specifically, I discuss how these communities can both benefit from – and even work together on building – highly accessible and interoperable corpora of scores and analyses when ambitious openness is factored into decision making from the outset.

Submitted 2020 April 30; accepted 2021 January 27.

Published 2021 December 10; <https://doi.org/10.18061/emr.v16i1.7644>

**KEYWORDS:** *harmony, corpus study, pedagogy, performance, community*

THE academic (and wider) world is rapidly becoming much more proactive about matters of diversity, accessibility, and openness. Within music theory and analysis, recent manifestos from the UK (Horton, 2017) and the USA (Campbell et al., 2014; Ewell, Rao, Straus & Hisama, 2019) recognize an urgent need to think again about what we really offer in universities and colleges, and for whom. Computational musicology can and should feature prominently in this debate. Let’s begin by introducing some hypothetical colleagues to motivate this particular study about openness in our field:

- Professor X is a music information retrieval scholar in a computer science department focusing on corpus study of symbolic data. She generally has to produce her own corpora of scores and/or analyses either individually or with a team of people paid by dedicated grants to undertake tasks that they would not otherwise do.
- Dr Y teaches an annual cycle of several courses and hundreds of students in the musicology department of a teaching-focused institution. He is constantly producing and marking human analyses, but not in a computer-readable format. He also occasionally type-sets a score for this purpose but does not share it publicly.
- Finally, Maestro Z is a conductor whose regular work includes university ensembles, but who has little other engagement with academia. Like Dr. Y, Maestro Z will occasionally type-set music for a particular purpose (to arrange a work for specific forces, or to provide a quick fix for an editorial problem) but would not think to share this work.

These characters have a great deal of musical training, understanding, and needs in common. Yet, at least when it comes to the creation and distribution of musical datasets, there is remarkably little connection between their worlds. This is partly a matter of the perennial difficulties with implementing meaningful connections between different institutions. For a start, it is relatively rare to see the same person both teach in a music department and conduct computational musicology research. On the fictional side, while Prof. X would have preferred not to choose between the labels ‘computational’ and ‘musicological’, institutional pressures soon saw her firmly into the former ‘camp.’ Turning to reality, witness the affiliations of those attending recent ISMIR conferences: very few have permanent positions in musicology departments.



This is simply an observation; there is no straightforward way to judge whether it is a ‘good’ or ‘bad’ situation overall. In broad strokes, we might say that it has the benefit of opening up more interdisciplinary connections and a wider range of viewpoints but also the detraction of potentially dissociating the field and burdening it with two very different disciplines to convince of its merits. The more useful and immediate question to ask is: how can we maximize the advantages and work on the challenges? This paper seeks to address those questions for the case of ‘openness’ in preparing and disseminating datasets. I will argue that Prof. X, Dr Y, and Maestro Z could work more closely together to the considerable benefit of all. I illustrate this through example projects which attempt to support and expand access to teaching, research, and music-making of many kinds.

I focus on my own projects to communicate some hard-won experience frankly, to illustrate that traction is possible, and to advance the conversation about best practice(s).[1] There are many other publicly available datasets of scores,[2] score representations in other formats,[3] and analyses.[4] I do not presume to judge the relative strengths and weaknesses of anyone else’s dataset here, but naturally we will discuss in general terms the relative ‘openness’ of the decisions that lie behind the creation of any such corpus.

The two main sections focus respectively on score and analysis corpora, specifically on the ‘OpenScore Lieder Corpus’ of songs,[5] and the ‘When in Rome’ corpus of Roman numeral analyses.[6] Beyond datasets, I also argue that the code we produce for relatively basic tasks within research projects can also readily serve wider communities. To that effect, this article also introduces new code and resources for: automatic retrieval from human analyses for creating music anthologies at scale,[7] and a free, public-facing app for ‘spell-checker’-style feedback on harmonic analyses.[8]

## SCORE ENCODING

The ‘F.A.I.R.’ principles (Wilkinson et al., 2016) argue that datasets should be ‘Findable, Accessible, Interoperable, and Reusable’ (hereafter ‘FAIR’). These are helpful terms, but they can be invoked more or less strictly. A minimum definition would require only that datasets be *potentially* FAIR. By contrast, a more genuinely ‘open’ reading would require that they be *easily* findable, and *readily* accessible, interoperable, and reusable for a *range* of purposes by *diverse* users.

As the potential methods for engaging with datasets are very broad, so too are the forms that this enhanced openness can take. Nevertheless, there are some core principles at work, including the relative openness to discovery directly by people, ‘manually’, on one hand, and to automatic searches on the other. Put even more directly, this generally speaks to a division between groups of people, between the more and less computationally minded. For instance, even where a dataset is easily accessible by anyone *in principle* (with a simple ‘download’ button as offered by a site like GitHub/Gitlab, for instance), that choice of hosting platform can still enforce the same division between those who would think to look there (developers in the case of GitHub/Gitlab) and those who would not (most amateur musicians).

Enhanced openness is desirable, but it is often both harder to realize and not mission-critical for the specific task at hand. Prof. X has a grant to support the preparation of a specific dataset but is only interested in pitch and does not bother encoding the rest reliably. She also has a preferred format that is convertible in principle, but that converter requires expert knowledge to operate and is not well maintained. Dr. Y and Maestro Z, in turn, produce encodings of excellent quality, but only do the sections they need, and although they use interoperable formats, they are not incentivized (or perhaps even aware of the option) to contribute that work to a communal effort. All parties might have benefitted from each other’s work but for the lack of communication and coordination. Improving this situation calls for a more collaborative approach, including sharing work of at least *decent quality* (a ‘minimal viable product’), in *interoperable formats*, and under a *clearly permissive license*. Of these three, licensing is the centerpiece; we can’t progress meaningfully without the ability to build on each other’s work. We will return to a more detailed look at what this entails at the conclusion of this section. Quality and interoperability require further discussion right away.

