scientific data

OPEN

DATA DESCRIPTOR

A corpus and a modular infrastructure for the empirical study of (an)notated music

Johannes Hentschel[1]✉, Yannis Rammos[2], Markus Neuwirth[1] & Martin Rohrmeier[2]

The present corpus is the outcome of a long-term collaborative effort to produce analytically annotated music scores suitable for the computer-assisted study of European compositions since 1600. With 1283 analytically annotated, symbolically encoded music scores by 36 composers, our corpus amounts to one of the largest published resources of its kind. At the same time, it provides a modular digital infrastructure for the accountable, collaborative curation of annotated scores ("sheet music"). All annotations were created and reviewed by a team of trained music theorists, who collaborated online using the `git` version control software according to a formally codified workflow. To improve the consistency of analytical practices given the diversity of represented eras and genres, the corpus has been automatically parsed for notational well-formedness and cross-reviewed by annotators for adherence to our music-analytical guidelines. The computational infrastructure has been designed with "data persistence" and open access in mind.

Background & Summary

The present corpus, hitherto referred to as *Distant Listening Corpus* (DLC)[1], is the outcome of a long-term collaborative effort to produce analytically annotated music scores suitable for the computer-assisted study of European music composition since 1600. By and large, notated music remains the primary source of evidence in theorizing Western music: (a) Scores are the sole form in which compositions survive as ontologically immutable entities, and indeed as historical documents (whether as manuscripts or "sheet music"); (b) Common Western Music Notation (CWMN) represents symbolic abstractions of a virtual infinity of potential performances and recordings; (c) CWMN establishes musical structures that may be aesthetically, cognitively, or perceptually salient but entirely irretrievable from the audio signal of a performance; (d) for centuries, score notation and annotations have been the principal means of communicating ideas about Western music in scholarship and pedagogy. The development of the DLC[1] was originally motivated by the broader discourse of "computational musicology" writ large, including prominent claims of an "empirical turn" towards corpus-driven music theory[2–7]. The corpus is equally in the service of an emergent body of scholarship on computer-assisted music analysis, music theory pedagogy, and "close reading" approaches[8,9].

Naturally, the size and availability of corpora of symbolically encoded digital music scores (e.g. MusicXML, MuseScore, or MEI) are indispensable for the computational study of such prominent topics as tonality and its historical transformations from the turn of the 17th century to this day[10–13]. Such sources may take the form of *comprehensive encodings*, i.e., digital editions that aim to represent the original manuscript or an authoritative engraving (*Urtext*) as faithfully as possible, or *selective encodings* confined to subsets of score elements (say, notes and rests only), which may be best represented with tables, token sequences, matrices, or graphs rather than music notation. Comprising ca. 1,370 comprehensively encoded scores of European compositions from the late 16th century until World War II, the DLC[1] is presently among the largest published resources of its kind.

Many central categories in music scholarship, including harmony, evade direct, unambiguous observation from scores—or from sound for that matter. While they are understood as intrinsic to a composition, their ontological status "within" it is the product of a listener's interpretive act, whether deliberate or reflexive. A common technique for making latent properties of human artifacts available for computational investigation is digital annotation[14–22]. Among the musical dimensions that are typically expressed and studied by means of

[1]Anton Bruckner University, Institute for Theory and History, Linz, 4040, Austria. [2]École Polytechnique Fédérale de Lausanne, Digital and Cognitive Musicology Lab, Lausanne, 1015, Switzerland. ✉e-mail: johannes.hentschel@bruckneruni.at

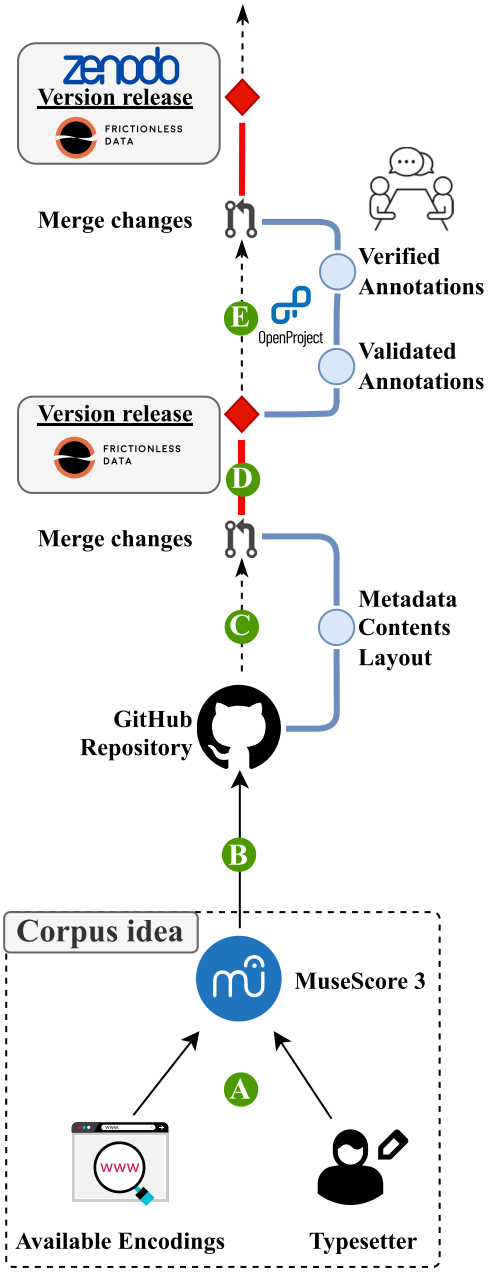

Fig. 1 Corpus curation pipeline. It begins at the bottom with **(A) the score procurement** according to the idea underlying the corpus. The collected scores are added to **(B) a corpus repository** for which a new **(D) version release** is issued each time changes are merged into the `main` branch. Such changes are either concerned with **(C) data curation** tasks or with our **(E) annotation workflow**.

annotations—both in corpus studies and in the analytical practice of music theorists and musicologists—, harmony, voice leading, and musical form are the most common. To date, 1,283 DLC[1] scores have been annotated with ca. 240,000 cross-reviewed labels that represent comprehensive harmonic analyses, also capturing aspects of voice leading, modulations ("key changes"), and certain form-related categories (keys, phrases, and cadences).

