

A Genre Classification Scheme for Metal-Music Corpus Studies: An Eleven-Bucket and Seventeen-Category Encoding of Encyclopaedia Metallum Genre Strings

Ignacio Soto-Silva ^{1,2} 

¹ Departamento de Humanidades y Artes, Universidad de Los Lagos, Osorno 5290000, Chile; ignacio.soto@ulagos.cl

² Millennium Nucleus in Musical and Sound Cultures, CMUS, Santiago 8320000, Chile

Abstract

This data descriptor documents a two-level genre classification scheme for metal-music corpus studies derived from Encyclopaedia Metallum's multi-label genre strings. The scheme assigns each band to one of eleven mutually exclusive primary buckets and, alternatively, to one of seventeen finer-grained sub-categories that separate closely related sub-styles. Classification is implemented by keyword priority on the lower-cased genre string and is fully deterministic given the published keyword table. The scheme was developed and tested on a corpus of 560 metal bands from Southern Chile (1988–2024), which serves as the development corpus throughout. On a stratified 60-band subset of this corpus, automatic assignments agreed with independent expert human coding at 81.7% (Cohen's $\kappa = 0.78$, 95% bootstrap CI [0.66, 0.88]); an out-of-sample application to a 60-band Norwegian sample, with the keyword tables left unchanged, retained the scheme's core logic at a 5.0% residual rate. The encoding rules, the keyword tables, and the mapping CSVs are released under CC-BY 4.0 to enable replication and adaptation across other metal-scene corpora.

Dataset: Soto-Silva, I. (2026). A keyword-priority genre classification scheme for metal-music corpus studies: eleven-bucket and seventeen-category encoding tables with a classified southern Chilean band corpus [Dataset]. Zenodo. <https://doi.org/10.5281/zenodo.20222458>.

Dataset License: CC-BY 4.0 (Creative Commons Attribution 4.0 International)

Keywords: metal music; corpus studies; genre classification; Encyclopaedia Metallum; popular music empirical research; music scenes; data classification; categorical encoding



Academic Editor: Maarten Marx

Received: 16 May 2026

Revised: 15 June 2026

Accepted: 18 June 2026

Published: 28 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Summary

Music is a core human communicative system, closely entwined with the development of language [1,2]. Genre, the vocabulary through which this system is organized into recognizable styles, is correspondingly complex, and its representation in fan-curated databases is rich but difficult to analyze quantitatively. The present descriptor addresses one tractable corner of that problem: the structured encoding of metal-genre labels for corpus research.

Encyclopaedia Metallum: The Metal Archives (hereafter Encyclopaedia Metallum) is the most comprehensive single-source repository of metal-music metadata, covering more than 197,000 bands worldwide as of the access date (28–29 April 2026), a count that is continually revised as new entries are approved [3]. The database's strength as a

research resource has been recognized in earlier quantitative work on metal corpora [4,5]; its principal liability for empirical research, however, is the format in which it encodes genre information. Each band is tagged with one or more genre strings of the form “Black/Death/Pagan Metal” or “Symphonic/Atmospheric Black/Folk Metal”, with no fixed vocabulary, no maximum number of labels, and no separation between primary and secondary stylistic affiliations. The format is well-suited to fan-curated description but ill-suited to operations that require mutually exclusive, machine-readable categories—multinomial regression, cross-bucket comparison, network analysis with categorical edge attributes, or the construction of category–typology indicators of the kind employed in network studies of music scenes.

The problem is a particular case of a more general one: polysemic uses of categorical vocabularies, where the same term operates across multiple analytic levels and across scenes, are recurrent in the study of situated musical practice [6–8]. The free-text labeling of metal styles in Encyclopaedia Metallum is descriptively rich but analytically opaque, and corpus-based work in metal-music studies would benefit from a documented, replicable mapping from those strings to a fixed categorical vocabulary.

In the absence of such a standard mapping, individual researchers have constructed their own bucketing schemes ad hoc, with varying granularity and varying treatment of hybrid bands. The result is a body of literature in which numerical results on the genre composition of metal scenes are often not directly comparable across studies. This data descriptor contributes a fully documented, replicable encoding scheme—implemented at two levels of granularity, with the seventeen-category refinement preserving the stylistic complexity that the eleven-bucket primary scheme compresses—to address that gap.

The dataset was developed within the framework of two ANID-funded research projects: the FONDECYT Regular Project “Representaciones territoriales emergentes en la música popular del sur de Chile” and the Millennium Science Initiative Núcleo Milenio en Culturas Musicales y Sonoras (CMUS). The theoretical framing of the dataset’s territorial scope and its place within the broader research program on music, territory and identity in Southern Chile is developed in Soto-Silva, Zamora, and Silva-Zurita [6]; a first analytic paper based on the classified corpus and its associated musician–band network is in preparation. Beyond the analytic uses pursued by the present author, publishing the encoding rules as a citable dataset is intended to enable parallel work by other researchers on cross-scene comparison, sensitivity analyses across granularities, and the construction of musician-level multi-genre indicators for studies of brokerage and cross-genre mobility in metal scenes worldwide.