With ‘quality’ we need to distinguish (somewhat crudely) between what is sometimes called the ‘semantic’ side (centered on notes and rests) and other matters (like layout). It is not always possible to firmly separate these two, but the principle is relatively clear: the most useful resource to share is an encoded score which conveys the core content of an existing, named, and easily accessed source edition. This immediately supports most users: those who simply want *some* version of the work can get going on their tasks; others seeking to produce scores edited and laid out in a specific fashion can focus on those matters of fine-tuning to complete a task that (even for their purposes) is already largely complete.

‘Interoperable’ is a potentially slippery term, though it can be considered self-defining: the most interoperable formats can be parsed by the greatest number of other software and code libraries, with the least ‘friction’ or ‘slippage’ between those representations. Currently, musicXML is the richest format that can make this claim. It is mutually convertible with most notation software (including commercial entities like Sibelius and Finale) and code libraries (music21, humdrum, ...) with relatively little loss of information.[9] By contrast, although MIDI is parseable by more software, it is considerably less rich in detail (losing all notion of pitch spelling, for instance). Newer initiatives like MEI and MNX that seek to improve on the existing XML format tend to be richer, but the specific question of their interoperability will depend on the provision of effective converters.[10] Even if we decide as a community to create converters for open-source applications and code libraries (a time-consuming and ill-rewarded task), interoperability with commercial software remains out of our hands, and companies can be slow on the uptake. It may seem odd to consider the commercial sector in a story about openness, but again, we must return to the central question: what kinds of openness are we fostering, and for whom? If new default formats do not serve as broad a range as musicXML does currently, then that is a loss for interoperability.

Even with a principled position on these matters, the practicalities of coordinating an encoding drive are subject to considerable challenges. To consider this in less abstract terms, we turn now to discuss the specific case of the ‘OpenScore Lieder Corpus’ and questions of editorial policy, recruitment, and collaborative workflow with which all such projects need to reckon (both in general, and specifically in terms of ‘openness’).

### Platform(s)

The OpenScore Lieder Corpus has explicitly set out to meet the criteria outlined above, providing a resource that scholars, teachers, and musicians can use for any purpose. The corpus now offers over 1,200 high-quality encodings of songs that can be viewed, played, transposed, and more either directly online at <https://musescore.com/openscore-lieder-corpus/sets> or once downloaded in any of several standard formats.[11] Usage statistics show that users do indeed engage with these scores hundreds of times per day (via views, plays, downloads, and more).

We reached this level of engagement by factoring it in from the outset: this is not exclusively a research dataset, but a score collection for which musicological research is one of several use cases. That simple fact informs every decision, including the primary host platform (musescore.com), which is set up for engagement with a wide musical community. Indeed, it is our attention to FAIR principles for the research side that has been slower to develop: downloading individual scores from musescore.com is easy, but not currently possible at corpus scale. To that effect, we have now developed GitHub mirror(s) of the entire corpus,[12] which in turn form the basis of an analytical corpus and a set of teaching resources to which we will return below.

Ultimately, an ideal platform would support all human and computational searches to find the desired content easily, whether that’s a PDF, a symbolic score, a recording (audio or video), or metadata about pieces (including information about movements and sections) and would even support diverse search terms for slippery concepts like genre and analytical properties. Sites like Musicbrainz.org and Doremus.org offer promising approaches to these matters of connected ‘music ontology’, but no single platform currently offers the full package. For now, the best we can do is to host scores on *complementary* platforms (musescore.com and GitHub in our case)[13], with clear and consistently structured metadata (for the mostly unequivocal items like title, composer, lyricist, and catalogue number), and with external links to collections like wikidata.org. This serves both the findability of those scores now and also the prospective incorporation of the corpus into future, more richly annotated and interconnected databases.

### Editorial ‘Policy’ and Recruitment

Any corpus-building effort is subject to the limited resources available and to a tradeoff between quantity and quality. As such, the ‘best’ approach depends on the project’s priorities, making a commitment to FAIR and open principles relevant even here. For the OpenScore Lieder Corpus, we have elected to copy specific, identifiable, public domain editions and we provide a link to that source for the sake of transparency and easy comparison. We aimed for reliable transcription quality, correcting only flagrant errors, and stopping far short of a scholarly edition. If we were optimizing for academic reward metrics such as publication, investing so much time in the scores’ quality without also producing a product recognizable as a scholarly edition

would be among the worst ways to proceed. We saw the problem differently: in order to maximize the number of scores reliably reusable, and to make best use of our amateur volunteers, it was important *not* to take on anything truly editorial.

That said, one cannot simply decide to ‘not do’ editing as each source poses specific problems in terms of a balance between fidelity to the reference source and consistency in the new corpus. In principle we have sought to emulate the original as closely as possible, with exceptions for two types of notational devices that affect openness. Firstly, guaranteeing openness to modern musicians requires the ‘updating’ of elements that have fallen decisively out of common usage. Examples include now-rare clefs or the positioning of pedal indications anywhere other than below the bottom staff. Secondly, we replaced notations which create undue complication in the encoded version, such as the splitting of a single measure across a system break. A representation of this is possible, but only with hidden, non-standard elements (two measures, of which only one is counted) which make conversions, interoperability and other computational processes more difficult. This comes back to balancing the prospective use cases. The measure-split may be a more direct reflection of the source and perhaps better for use in training optical musical recognition software (hereafter OMR), but is much worse for most other computational purposes and is also increasingly unfamiliar to musicians today.

To make these kinds of judgement calls effectively, we need to understand different users’ needs. We can arrange consultations to this effect, but the most direct and meaningful engagement is to be had by involving a range of users in the building of the corpus itself, very much including amateurs alongside professional musicians and musicologists. Moreover, apart from bringing complementary priorities and insights, we have found that some of the most committed and successful contributors have begun as musical novices, perhaps because the tasks are proportionately more interesting and valuable for personal, musical development at that stage.[14] With a little encouragement from more experienced members of the team, we have helped these contributors to develop their own skills (a worthwhile service in itself) and benefitted from the transcriptions they subsequently produce.