The 40 sub-corpora contained in the DLC[1] have been initiated, annotated, notationally validated, musically cross-reviewed, and archived according to the sequence ("pipeline") schematized in Fig. 1 and described in the Methods section. The schema lays out a distributed corpus-creation workflow, which was familiar to all contributors in advance, and is defined in a publicly available online document (dcmlab.github.io/standards/pipeline). The purpose of this document is: to facilitate the curation process by providing a step-by-step guide for the curator, to guarantee notational consistency and computational interoperability of the corpora, and to ensure that our corpus, considered as data, is compliant with the FAIR principles[23,24] at the moment of its public release, even when curated by multiple contributors at different times. The "pipeline" segment responsible for the collaborative creation, validation, and verification of analytical annotation labels has been introduced in greater detail in Hentschel *et al.*[25]. Based on the version control software `git`[26] and the collaboration platform

Github, the pipeline aims to ensure transparency in the genesis of each annotated score, allowing users to trace the authorship and revision history of each annotation.

From a research-planning point of view, any corpus-development project needs to pursue a balance between precision, size, and curation costs —especially when human annotations are involved[27]. To this end, our "pipeline" includes a software layer which facilitates non-interpretive curatorial tasks through automation. For example, every new version is automatically archived on the Zenodo platform which guarantees its survival for at least 20 years.

Methods

The DLC[1] provides curated research data in the form of music scores in MuseScore format, which contain well-formed, musically cross-reviewed analytical annotations. The data is partitioned into sub-corpora, each of which takes the form of a `git` repository with its own version history. Terminologically, we call 'corpus' any collection of physical or digital materials which is unified according to a meaningful musicological criterion, or representative of a larger collection (e.g., the style of J.S. Bach's four-voice chorales)[28,29]. Any corpus can serve as 'sub-corpus' to a 'meta-corpus'; the latter may include, for example, multiple repositories which together represent the style of German four-voice chorale writing before 1750. In that sense, we consider the DLC[1] as, to some extent, representative of certain aspects of European composition since 1600. In the following sections we describe the curation "pipeline" (Fig. 1) that the individual sub-corpus repositories underwent.

Score procurement. Symbolized by rectangle (A) at the bottom of Fig. 1, each sub-corpus came into existence as an underlying idea of what it should represent. Such an idea often takes the form "composer X's output for instrumentation(s) Y and/or in genre Z" and therefore usually consisted of a group of compositions by a single composer. Based on this idea, a curator compiled a list of desired pieces for the sub-corpus by drawing from one or several work catalogs, potentially taking into account which pieces are already available as digital symbolic encodings with permissive licenses. Multiple pathways could then lead from this list to an organized collection of files in the target format —one per composition or part of work— each of which is guaranteed, within the margins of editorial error, to correspond to its original source (that is, to a scan of a manuscript or a print edition, preferably of scholarly quality, from the public domain). Pieces for which symbolic encodings were unavailable under an open license or inaccurate were commissioned from typesetters who engraved them in MuseScore 3 transcribing a reliable printed source. (This yields the best results in terms of accuracy, but also entails the highest curation cost.) In all other cases, whether pieces that were already available in reliable digital editions or outputs of Optical Music Recognition (OMR) software, the symbolic encodings needed to be compared carefully to the original source. (Although this is a less costly avenue, in our experience it is also more error-prone and tedious, especially as mistakes generally abound in automatic conversions between formats[30]). Since some errors would require painstaking proofreading of the source code of the encoding, and are invisible in a rendered score (e.g., a slur masquerading as a tie), creating a symbolically encoded score from scratch in a format that does not require conversion is clearly the preferable option.

Once a collection of MuseScore files for the sub-corpus was compiled, the files were then renamed using an appropriate naming scheme. Ideally, such a scheme

- enables the meaningful ordering of files when sorted lexically (e.g., by using leading zeros as in `001`);
- contains components that allow music researchers to recognize each piece (such as catalog numbers);
- is structured in a way that makes its components retrievable with a relatively simple regular expression;
- uses a minimal set of characters and no spaces (e.g., no diacritics, ligatures, or special characters other than dashes and underscores).

It proved beneficial to settle on a robust naming scheme at an early point in the development of the DLC[1] because, ideally, the name of a file should not change during its lifetime for reasons of traceability. As explained below, DLC file names also serve as identifiers grouping together the various files which represent any single piece. Once the scores were renamed, a new `git` repository for the sub-corpus was initialized from a template.