Platforms with an ‘open to all’ policy and social-media-esque stylings can help to build this kind of collective transcription community by gathering like-minded people together and promoting all-important personal connections. This has been central to the Lieder Corpus, both for inviting contributors to work on new scores, and (in some cases) for adopting existing ones by reaching out to the transcriber via the site. Where those existing scores are already released under a suitable license, we can thank the transcriber for their work and get to work on any necessary improvements immediately (either separately, or in collaboration with the original transcriber as they prefer). Where the license is not suitable, often the transcriber has simply not considered this aspect: on several occasions, a short exchange has confirmed their preparedness to contribute their work. In some cases, this has been the initial moment of recruitment for highly productive and valuable contributors. This speaks to the benefits of personal contact and open channels of communication with (potentially) interested parties. It also affects the best-practice in workflow.

## Workflow

Coordinating multi-author transcription projects requires careful coordination and distribution of tasks. In the Lieder Corpus we have been able to keep this reasonably simple: we take every song as an atomic unit, allocating exactly one transcriber and one (different) reviewer to each,[15] and keep track of progress on a spreadsheet viewable by all but editable only by the review team. More complex projects than ours may require a more sophisticated version of this process, for instance, distributing the task of transcribing among different parties according to their experience and time availability, ranging from simple, one-click reCAPTCHA-style tasks (‘do the pitches in these two score excerpts match?’) to expert editorial judgement calls. There are promising ideas for implementing this,[16] which if combined with McGill’s forthcoming mass OMR conversion of the entire IMSLP catalogue, could be a strong basis for coordination and potentially a watershed moment in the score encoding effort.

That said, we would do well to synthesize project management efficiency with what contributors will enjoy doing. While there is a great enthusiasm for encoding, there seems to be less of an appetite for correcting existing encodings. This may be partly for want of a coordinated system and lack of clarity over licenses, but it would seem to run deeper. For instance, we should not expect wholesale enthusiasm for a set-up in which contributors are expected to fix OMR.[17] OMR has advanced to a point that it is no longer so easy to dismiss revising it as being more time-consuming than transcription from scratch, and yet that argument persists. To engage with that viewpoint, we need to establish whether it has simply become a

convenient distraction from a simple preference for encoding from scratch, and (if so), what really motivates that preference. There may be merit to this preference – for instance, correcting OMR may be a ‘less musical’ task that does not offer the same (or equivalent) pedagogical benefits as transcription-from-scratch. In short, even a sophisticated infrastructure for coordination should perhaps accommodate the option of doing this task *ab initio*, ‘the hard way.’

Correcting OMR is one matter; reviewing each other’s work is another. In academic corpus creation this too is mostly conducted behind closed doors, with no release until the final product is ready. This is an understandable position for contexts unaccustomed to sharing work in progress, though error-checking is one task for which the ‘many heads are better than one’ approach of crowdsourcing is ideally suited. A change of focus from discrete, disconnected products to many ‘micro-contributions’ model may help here too by diminishing the expectation of perfection prior to release and thus encouraging all-comers to contribute what they have, proof-read and polished or otherwise. Quite apart from the extensive and variable quality of what is hosted online, who can say what treasure troves may be hiding on hard drives around the world, ready to swell the ranks of the world’s collection of FAIR and open encoded scores?

### **Moving Forward Together**

While there remain many open questions over the best technical and inter-personal methods for corpus building, some considerations are abundantly clear. First, we are strongly incentivized to work together, so that we can build meaningfully on each other’s work and produce large, open corpora for all. Second, we can and should find ways to involve members of as broad a range of musical communities as possible not only in the *use* but also the *creation* of corpora. To that effect, organizers should be sensitive to the kinds of tasks that people actually want to do, even if this does not appear to optimize efficiency. Many people around the world happily transcribing from scratch is better than a few people miserably correcting OMR. Third, and most important, we need widespread support for the philosophy of open science extolled in this special edition of EMR: in short, we should release score encodings under the maximally permissive licenses (CC0, PD, or equivalent). Any formal restrictions are a serious barrier to progress.

This is no small matter. Open licenses involve a real commitment to corpus-building as a collective rather than individual effort. To take the thorniest issues head on, consider the ‘non-commercial’ and ‘attribution’ options (to use Creative Commons’ language). ‘Non-commercial’ may sound virtuous, but while score encoders opting for this license may imagine themselves fending off a monolithic, exploitative commercial entity, it is much more likely to cast doubt on well-meaning use cases like performance of the score at a charity fund-raiser concert. Even if that specific use is (or would have been) permitted, the lack of clarity means would-be users waste time and energy wondering and will quite likely decide not to risk it.

‘Attribution’, in turn, is something we should all strive for: we certainly should acknowledge people who put time and energy into projects for the common good where that is practical and important, but here too the strict formality is contrary to broad openness. Wouldn’t we do better to commit informally to acknowledging one another where that is important and practical (such as the full disclosure of datasets used in a scholarly study) but not require it formally in all cases (like that amateur concert)? The answer depends in part on our attitudes to engagement with wider musical communities, our openness to all.

As a step towards a more meaningful and productively coordinated effort, it would be helpful to agree on a minimal ‘manifesto’ or value statement in which we:

- recognize the value and importance of digital scores for musicians, teachers, and scholars to adapt for various purposes,
- likewise, recognize the considerable effort it can take to produce these scores,
- informally undertake to attribute that work where it is practical and meaningful to do so,
- nevertheless consider intellectual property claims inappropriate for the direct transcription (without significant, new editorial intervention) of works in the public domain,
- accordingly believe that the formal license for such transcriptions should be maximally permissive (CC0, PD, or equivalent) and commit to this in our own work.