The corpus repository. The history of each sub-corpus in the DLC[1] starts with the initialization of a `git` repository (segment (B) of Fig. 1) whose name serves as corpus identifier. A template was used to initialize new sub-corpus repositories; it contains auxiliary files and various template files. The former control the automated aspects of our curation pipeline, whereas the latter serve as templates for metadata files (such as a descriptive `readme` file), the placeholders ("variables") of which can be automatically filled-in for each new corpus. Once registered with our project management infrastructure and uploaded to GitHub, the given sub-corpus was ready to undergo the remainder of our curation pipeline.

Curatorial guidelines. The guidelines described in this section are extracted from a publicly available web document (dcmlab.github.io/standards/pipeline) that all collaborators were familiar with. Apart from spelling out concrete steps to prepare a new sub-corpus as described above, they provide details on the metadata fields ("keys") that need to be encoded for every score, the format these values should have, and the use of the `ms3` parser[31] to batch-update multiple MuseScore files at once. Although in Fig. 1 these curatorial errands are indicated with the sign of a single commit (segment (C)), they are in fact repeated within an iterative process that is generally distributed among multiple merge requests throughout the revision history of a sub-corpus repository. It is for this reason that the guidelines conclude with a long section on corpus finalization, which can indeed serve as a checklist of tasks prior to public release. Since most repositories remain private until that point, this section

of the guidelines also outlines the activation of automatic uploads to CERN's data registry zenodo.org, a process which also assigns a Digital Object Identifier (DOI) to each published revision of the sub-corpus.

Version releases. The auxiliary files that are part of each sub-corpus repository, and indeed of the repository template, configure pre-commit hooks and webhooks. The former support our annotation workflow (see below), whereas the latter automate the creation of version releases. A "version release" is a "snapshot" of the sub-corpus at a given point in its revision history, labeled with a version tag. The webhook creating such a snapshot is triggered each time a `git` branch is merged into a repository's `main` branch (symbolized by segment (D) in Fig. 1). Putting to use GitHub's continuous-integration platform ("GitHub Actions"), such a merge event entails the creation of a virtual machine (a "runner") on which the following steps are executed:

1. The sub-corpus repository is checked out (including any submodules).
2. The next version number is determined based on the latest version tag and the presence of particular labels attached to the merge request.
3. A separate account ("bot") is used to create a `git` commit with the updated metadata, auxiliary files, and version tag.
4. The parsing library `ms3` is installed on the runner and used to generate a so-called "Frictionless data package"[32], which consists of a ZIP archive and a JSON descriptor file, and represents a snapshot of the sub-corpus.
5. The new version is released on GitHub, stating the title and a description of the merge request, and including the datapackage as a release asset (alongside an error report if the validation of the data package failed).
6. If the repository is public, the release will also propagate to the Zenodo platform where a new DOI will be minted.

The DLC[1] adopts the practice of "semantic versioning", reduced to two positions only (without the patch level). By default, the minor-version number is increased by one (for instance, from `v2.11` to `v2.12`). Only substantial, potentially "breaking" changes warrant a new major-version number (e.g., from `v2.11` to `v3.0`). By "breaking changes" we construe those that may render existing evaluation code nonfunctional—for example due to renamed or removed files—or that may substantially change previous evaluation results, for example after adding missing bars to a score. A new major version of a sub-corpus therefore implies that the next update of every parent meta-corpus (and every respective meta-repository, which contains the sub-corpus as a `git` submodule), will also result in a new major version. The first release typically includes the scores in their original state without annotations, as delivered by the engraver or produced by the OMR software, alongside the basic set of metadata available at that stage (e.g., the name or identifier of the engraver). Subsequent versions in the history correspond either to curatorial errands (addressed in the previous section), or to annotation label additions and revisions, as explained below.

Annotation workflow. The DCML annotation workflow is the backbone of a formally defined framework for the collaborative, accountable creation and curation of annotation labels. Since the task of building and maintaining a corpus of well-formed, cross-reviewed expert annotations quickly surpasses the capacities of a single curator, the workflow has been designed to distribute the process among several actors, while pursuing a coordination of automated repetitive tasks (such as annotation "spelling checks") and the interpretive work of music theorists (individually and in online discussions). The workflow has been formally described in Hentschel *et al.*[25] and, since then, seen some of its components move from the cloud (*GitHub Actions*) to the annotator's personal computer (pre-commit hooks, see pre-commit.com), without changes to its algorithmic underpinnings. Segment (E) in Fig. 1 schematizes the transition of a not-yet-annotated piece (or one whose annotations warrant corrections) to a fully annotated piece containing a complete set of labels that have been:

- algorithmically validated as well-formed and thus guaranteed to be processable without parsing errors;
- cross-reviewed through consensus between at least two different music theorists, thus compliant with the annotation guidelines.

The workflow iterations for the pieces of a corpus were performed on parallel `git` branches—one per piece— and resulted in a new version release when merged (as discussed in the previous section). The blue logo in the schema stands for the open-source project management software OpenProject, which we used to solicit annotations and reviews from our music theorists, easily tally the amount of work carried out, and facilitate invoicing.

As outlined in the previous section, the latest version of a sub-corpus is stored as a "Frictionless data package". Each package has the form of a ZIP file containing five tabular files (TSV), alongside a JSON-formatted descriptor with metadata on these tables, including the data type of each column. For any given sub-corpus, and for the DLC[1] corpus as a whole, these aggregate TSV files are concatenations of the respective piece-specific files previously extracted.