## ANALYSIS ENCODING

Encoding the world’s collection of public domain scores is a very large task, but a finite one, and with technological improvements to speed us up, it may be a relatively short-term issue. The broader question of how we can engage with wider musical communities both as creators and users of open resources has much wider ramifications. For example, there are comparable prospective benefits of also encoding and sharing *analyses* in FAIR and open ways. Here too, computational methods afford us opportunities to share our work in a way that enhances the public provision, connects musical and scholarly communities, and achieves a more genuine kind of ‘openness’.

‘When in Rome’ is the title of a centralized repository in which I have brought together every human harmonic analysis in an encoded format that both humans and computers can read and work with. The repository now offers:

- Over 400 analyses and 85,000 Roman numerals.
- Those analyses in multiple formats that are easy to read and edit without computational experience, including both a lightweight text format (the ‘RomanText’ of Tymoczko, Gotham, Cuthbert & Ariza, 2019) and musical notation (aligned with the score where available – see Figure 1 for an example).[18]
- Code for pedagogical resources that build on this corpus to serve wider musical uses and communities and which also stand to harness the power of crowdsourcing to build and expand the datasets further. That is the subject of this section.

### Open Music Theory Textbook and Anthology

While many textbooks now include a parallel online provision, they generally remain limited in several important ways. First, the number of examples still tends to be rather meagre, despite the fact that page constraints of the printed medium no longer apply. Second, those examples are still very often static images that offer little versatility or functionality beyond their print equivalents. Third is persistence of prohibitive cost. There is a great public appetite for music theory (as attested by vigorous discussion on social media sites and YouTube channels, for instance), but only a tiny percentage of those interested parties would be prepared to pay \$100 or more for a textbook where some version of the same content can be accessed for free online. These considerations are part of the motivation for *Open Music Theory* (Gotham et al., 2021): a textbook, anthology and workbook project in which we have sought to provide free, high-quality material that is both comparable to the traditional print textbook but also benefits from the additional reach and functionality of the online medium.

5 Lieder, Op.10  
1. Nach Süden

Fanny (Mendelssohn) Hensel

The figure shows a musical score for 'Nach Süden' by Fanny Hensel. It consists of three staves: Voice, Piano, and Analysis. The Voice staff has a treble clef and a key signature of three sharps (F#, C#, G#). The lyrics are 'Von al - len Zwei - gen schwin - gen sich'. The Piano staff has a grand staff (treble and bass clefs) and a key signature of three sharps. The Analysis staff has a treble clef and a key signature of three sharps, showing Roman numerals: E: I, vi64, ii, viio6/ii.

**Fig. 1.** Fanny Hensel (née Mendelssohn)’s 5 Lieder, Op.10 No.1: ‘Nach Süden’ as an example of an OpenScore Lieder Corpus score with aligned harmonic analysis part added.

Among our hypothetical colleagues, Prof. X would be the one most likely to produce or use the ‘When in Rome’ corpus, and Dr Y would be more oriented towards the textbook, but here too we stand to benefit from thinking across those boundaries. For instance, one of the most fundamental tasks of both a textbook-anthology and a corpus study of analyses is to explore specific instances of a certain chord or progression. This may be in the service of other matters, but even the simple retrieval of examples is of value in itself. Harmonic analysis datasets like ‘When in Rome’ enable a relatively easy retrieval of this kind of information and can be used to produce musical anthologies of examples at scale.[19]

I have included new code in the ‘When in Rome’ repository for building such an anthology,[20] and implemented one version of this as the ‘Harmony Anthology’ in *Open Music Theory* (<https://viva.pressbooks.pub/openmusictheory/chapter/anthology-harmony/>). This anthology provides lists of examples by topic (standard anthology fare such as augmented sixth chords), *automatically* extracted from *human* analyses of CC0 scores in the Lieder Corpus. It bears repeating that while the processing is automatic, the analyses themselves are by people. Among other benefits, the scale of the provision stands to help students contextualize assertions made about so-called ‘normal’ resolutions. While we may still opt to highlight a few examples in the first instance (as the textbook does), partnering them with links to hundreds of examples – very much including borderline cases – provides a much stronger sense of how exemplary those cases really are of wider practice. It will cost nothing to scale this anthology up as score and analysis corpora grow in the future.

Further, when the dataset of harmonic analyses eventually includes *multiple analytical readings* of the same works for a larger number of pieces,[21] we will be able to highlight the areas where analysts disagree. These divergences provide important insights both in relation to the particular work at hand, and of the underlying theoretical foundation of the analyst’s approach in a more abstract sense. In the simplest case, finding these divergences is readily automatable as the difference between two files.[22] More sophisticated versions can point to the musical nature of the difference between two sources, a subject to which we now turn.

### **Openness to all: ‘The Roman Umpire’ code and ‘Working in Harmony’ app**

Scores and analyses hosted publicly under open licenses and amenable formats are valuable, but if we wish to help people develop their skills in tasks like Roman numeral analysis, then we need to offer more. Here, too, routine computational operations developed in the service of research projects can be readily repurposed for these broader uses and may serve, in turn, to grow the corpora.

‘The Roman Umpire’ is a new code module providing basic feedback on harmonic analyses. For instance, one feedback option involves comparing the pitches implied by each harmony with those in the score at the corresponding moment and alerting the user to moments where the pitch content and/or bass notes diverge. The choice remains with the analyst, the code simply points out moments they may wish to review. I initially developed this code for proof-reading during corpus building; as it proved so useful for experts checking their work, it stands to be doubly useful in pedagogical contexts with a greater need for such guidance. Furthermore, as the process of generating feedback is completely automatic, we can readily offer this as a service that anyone can use as much or little as they chose, whenever they like, via a simple web app.

We have created an initial, alpha version of such an app called ‘Working in Harmony’ at <https://fourscoreandmore.org/working-in-harmony/>,[23] initially offering feedback on a user’s analysis of any score from the Lieder Corpus. This will prospectively serve to help students enrolled in classes proofread their work more effectively, freeing up class time and instructor feedback for more interesting, musical matters. At least as importantly, it will offer some provision, however rudimentary, to those who wish to learn Roman numeral analysis but have no such class enrollment.

This too can contribute to the creation of FAIR and open datasets and the specific question of involving diverse contributors. We offer this service for free to anyone and invite users to return value as citizen scientists by submitting their analyses (anonymously, unless they choose otherwise) into a corpus which may prove useful to computational musicologists, not least in learning more systematically about the kinds of errors and improvements that students of Roman numeral analysis make in the course of their work. This provides a potential for symbiosis, in which simple code developed in the service of research projects can serve the wider musical public, who in turn help to build up corpora for a range of research projects including those with a focus on developing public-facing pedagogy.[24]

## Openness to adjustment

Simple though these computational operations may be in principle, there are always procedural questions and challenges to address in the design of such a system which have a direct bearing of the prospective adoption by musical communities, and thus to whom they are truly open. The final parts of this paper consider how computational formats and systems can both promote – but also constrain – openness.

Just as encoded scores (like musicXML) are preferable over ‘fixed’ formats (PDF) on the basis of their FAIR attributes, so too is that true of analyses: among other tasks, we need to be able to access the constituent information of an analysis in a structured way and to adjust that work in place. For simpler musical passages, analysts often agree about their preferred reading of a passage. That being the case, one of the main benefits of sharing analyses in easily edited formats is the opportunity for other analysts to add a variant reading either within the same file or by making a copy and adjusting only the passages (often few and far between) that they see differently. Editable files speed up the process of producing those analyses by avoiding duplicated work.

In cases of more sustained disagreement, editing an existing analysis in place may not be helpful, or at least not without a more sophisticated user interface. This is not only a question of the number, but also the types of changes. For instance, editing an existing analysis arguably presupposes a method by which the analyst works through from start to finish in a linear fashion, but many analysts do not work this way, preferring instead to start at a kind of middle-ground level, seeking only the main moments indicative of the three harmonic functions (tonic, pre-dominant, dominant) in preparation for a cadence.[25]

Engaging with this use case would probably require some kind of multi-layered analysis for harmonic motion at different ‘levels’, perhaps vertically stacked (in a quasi-Schenkerian style) and aligned with the score, thus enabling users to weave freely between levels. The system could perhaps offer an option to auto-complete certain parts, for instance by extracting the ‘big picture’ readings from detailed harmonic analyses using a system such as that proposed by Rohrmeier (2011, 2020). This could potentially provide a valuable framework for building students’ metacognitive understanding of harmonic analysis. That said, the potential for confusion is high: for instance, when you adjust a detail in the foreground analysis, should the other levels auto-update? User-interface questions of this kind are among the most important for determining how FAIR and open a resource is, and to whom.

## On the relative openness of syntax

Openness to adjustment is a high priority for both analysts and for the user-interface of any public-facing app. At the same time, any computational system requires at least some strict syntactical constraints. Specifically, in this case of Roman numeral analysis, we need symbols to have a single, unambiguous meaning. Implementing this requires some careful consideration and warrants discussion.

First, there is some variation in how Roman numerals are encoded, for instance, in the encoding of inversions by Arabic numerals (V42, I6, ...) or Latin-alphabet letters (so V42 becomes Vd and I6 is Ib). While the former is probably more common in musicology today, the latter is still in active use.[26] Nevertheless, there is no internal ambiguity: it is clear within either system which pitches are invoked.

Similarly, even within the figured Roman system, there are different conventions, for instance in handling the sixth and seventh degrees in minor. Should ‘vii<sup>o</sup>’ refer to a chord on the (raised) leading note, or should it require a leading sharp (‘#vii<sup>o</sup>’)? There are different approaches here, but still no internal ambiguity. Music21’s handling of all common variants has improved recently in tandem with our presentation of the ‘Roman Text’ standard (Tymoczko et al., 2019)[27] and has developed further since, notably through the integration of Nápoles López and Fuginaga (2020)’s unifying regular expression.

Nevertheless, while most of the headaches of apparent ‘ambiguity’ in Roman numeral analysis are actually attributable to variant, interoperable styles, genuine ambiguity does still creep in at the edges. Abbreviations clarify where this line falls. Perhaps the most common abbreviations in use in Roman numeral analysis handle the cases of:

1. Augmented 6<sup>th</sup> chords: for instance, ‘Ger65’ stands in directly for #ivo65[b3], compressing a complex symbol into a simple shorthand and identifying a canonical chromatic chord. (This class of abbreviation also helps to side-step some commonly confusing matters such as the ‘real root’ of the so-called ‘French’ sixth chord).



2. Neapolitan 6ths: ‘N6’ replaces ‘bII6’. The emphasis here is less on compression and more on highlighting a special status for this particular chordal configuration.
3. The Cadential 6/4: an abbreviation like ‘Cad64’ likewise serves to side-step ongoing arguments, this time over whether the Roman numeral for this chord ‘should be’ I64 or V64. This abbreviation can also contribute to a more contextual reading by distinguishing *cadential* from *passing* usages of the same chord.
4. Common-tone diminished seventh chords: sometimes afforded the label ‘Cto7’ explicitly to avoid ascribing tonal function (implied by Roman numerals) to a non-functional diminished seventh that shares at least one pitch with at least one of the preceding and following chords.

The first three cases here involve simple, direct, one-to-one mapping between the chord and the abbreviation, and thus are readily included. The ‘Cto7’ label, by contrast, does not unambiguously commit to specific pitches and pitch spellings. Often these pitches and spellings can be deduced from context, but that is not universally reliable and so the label is somewhat problematic from a computational perspective.