Data Records

The DLC[1] is available at Zenodo (https://doi.org/10.5281/zenodo.15150283), with the present section being the primary source of information on the availability and content of the data described. Each of the 40 constituent sub-corpora of the dataset is intended to represent a particular European composition style, spanning from the late 16th century to World War II. Of the ca. 1,330 encoded scores that it currently comprises (roughly 1.7 million notes), 1,283 are annotated with comprehensive analyses of harmony, voice-leading, and form-related

Sub-corpus	Pieces	Measures	Length	Notes	Labels
Bach Solo	68	3849	12465	36365	7181
Bach Suites	89	3580	12607	56398	11312
Bartok Bagatelles	14	896	2307	9469	1191
Beethoven Sonatas	64	11662	35871	162068	21963
Beethoven String Quartets	70	15731	48767	232800	28089
C Schumann Lieder	12	509	1464	10248	1326
CPE Bach Keyboard	66	5129	14386	63827	11191
Chopin Mazurkas	55	5089	14658	55966	9125
Corelli Trio Sonatas	149	4777	18145	68143	14314
Couperin Clavecin	9	230	779	3422	717
Couperin Concerts Royaux	84	2945	9834	34887	8755
Debussy Suite Bergamasque	4	421	1616	7772	1013
Dvor^ák Silhouettes	12	674	1852	10105	1539
Frescobaldi Fiori Musicali	47	1486	10022	18503	5119
Grieg Lyric Pieces	66	5414	16496	62431	8236
Handel Keyboard	6	53	221	1701	350
JC Bach Sonatas	29	2416	8306	37688	5063
Kozeluh Sonatas	49	6946	20392	108518	16598
Liszt Années	19	2625	9709	58553	5070
Mahler Kindertotenlieder	5	438	1834	5582	595
Medtner Tales	19	2464	6604	41144	6508
Mendelssohn Quartets	24	6253	22123	96071	14758
Monteverdi Madrigals	19	1568	8545	16899	3289
Mozart Piano Sonatas	54	7488	22445	103553	15272
Pergolesi Stabat Mater	7	470	1139	6772	1189
Peri Euridice	6	1120	9015	11009	2884
Pleyel Quartets	6	823	3038	13838	1567
Poulenc Mouvements Perpetuels	3	94	319	1527	278
R Schumann Kinderszenen	13	392	940	5012	948
R Schumann Liederkreis	12	495	1416	9522	892
Rachmaninoff Piano	22	415	1541	9880	1141
Ravel Piano	3	457	1452	13465	861
Scarlatti Sonatas	69	5560	13706	64287	12490
Schubert Winterreise	24	1417	4052	26331	3100
Schulhoff Suite Dansante En Jazz	6	286	1007	5390	488
Schütz Kleine Geistliche Konzerte	55	5573	24698	47554	11709
Sweelinck Keyboard	1	196	784	2639	501
Tchaikovsky Seasons	12	1250	3919	18169	3059
WF Bach Sonatas	9	631	1914	8551	1753
Wagner Overtures	2	333	1221	7056	1433
sum	**1283**	**112155**	**371621**	**1553115**	**242867**

Table 1. Extent of the DLC[1] by sub-corpus. Lengths are expressed in quarter notes and represent the summed duration of the pieces in a given corpus.

aspects by trained music theorists. Table 1 summarizes, for each of the 40 constituent sub-corpora, the number of pieces, measures, notes, and annotation labels included (for a more detailed description, see below), as well as the overall length in quarter notes. Figure 2 shows the DLC's[1] dimensions on the historical timeline, aggregated to differentiate between the 12 sub-corpora already published in four previous reports[25,33–35] and the 28 new sub-corpora that the DLC now makes available to the public. The four previously published datasets are the *Annotated Beethoven Corpus* (https://doi.org/10.5281/zenodo.7441343), the *Annotated Mozart Sonatas* (https://doi.org/10.5281/zenodo.7424962), *36 Trio Sonatas by Arcangelo Corelli* (https://doi.org/10.5281/zenodo.7504011), and *An Annotated Corpus of Tonal Piano Music from the Long 19th Century* (https://doi.org/10.5281/zenodo.7483349).

Like its constituent sub-corpora, the DLC[1] is provided in the form of a "Frictionless data package", that is, a ZIP file containing five TSV-formatted tables alongside a JSON descriptor of their content types. Each of the following subsection describes one of these TSV files. The exact meaning and data type of each column are also codified, aside from the aforementioned JSON descriptor, in the documentation of the parsing library `ms3` (ms3.readthedocs.io/columns).

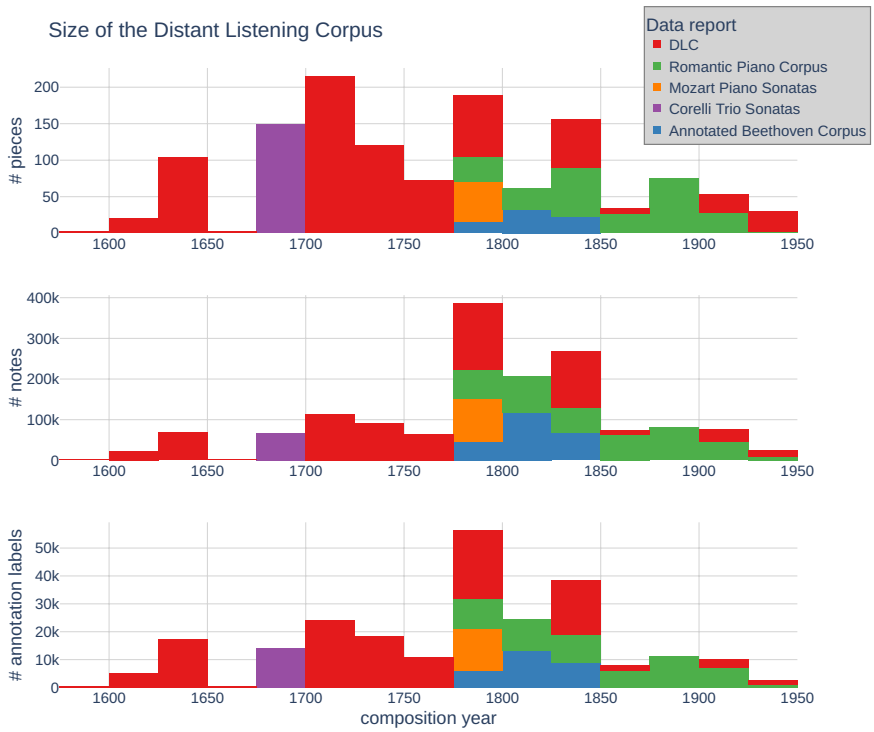

Fig. 2 Size and temporal coverage of the DLC[1]. Colors indicate which parts of the dataset are made public for the first time with this publication (red) and which ones have been described in other data reports, namely the Annotated Beethoven Corpus[33] (blue), the Mozart Piano Sonatas[34] (orange), Arcangelo Corelli's Trio sonatas[25] (violet), and the Romantic Piano Corpus[35] (green).

Metadata. The file `distant_listening_corpus.metadata.tsv` contains one row per piece. The columns of this file can be grouped into several categories. The first group compiles information about the annotated score file such as its path, its ID, and the ID of its sub-corpus. The IDs are relevant in correctly attributing information from different TSV files to the relevant piece. A second group of columns provides descriptive statistics about each score, such as its number of measures, its total length, its ambitus (i.e., the lowest and highest note), the number of notes and annotation labels, or the time and key signatures. A third group of columns corresponds to the default metadata fields of the MuseScore software. These include fields for identifying the piece (such as 'composer', 'movementTitle', or 'workNumber') or the encoded edition (such as 'source' or 'copyright'). The last group consists of custom metadata keys, which we add to enrich our corpus. (These fields can equally be managed within the MuseScore environment.) These include, in particular, the names of annotators and reviewers, version numbers of the annotation standards applied and software used, and URIs (e.g., from viaf.org, www.wikidata.org, or musicbrainz.org, depending on availability) that identify the encoded music within a Linked Open Data paradigm. Taken together, this TSV file provides means for locating corpus files, for grouping them according to their properties, and for integrating them with the "semantic web" through URIs.

Measures (bars). The file `distant_listening_corpus.measures.tsv` contains one row per measure per piece (more precisely, each row represents a *notated* measure, which may be different from a *logical* one, as outlined in Gotham *et al.*[36]). Measures (British English: "bars") represent musical segments and play an important role in CWMN, especially with respect to the metric and rhythmic organization of the notated music. Such a division of the score in short numbered segments helps locate ("addressing") musical events, both in communication between musicians and in the context of machine encodings, such as those of the DLC[1]. The most characteristic properties of a measure are arguably its position within the score (ordinal number), its length in notated musical time (note values), and its time signature (which has important implications for the perception of rhythm within the measure). A measure's number is an integral component of every "timestamp" in the corpus. It is essential for unambiguously locating the temporal onset of all score elements that are described in the remaining three TSV files. Whereas musicians commonly locate score symbols in terms of a measure number and a small relative offset from the beginning of that measure (e.g., "measure 10, beat 2"), some computer programs do so in terms of large absolute offsets from the beginning of the score. For this reason, the column 'quarterbeats' contains such absolute offsets, from which relative offsets can be derived. Apart from these two ways of expressing temporal information, the remaining columns serve to express the repetition structure of a piece as encoded in the MuseScore file. This structure is defined by repeat signs, jump marks (such as *da capo al segno*), section breaks, and alternative endings. In order to decode this information, users are advised to consult the column 'next' which contains, for each measure, the number of each measure that can follow it. The processing library `DiMCAT`[37] uses

this column whenever the user requests an 'unfolded' version of a table which reflects the series of events in a performance, including all repeats and jumps, rather than a mere succession of measures as printed in the score.

Notes. The file `distant_listening_corpus.notes.tsv` contains one row per note head per piece. Apart from a timestamp, each note is also defined by its tonal pitch class (note name), MIDI pitch (key on the piano), and duration. The columns 'staff' and 'voice' indicate the notational layer in which a note occurs, depending on the specific instrument in question, while 'chord_id' links the note to all other vertically aligned score elements—including other note heads—in the same layer. Additional note properties include various types of grace notes, the presence of ties to other notes, or membership in a tremolo.

Other score elements. The file `distant_listening_corpus.chords.tsv` comprises all score elements that are neither notes, nor annotation labels. They have been included for the sake of completeness and can be filtered on the column 'event' for a given use case. The main event category is 'Chord' which corresponds to a set of at least one note sharing the same 'chord_id' and occurring at a given position in the same notational layer. This is relevant because MuseScore, in adherence with music notation and engraving rules, attaches articulation semantics (such as staccato) and lyrics to an entire set of notes rather than the individual note heads. Other event types include 'Dynamic' (dynamic markings), 'FiguredBass' (thoroughbass figures), 'Spanner' (symbols extending across notated timespans, such as crescendo wedges, pedal, or *all'ottava* markings), 'StaffText' (text concerning a single staff, e.g. *espressivo*), 'SystemText' (text concerning all staves, such as section titles), and 'Tempo' (metronome markings and verbal tempo indications). These events can appear at score positions independent of notes, and are represented in the chord tables with separate sets of columns that define their properties.