That being the case, rather than entering ‘Cto7’ and later attempting to deduce the pitches from there, we may do better to work the other way around, using unambiguous chordal designations at the point of creating analyses and then deducing the presence of a ‘Cto7’. For instance, perhaps the archetypical form of the Cto7 sees it prolong the tonic chord (I-Cto7-I), with the root of that prolonged tonic as the common tone, and the Cto7’s pitches equivalent to #iio7 (or another spelling with enharmonically equivalent pitches). Some analysts may balk at using any Roman numeral here due to the apparent harmonic function implied, but at least the chord is clear (with or without context) and it is still easy to extract all-and-only these moments for a study of common-tone progressions.[28] Moreover, if this is at the limits of what functional harmony and Roman numerals can express, then it is perhaps a natural limit for what a computational implementation should seek to support.

Some readers will find this a trivial aside, but analysts committed to using ‘Cto7’ may not be prepared to do without it and may thus reject any resource that does not include provision for it. In short, while we clearly want to support different users and their preferences, we may have to make concessions of this kind in some cases to support a common framework and/or make difficult procedural decisions in the design phase. That said, nothing prevents flexible users from reintegrating these notations into their non-communal uses of the corpora, especially when the computational means to do so are provided (as in the case of editable formats discussed here). We should aspire to create inclusive, open, FAIR datasets, but as discussed above in relation to scores, we should focus on sharing the most mutually useful resources from which more specialised individual projects can grow. Not everything has to be shared.

## SUMMARY AND PARTING WORDS

This paper argues that a greater integration of computational musicology with wider musical communities would be to the benefit of all in terms of growing datasets, learning from different perspectives, and fostering meaningful outreach. In the course of this, I have reported on my own projects which seek to contribute to that effort: the ‘OpenScore Lieder Corpus’, the ‘When in Rome’ analysis corpus, the ‘Open Music Theory’ textbook-anthology and the ‘Roman Umpire’ module for the ‘Working in Harmony’ app. I offer some suggestions for how we might go about collaborating on producing these kinds of corpora and resources, but I do not purport to have all the answers – indeed, I argue that we will need to collaborate on even planning the collaboration. I hope at least to have raised the profile of these questions and to have helped move the conversation along.

I have also acknowledged that this kind of work can be demanding and is not always well supported by academic reward structures focussed on particular kinds of research outputs. That said, the improved coordination needed here is desirable partly because it avoids the duplication of effort we currently see and that more streamlined approach can serve to accelerate the pace of research activities. Finally, I have argued for the widest possible participation in the creation and use of these resources, though I should close by acknowledging that it takes a very particular skill set to lead on the planning and implementation of such systems – a skill set which readers of this journal are uniquely placed to offer, whether they identify primarily with Prof. X, Dr Y, or Maestro Z.

## ACKNOWLEDGMENTS

Thank you, as always, to all who have contributed to the corpus-building projects discussed in this paper as well as to those friends and colleagues who read and provided insightful comments on earlier versions of this paper. This article was copyedited by Gabriele Cecchetti and layout edited by Diana Kayser.

## NOTES

[1] See McKay, Cumming, and Fujinaga (2020) for another current and commendably frank account of a similar kind (though one strictly limited to the research community).

[2] Among the examples from citable scholarship, see, for instance, CCARH (2001), Cuthbert and Ariza (2010), Rodin and Sapp (2010), Tilley (2017); more often these initiatives come from outside the academy – a central theme of this article.

[3] See, for instance, White and Quinn’s YCAC dataset of score ‘slices’ (2016) and Gotham’s code (2019) for producing more data of this kind.

[4] For the case of Roman numeral analyses discussed later in this paper, see for instance Devaney, Arthur, Condit-Schultz, and Nisula (2015) (hereafter ‘TAVERN’), Neuwirth, Harasim, Moss, and Rohrmeier (2018), and Tymoczko, Gotham, Cuthbert, and Ariza (2019).

[5] The public-facing website for the corpus is <http://musescore.com/openscore-lieder-corpus/> (Gotham et al. 2018), and we have just released a GitHub mirror of the corpus as a dataset for MIR research at <https://github.com/OpenScore/Lieder> (Gotham & Jonas, 2021).

[6] Code and corpus at <https://github.com/MarkGotham/When-in-Rome>.

[7] The code is provided as part of the ‘When in Rome’ repository as ‘anthology.py’ and the public facing implementation forms part of the ‘Open Music Theory’ textbook, (Gotham et al., 2021).

[8] Again, the backend code (inevitably called ‘The Roman Umpire’) is provided as part of the ‘When in Rome’ repository. The public-facing implementation is a free app accessible at <http://fourscoreandmore.org/working-in-harmony/>.

[9] Moreover, those ‘friction losses’ in conversion center on matters of layout, reinforcing the prioritization of other matters when encoding a corpus with a view to its interoperability.

[10] This may help to explain why the long-anticipated replacement of musicXML with MNX is yet to materialise as planned, with musicXML now headed for a version 4.0 in parallel to the development of MNX.

[11] The central format which the transcribers work in is MuseScore, not least because MuseScore is free (and thus available to all), open source, and includes relatively robust internal (batch) converters for export to musicXML, MIDI, PDF and MP3.

[12] The official mirror is at <https://github.com/OpenScore/Lieder>. There has also been a *de facto* mirror as part of the ‘When in Rome’ repository for some time. We are exploring how best to maintain version control across and auto-updating between sites. Currently the musescore.com copy is the ‘version of record’ and GitHub is ‘downstream’ from there, but clearly this fails to make use of git’s power in keeping track of version control. In this connection, see Foscarin, Jacquemard, and Fournier-S’niehotta (2019) for a promising diff method that (as the authors note) stands to serve ‘collaborative score editing, version control and evaluation of optical music recognition (OMR) or music transcription.’

[13] One might even imagine users of these scores exploring the origins of the encoding project and discovering the field of MIR research, which even many musically active people simply do not know exists. Closer integration between all the things a score can be through digital edition or the like will help foster this kind of mutually beneficial cross-pollination.