Annotation labels. The file `distant_listening_corpus.expanded.tsv` contains one row per annotation label per piece. The DLC[1] contains 242,867 labels which adhere to the DCML harmony annotation standard[33–35]. This standard allows annotators to encode their music-theoretically grounded analyses in plain-text labels that obey the standard's syntax and therefore can be parsed into its accompanying tabular model. The annotation syntax is defined as a regular expression; we consider a label valid if and only if it matches this expression (see the section on data validation below). At the same time, the parsing library `ms3` uses the capture groups of this regular expression to split each label into its various components and organize them in separate columns. Apart from the full annotation labels, their timestamps, and their parsed components, the TSV files described here contain additional columns which are computationally derived from these features, most notably the chord tones implied by each label.

On a theoretical level, the DCML harmony annotation standard allows music analysts to account for four distinct layers —keys, chords, phrases, and cadences—which correspond to the four main components of the regular expression. Granted that in actual analytical practice these features are highly interdependent, and that an annotator would typically determine them collectively, as pieces of information they can be named and annotated independently of each other, and are amenable to "tabularization". Indeed, these four layers have their own set of columns and are therefore easy to inspect separately. To retrieve more compact representations of an individual layer, users may consult the `DiMCAT` library, which can create them.

Key segments. Inasmuch as the DLC[1] is a corpus of tonal music, we decided to annotate each constituent piece with a proposed segmentation into global and local tonal centers (or "keys"). In doing so, we acknowledge the prominence of keys in musical discourse while being aware of the high degree of subjectivity involved in local-key identification (e.g.,[38]). Key annotations are given in the columns 'globalkey' and 'localkey' and represent contiguous segmentations (i.e., one without gaps) of a piece. Since the actual tones implied by a chord label (see below) are inferrable only in relation to these two columns, the applicable global and local keys are repeated in every row, even though, strictly speaking, they are only necessary when their value changes.

Chord symbols. Among the 242,867 annotation labels, 240,131 (98.89%) are chord symbols. They partition scores into segments, each of which was identified by the annotators as an expression of a distinct harmonic entity. Accordingly, chord labels are provided in the TSV file with their timestamps and their durations, as derived from the timestamps of their successor label. Chord labels must always match the relevant part of the regular expression used in our automatic validation scripts. They adhere to a Roman-numerals-based standard of harmony annotation that aims at a high degree of analytical detail with regard to non-triadic tones, non-diatonic chord notes, and various advanced harmonic phenomena.

In our tabular format the structured labels appear both in their original form (column 'chord'), and separated into their parsed syntactical components. The choice of representation depends, obviously, on the research question and ergonomic considerations (musicians may find the compact labels more readable, whereas computational analysts may prefer fine-grained parsed information). Label rows are also enriched with additional information that is derived from the label components. An additional set of columns represents chord labels as collections of the scale degrees (or "scale steps") that they represent, thus facilitating their matching with notes in the corresponding score segments. Scale degrees are of central importance to various schools of music analysis, as they help attribute, in multifarious ways, specific functions to each note in the score.

Cadences. Cadence labels highlight points of arrival or closure, or articulatory moments where the expectation of such closure was thwarted. In most cases, they coincide with the last harmony label within a harmonic progression that is interpreted by the annotators as "cadential". Similarly with keys, the annotation of cadences is a highly interpretive and theoretically charged process. Our annotation guidelines are generally aligned with

contemporary "textbook music theory", which calls for a consideration of the interaction between key, harmonic content, phrase length, melodic motion, and textural features in identifying cadences. Cadence labels are presented in the column 'cadence'.

Phrases. Phrase annotations in the DLC[1] are contiguous segmentations of a score into musical spans that exhibit some kind of closure. Attempting to circumvent prescriptive theoretical criteria of phrase boundaries, we decided to defer to our annotators' judgment. Generally speaking, phrases in the DLC are meant to correspond to an underlying sense of harmonic and melodic closure. As with keys and cadences, certain phrase annotations in the corpus may raise questions for some music theorists, and indeed for some of the authors of this report. Nonetheless, in the interest of integrity, we decided not to tamper with the annotators' decisions and to refrain from ad hoc departures from the formal workflow, which entrusts analytical decisions exclusively to the annotators. The 1,283 annotated pieces in the DLC[1] have been split by the annotators into 15,576 phrases, as expressed by column 'phraseend'. Phrase annotations may be used to process corpus contents in the form of smaller, simpler yet syntactically coherent musical enunciations.

Technical Validation

The scores and the ca. 240,000 labels they contain have been automatically validated as well-formed by a parser, and cross-reviewed for musical coherence by our team of music theorists, in adherence to our annotation guidelines. The annotation workflow triggers the `ms3 review` command before each `git` commit, so that any modified scores and annotation labels therein are checked for certain notational and musical inconsistencies. For the scores, such errors may involve inconsistent repeat signs, alternative endings, or measure numbers. Annotation labels that do not match the standard's regular expression are rejected as ill-formed. In cases when an annotator or curator decided to commit an inconsistency to the `git` history, this is documented in an additional text file with the file extension `.warnings`. If, on the other hand, a warning was deemed a false positive, it is copied into a text file called `IGNORED_WARNINGS`, alongside a human-readable rationale for the exception. Doing so suppresses the warning in the future. This was particularly relevant for warnings resulting from a particular class of automated validations, in which the chord tones implied by a chord annotation must sufficiently match the notes in the corresponding score segment: Our algorithms automatically reject a chord label if the aggregate duration of the pitches that it accounts for is less than 60 % of the total duration of all pitches in the respective score segment. We arrived at this threshold through a combination of trial-and-error and music-theoretical considerations that are beyond the scope of this report.

A validation step also accompanies the concatenation of TSV files into a "Frictionless data package". Once these files are stored (whether in a ZIP archive or as individual "Frictionless resources") and the JSON descriptor is created, the latter is used to validate the former. This validation steps ensures that the columns of the corpus TSV files are of the correct data type and format. Validation errors encountered at this step are stored in a separate text file with the extension `.errors`.

Usage Notes

Users of the DLC[1] may inspect its contents by opening individual scores via the open-source score editor MuseScore (musescore.org). In addition, the relevant information is extracted from the annotated scores and provided in the tabular form of tab-separated-value (TSV) files to facilitate further computational work. Each TSV file is accompanied by an aforementioned machine-readable descriptor file in JSON format, which specifies the meanings and data types of the table fields. The Frictionless specification[32] provides instructions for loading and validating the tables. The open-source library `DiMCAT`[37] can load the DLC[1] and may be used to explore its contents or carry out further research with it. In fact, `DiMCAT` is used to automatically generate the static homepage and interactive figures that supplement each DLC sub-corpus (see, for example, https://dcmlab.github.io/chopin_mazurkas/).

Code availability

The code that implements the automated pipeline steps is part of the corpus repository template and available at github.com/DCMLab/annotation_workflow_template. Both the pre-commit hook and the "GitHub Actions" webhook make use of the MuseScore parsing library `ms3`, which is published as Hentschel & Rohrmeier (2023)[31] and available via pypi.org/project/ms3. The entire infrastructure, which incorporates the DLC[1] alongside private sub-corpora at various stages of development, is managed via a central "project management" repository that is not open to the public. It includes code for executing certain maintenance tasks on selected or all sub-corpora, such as updates to their workflow components or static homepages. As already mentioned, the latter task involves the Digital Musicology Toolkit `DiMCAT` which has been published as Hentschel *et al.*[37] and is available via pypi.org/project/dimcat.

Received: 23 October 2024; Accepted: 9 April 2025;
Published online: 23 April 2025

References

1. Hentschel, J., Rammos, Y., Neuwirth, M. & Rohrmeier, M. The Distant Listening Corpus. *Zenodo* https://doi.org/10.5281/zenodo.15150283 (2025).
2. Huron, D. The new empiricism: Systematic musicology in a postmodern age. Music and Mind: Foundations of Cognitive Musicology 3, The 1999 Ernest Bloch Lecture, University of California, Berkely (1999).
3. Temperley, D. *Music and Probability* (The MIT Press, 2007).
4. Volk, A., Wiering, F. & van Kranenburg, P. Unfolding the potential of computational musicology. In *Proceedings of the 13th International Conference on Informatics and Semiotics in Organisations*, 137–144 (Leeuwarden, 2011).