[14] Samiotis et al.’s (2020) study of a comparable crowd-sourcing task found that ‘workers with lower [musical experience] still achieved good performance, with a small additional cost in terms of cognitive load’.

[15] We have a smaller number of approved reviewers. These reviewers are also active as transcribers, but we do not allow reviewers to review their own transcriptions.

[16] See Shanahan and Albrecht (2013), Weigl et al. (2019), and especially Samiotis et al. (2020) for the most detailed pilot projects to date. Additionally, see Foscarin et al.'s diff method (2019) as discussed in end note 12 above.

[17] For a recent survey of multi-stage correction of OMR, see deGroot-Maggetti et al. (2020).

[18] This is currently available through the 'When in Rome' repository on GitHub, but again (as with the Lieder Corpus scores) we intend to duplicate the provision through a more public-oriented site like musescore.com or fourscoreandmore.org soon. There are technical hurdles we will need to overcome before this is possible, for instance, musescore.com does not currently enable bulk upload.

[19] For an exploration of alternative methods for digital-age anthologies, see Gotham (2019).

[20] When-in-Rome/Code/anthology.py. Please note that much of the musical logic is getting centralized within the main music21 repository to enable wider use cases (see for instance the roman.isMixture() method).

[21] Currently, only analyses from the TAVERN dataset include multiple readings. The initial priority is on increasing coverage of different scores, different analyses of the same scores will follow in due course.

[22] TAVERN's variant analyses exhibit mostly small differences, easily identified with the diff utility (Hunt & McIlroy, 1976). This is perhaps to be expected of the repertoire in question.

[23] The same backend functionality has also been incorporated into the <http://roman.algomus.fr/> app for automatic harmonic analysis based on our recent TISMIR paper (Micchi et al., 2020).

[24] See chordify.net for a somewhat similar provision, though in the context of a commercial product and lead-sheets.

[25] As taught in Laitz (2016), to take just one example.

[26] Notably, in datasets like TAVERN using the `**harm` extension of the `**kern` language. See the `harm2kern` command in the 'Humdrum Extras' package for converting `**harm` into `**kern` data and Nápoles López and Fuginaga (2020) for wider conversion as discussed in the main text.

[27] See the `roman.Minor67Default()` class.

[28] Again, the 'When in Rome' code base's 'anthology' script provides the functionality for this.

## REFERENCES

Brown, A. (2018). Algorithms and computation in music education. In Dean & McLean (Eds.). *The Oxford Handbook of Algorithmic Music*. New York, NY: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190226992.013.17>

Campbell, P. S., Myers, D., Sarath, E., Chattah, J., Higgins, L., Levine, V. L., Rudge, D., & Rice, T. (2014). *Transforming Music Study from Its Foundations: A Manifesto for Progressive Change in the Undergraduate Preparation of Music Majors. Report of the Task Force on the Undergraduate Music Major*. The College Music Society. Retrieved from <http://www.mtosmt.org/issues/mto.16.22.1/manifesto.pdf>

CCARH (2001). *The MuseData Database* (<http://www.musedata.org>). Stanford, CA: Center for Computer Assisted Research in the Humanities.

Clendinning, J. P., & Marvin, E. W. (2016). *The Musician's Guide to Theory and Analysis Anthology* (3rd ed.). New York, NY: W. W. Norton & Company.

Cuthbert, M. S., & Ariza, C. (2010). Music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the 11th International Society for Music Information Retrieval Conference* (pp. 637–642), Utrecht, Netherlands: ISMIR. <https://doi.org/10.5281/zenodo.1416113>

- Devaney, J., Arthur, C., Condit-Schultz, N., & Nisula, K. (2015). Theme and Variation Encodings with Roman Numerals (TAVERN): A new data set for symbolic music analysis. In Müller, M., & Wiering, F. (Eds.). *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 728–734). Malaga, Spain: ISMIR. <https://doi.org/10.5281/zenodo.1417496>
- Duker, P., Gawboy, A., Hughes, B., & Shaffer, K. (2015). Hacking the music theory classroom: Standards-based grading, just-in-time teaching, and the inverted class. *Music Theory Online* 21(1). <https://doi.org/10.30535/mt.21.1.2>
- Ewell, P., Rao, N., Straus, J. N., & E. Hisama, E. (2019). Reframing music theory. Plenary session presented at the 42nd Annual Meeting of the Society for Music Theory. Columbus, OH.
- Foscarin, F., Jacquemard, F., & Fournier-S'niehotta, R. (2019). A diff procedure for music score files. In *Digital Libraries for Musicology (DLfM '19)* (pp. 58–64). New York, NY: Association for Computing Machinery. <https://doi.org/10.1145/3358664.3358671>
- Gotham, M., Jonas, P., Bower, B., Bosworth, W., Rootham, D., & VanHandel, L. (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* (New York, NY, USA, 2018), DLfM '18, ACM, (pp. 87–95). <https://doi.org/10.1145/3273024.3273026>
- Gotham, M. (2019). Moments Musicaux: Towards comprehensive catalogues of real repertoire examples for teaching and research. In *6th International Conference on Digital Libraries for Musicology (DLfM '19)* (pp. 70–78). New York, NY: Association for Computing Machinery. <https://doi.org/10.1145/3358664.3358676>
- Gotham, M. & Ireland, M. (2019). Taking form: A representation standard, conversion code, and example corpus for recording, visualizing, and studying analyses of musical form. In A. Flexer, G. Peeters, J. Urbano, & A. Volk (Eds.). *Proceedings of the 20th International Society for Music Information Retrieval Conference* (pp. 693–699). Delft, Netherlands: ISMIR. <https://doi.org/10.5281/zenodo.3527903>
- Gotham, M., Gullings, K., Hamm, C., Hughes, B., Jarvis, B., Lavengood, M. & Peterson, J. (2021). *Open Music Theory*. Version 2. VIVA Pressbooks. <https://viva.pressbooks.pub/openmusictheory>
- Gotham, M., & Jonas, P. (2021, forthcoming). The OpenScore Lieder Corpus. In *Music Encoding Conference 2021*.
- deGroot-Maggetti, J., de Reuse, T. R., Feisthauer, L., Howes, S., Ju, Y., Kokubu, S., Margot, S., Nápoles López, N., & Upham, F. (2020). Data quality matters: Iterative corrections on a corpus of Mendelssohn string quartets and implications for MIR Analysis. In *Proceedings of the 21st International Society for Music Information Retrieval Conference* (pp. 432–438). Montréal, Canada: ISMIR. <https://doi.org/10.5281/zenodo.4245459>
- Giraud, M., Groult, R., & Leguy, E. (2018). Dezzann, a Web Framework to share music analysis. In *International Conference on Technologies for Music Notation and Representation (TENOR 2018)* (pp. 104–110).
- Horton, J. (2017). April 2017 Newsletter: 'The SMA's Strategy Document'. *Society for Music Analysis Newsletter*. Retrieved from [http://www.sma.ac.uk/wp-content/uploads/2017/04/SMA\\_newsletter\\_April-2017.pdf](http://www.sma.ac.uk/wp-content/uploads/2017/04/SMA_newsletter_April-2017.pdf)
- Huron, D. (1999). *Music Research Using Humdrum: A User's Guide*. Stanford, CA: Center for Computer Assisted Research in the Humanities.
- Hunt, J. W., & McIlroy, M. D. (1976). *An Algorithm for Differential File Comparison*, Bell Telephone Laboratories CSTR #41. <https://www.cs.dartmouth.edu/~doug/diff.pdf>
- Laitz, S. G. (2016). *The Complete Musician: An Integrated Approach to Theory, Analysis, and Listening* (4th ed.). New York, NY: Oxford University Press.

- McKay, C., Cumming, J., & Fujinaga, I. (2020). Lessons learned in a large-scale project to digitize and computationally analyze musical scores. *Digital Scholarship in the Humanities* (2020). <https://doi.org/10.1093/lc/fqaa058>
- Micchi, G., Gotham, M. & Giraud, M. (2020). Not all roads lead to Rome: Pitch representation and model architecture for automatic harmonic analysis. *Transactions of the International Society for Music Information Retrieval*, 3(1), 42–54. <https://doi.org/10.5334/tismir.45>
- Murdock, M., & Parsell, B. (2018). Music Theory Examples by Women. Retrieved from <https://musictheoryexamplesbywomen.com/>
- Nápoles López, N., & Fujinaga, I. (2020). Harmalysis: A language for the annotation of Roman numerals in symbolic music representations. In De Luca, E. & Flanders, J. (Eds.). *Music Encoding Conference Proceedings 2020* (pp. 83–85). <https://doi.org/10.17613/380x-dd98>
- Neuwirth, M., Harasim, D., Moss, F. C., & Rohrmeier, M. (2018). The Annotated Beethoven Corpus (ABC): A dataset of harmonic analyses of all Beethoven string quartets. *Frontiers in Digital Humanities*, 5. <https://doi.org/10.3389/fdigh.2018.00016>
- Rodin, J., & Sapp, C. (2010). *The Josquin Research Project*. <http://josquin.stanford.edu/>
- Rohrmeier, M. (2011). Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music*, 5(1), 35–53. <https://doi.org/10.1080/17459737.2011.573676>
- Rohrmeier, M. (2020). The syntax of Jazz harmony: Diatonic tonality, phrase structure, and form. *Music Theory and Analysis*, 7(1), 1–63. <https://doi.org/10.11116/MTA.7.1.1>
- Roland, P. (2002). The Music Encoding Initiative (MEI). In *Proceedings of the First International Conference on Musical Applications Using XML* (pp. 55–59).
- Samiotis, I. P., Qiu, S., Mauri, A., Liem, C., Lofi, C., & Bozzon, A. (2020). Microtask Crowdsourcing for Music Score Transcriptions: An Experiment with Error Detection. In *Proceedings of the 21st International Society for Music Information Retrieval Conference* (pp. 901–907). Montréal, Canada: ISMIR. <https://doi.org/10.5281/zenodo.4245579>
- Shanahan, D., & Albrecht, J. (2013). The Acquisition and Validation of Large Web-Based Corpora. Presented at the *Conference for the Society for Music Perception and Cognition*, Toronto, Canada.
- Tilley, J. (Ed.) (2017). *Web Library of Seventeenth-Century Music*. Society for Seventeenth-Century Music. <http://www.sscm-wlscm.org>
- Tymoczko, D., Gotham, M., Cuthbert, M. S., & Ariza, C. (2019). The Romantext Format: a flexible and standard method for representing Roman Numeral analyses. In A. Flexer, G. Peeters, J. Urbano, & A. Volk (Eds.). *Proceedings of the 20th International Society for Music Information Retrieval Conference* (pp. 123–129). Delft, Netherlands. <https://doi.org/10.5281/zenodo.3527755>
- Weber, G. (1832). *Versuch einer geordneten Theorie der Tonsetzkunst*. Mainz: Schott.
- Weigl, D.M., Goebel, W., Crawford, T., Gkiokas, A., Gutierrez, N. F., Porter, A., Santos, P., Karreman, C., Vroomen, I., Liem, C.C.S., Sarasúa, Á., & Van Tilburg, M. (2019). Interweaving and Enriching Digital Music Collections for Scholarship, Performance, and Enjoyment. In *6th International Conference on Digital Libraries for Musicology (DLfM '19)* (pp. 84–88). New York, NY: Association for Computing Machinery. <https://doi.org/10.1145/3358664.3358666>
- White, C., & I. Quinn (2016). The Yale-Classical Archives Corpus. *Empirical Musicology Review*, 11(1). <https://doi.org/10.18061/emr.v11i1.4958>
- Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., . . . , Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3. <https://doi.org/10.1038/sdata.2016.18>