5. Meredith, D. (ed.) *Computational Music Analysis* (Springer, New York, 2016).
6. Mor, B., Garhwal, S. & Kumar, A. A systematic literature review on computational musicology. *Archives of Computational Methods in Engineering* **27**, 923–937, https://doi.org/10.1007/s11831-019-09337-9 (2020).
7. Shanahan, D. What the history of computational musicology can tell us about the future of corpus studies. *Future Directions of Music Cognition Virtual Speaker Series* https://doi.org/10.17605/OSF.IO/CB26R (2022).
8. Snarrenberg, R. Westerparse: Species counterpoint online. https://westerparse.readthedocs.io/en/latest/ (2020).
9. Marsden, A. Schenkerian analysis by computer: A proof of concept. *Journal of New Music Research* **39**, 269–289, https://doi.org/10.1080/09298215.2010.503898 (2010).
10. Schreibman, S., Siemens, R. G. & Unsworth, J. (eds.) *A Companion to Digital Humanities*. No. 26 in Blackwell Companions to Literature and Culture (Blackwell Pub, Malden, MA, 2004).
11. Huron, D. On the virtuous and the vexatious in an age of big data. *Music Perception* **31**, 4–9, https://doi.org/10.1525/mp.2013.31.1.4 (2013).
12. Shanahan, D., Burgoyne, J. A. & Quinn, I. (eds.) *The Oxford Handbook of Music and Corpus Studies* (Oxford University Press, Oxford, 2022).
13. White, C. *The Music in the Data: Corpus Analysis, Music Analysis, and Tonal Traditions* (Routledge, New York, 2022).
14. Unsworth, J. Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this? (2000).
15. Ide, N. Preparation and analysis of linguistic corpora. In Schreibman, S., Siemens, R. & Unsworth, J. (eds.) *A Companion to Digital Humanities*, 289–305, https://doi.org/10.1002/9780470999875.ch17 (Blackwell Publishing Ltd, Malden, MA, USA, 2004).
16. Smith, J. B. L., Ashley Burgoyne, J., Fujinaga, I., De Roure, D. & Downie, J. S. Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, 555–650 (Miami, Florida, 2011).
17. Borek, L., Dombrowski, Q., Perkins, J. & Schöch, C. Tadirah: A case study in pragmatic classification. *Digital Humanities Quarterly* **10**, (2016).
18. Hughes, L., Constantopoulos, P. & Dallas, C. Digital methods in the humanities: Understanding and describing their use across disciplines. In Schreibman, S., Siemens, R. & Unsworth, J. (eds.) *A New Companion to Digital Humanities*, 150–170, https://doi.org/10.1002/9781118680605 (John Wiley & Sons, 2016).
19. Rizo, D. & Marsden, A. A standard format proposal for hierarchical analyses and representations. In *Proceedings of the 3rd International Workshop on Digital Libraries for Musicology*, 25–32, https://doi.org/10.1145/2970044.2970046 (ACM, New York USA, 2016).
20. Terras, M. Crowdsourcing in the digital humanities. In Schreibman, S., Siemens, R. & Unsworth, J. (eds.) *A New Companion to Digital Humanities*, 420–438, https://doi.org/10.1002/9781118680605 (John Wiley & Sons, 2016).
21. Gotham, M. & Ireland, M. T. Taking form: A representation standard, conversion code, and example corpus for recording, visualizing and studying analyses of musical form. *20th International Society for Music Information Retrieval Conference, Delft* 633–699 (2019).
22. Weigl, D. M. *et al*. Notes on the music: A social data infrastructure for music annotation. In *8th International Conference on Digital Libraries for Musicology*, DLfM '21, 23–31, https://doi.org/10.1145/3469013.3469017 (Association for Computing Machinery, New York, NY, USA, 2021).
23. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, https://doi.org/10.1038/sdata.2016.18 (2016).
24. Arguillas, F., Christian, T.-M., Gooch, M., Honeyman, T. & Peer, L. 10 things for curating reproducible and fair research, https://doi.org/10.15497/RDA00074 (2022).
25. Hentschel, J., Moss, F. C., Neuwirth, M. & Rohrmeier, M. A. A semi-automated workflow paradigm for the distributed creation and curation of expert annotations. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 262–269, https://doi.org/10.5281/ZENODO.5624417 (Zenodo, Online, 2021).
26. Chacon, S. & Straub, B. *Pro Git*. The Expert's Voice in Software Development (Apress, New York, NY, 2014).
27. Stefanowitsch, A. *Corpus Linguistics: A Guide to the Methodology* (Language Science Press, Berlin, 2020).
28. Biber, D. Representativeness in corpus design. *Literary and Linguistic Computing* **8**, 243–257 (1993).
29. Piotrowski, M. Historical models and serial sources. *Journal of European Periodical Studies* **4**, https://doi.org/10.21825/jeps.v4i1.10226 (2019).
30. Nápoles López, N., Viglianseni, G. & Fujinaga, I. The effects of translation between symbolic music formats: A case study with humdrum, lilypond, mei, and musicxml. In *Music Encoding Conference* (Vienna, 2019).
31. Hentschel, J. & Rohrmeier, M. ms3: A parser for MuseScore files, serving as data factory for annotated music corpora. *Journal of Open Source Software* **8**, 5195, https://doi.org/10.21105/joss.05195 (2023).
32. Fowler, D., Barratt, J. & Walsh, P. Frictionless data: Making research data quality visible. *International Journal of Digital Curation* **12**, 274–285, https://doi.org/10.2218/ijdc.v12i2.577 (2018).
33. Neuwirth, M., Harasim, D., Moss, F. C. & Rohrmeier, M. The Annotated Beethoven Corpus (ABC): A Dataset of Harmonic Analyses of All Beethoven String Quartets. *Frontiers in Digital Humanities* **5**, 1–5, https://doi.org/10.3389/fdigh.2018.00016 (2018).
34. Hentschel, J., Neuwirth, M. & Rohrmeier, M. The annotated Mozart sonatas: Score, harmony, and cadence. *Transactions of the International Society for Music Information Retrieval* **4**, 67–80, https://doi.org/10.5334/tismir.63 (2021).
35. Hentschel, J., Rammos, Y., Neuwirth, M., Moss, F. C. & Rohrmeier, M. An annotated corpus of tonal piano music from the long 19th century. *Empirical Musicology Review* **18**, 84–95, https://doi.org/10.18061/emr.v18i1.8903 (2024).
36. Gotham, M. *et al*. The 'Measure Map': An inter-operable standard for aligning symbolic music. In *Proceedings of the 10th International Conference on Digital Libraries for Musicology*, 91–99, https://doi.org/10.1145/3625135.3625136 (Milan, Italy, 2023).
37. Hentschel, J., McLeod, A., Rammos, Y. & Rohrmeier, M. Introducing DiMCAT for processing and analyzing notated music on a very large scale. In *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*, 516–523 (Milan, Italy, 2023).
38. Schachter, C. Analysis by key: Another look at modulation. *Music Analysis* **6**, 289–318 (1987).

Acknowledgements

This research was supported by the Swiss National Science Foundation within the project "Distant Listening - The Development of Harmony over Three Centuries (1700–2000)" (Grant no. 182811). This project was being conducted at the Latour Chair in Digital and Cognitive Musicology, generously funded by Mr. Claude Latour.

Author contributions

JH conceived the protocol and workflow toward the creation of this dataset, contributed reviews, implemented the workflow and data parser, developed the analysis library, trained annotators, and wrote the article. YR made contributions and emendations to JH's original draft of the article. JH and YR coordinated the pool of annotators, commissioning the creation and version upgrades of score corpora and annotation labels. MN and MR selected the corpora and supervised the entire project. MR made this publication possible through the acquisition of funds.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04976-z.

Correspondence and requests for materials should be addressed to J.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.