2. Data Description

The deposited dataset consists of nine files: The file (i) `keywords_11bucket.csv` lists the keyword-priority table for the primary eleven-bucket scheme, with columns `bucket`, `keyword`, `priority_rank`. The file (ii) `keywords_17category.csv` lists the keyword table for the seventeen-category refinement, with columns `category`, `parent_bucket`, `keyword`, `priority_rank`. The file (iii) `bands_classified.csv` is the application of both schemes to the 560-band development corpus, with columns `band_id`, `raw_genre_string`, `bucket_11`, `category_17`, `residual_flag`; `band_id` is a non-sequential, pseudonymized numerical identifier (see the Institutional Review Board Statement), `raw_genre_string` is the verbatim Encyclopaedia Metallum genre annotation, `bucket_11` and `category_17` are the assignments under the two schemes, and `residual_flag` is a binary indicator coded 1 when no keyword matched (the band falling in the residual “Other” category) and 0 otherwise. The file (iv) `classify.py` is a Python module (Python 3.14; Python Software Foundation; <https://www.python.org>) implementing the keyword-priority logic; given a raw genre

string and a keyword table, it returns the bucket and category assignments together with a residual flag for strings that fail to match any keyword. The file (v) *validation_subset.csv* contains the 60-band inter-coder validation subset described in Section 3.4, comprising the automatic and independent manual bucket assignments and an agreement flag. The file (vi) *pilot_norway_classified.csv* contains the out-of-sample portability demonstration described in Section 4, a 60-band Norwegian sample classified with the keyword tables unchanged. The files (vii) *bands_input_template.csv* and (viii) *bands_classified_template.csv* provide a minimal worked example of the expected input and output formats. The (ix) *README.md* file documents usage, the rationale for priority ordering, and procedures for adapting the keyword tables to other regional metal-scene corpora.

Table 1 reports the distribution of the 560-band development corpus across the eleven primary buckets (the keyword-priority basis on which these buckets are defined, including how related styles such as Black and Death are distinguished, is set out in Section 3.2). The largest buckets are Black ($n = 226$), Death ($n = 150$), and Thrash/Speed ($n = 60$); the smallest non-residual bucket is Industrial/Avant ($n = 1$). The residual “Other” bucket contains five bands (~0.9%), all of which are tagged with genre strings sufficiently idiosyncratic to fall outside the keyword-priority table; these are listed in *bands_classified.csv*, with the *residual_flag* column set to 1, for case-by-case inspection.

Table 1. Distribution of the 560-band Southern Chilean development corpus across the eleven primary buckets of the classification scheme. Source: *bands_classified.csv* (deposited in the dataset).

Primary Bucket	n (Bands)	% of Corpus
Black	226	40.4
Death	150	26.8
Doom/Sludge	29	5.2
Thrash/Speed	60	10.7
Heavy/Trad	51	9.1
Power	6	1.1
Folk/Pagan/Viking	2	0.4
Grindcore/Gore	22	3.9
Progressive	5	0.9
Metalcore/Deathcore	3	0.5
Industrial/Avant	1	0.2
Other (residual)	5	0.9

Under the seventeen-category refinement, the bucket distribution shifts as follows: the Black bucket decomposes into Atmospheric/Symphonic Black ($n = 13$) and Black ($n = 213$); Death decomposes into Brutal Death/Slam ($n = 8$), Technical Death ($n = 3$), and Death ($n = 139$); Doom/Sludge decomposes into Funeral Doom/Sludge ($n = 3$), Stoner ($n = 13$), and Doom ($n = 13$); and Thrash/Speed decomposes into Speed ($n = 7$) and Thrash ($n = 53$). The remaining seven buckets are preserved unchanged across granularities.

3. Methods

3.1. Source Data and Corpus Construction

The development corpus consists of the genre strings associated with 560 metal bands of Southern Chilean origin retrieved from Encyclopaedia Metallum on 28–29 April 2026. The geographic scope comprises the five southernmost administrative regions of continental Chile—La Araucanía, Los Ríos, Los Lagos, Aysén del General Carlos Ibáñez del Campo, and Magallanes y Antártica Chilena—covering roughly 1500 km of mainland coast, together with the Chiloé and Tierra del Fuego archipelagos. The region is characterized by a peripheral position relative to the metropolitan music industry concentrated in Santiago,

a multilingual cultural ecology in which Mapuche and Huilliche linguistic and musical traditions coexist with European-descent metal practice, and infrastructural constraints on inter-city circulation that make the analysis of brokerage patterns substantively informative. Heavy metal, as a regional musical practice, has been the focus of sustained scholarly attention in Latin America in recent years [9], with substantial cross-country variation in scene composition that makes documented, replicable classification schemes a precondition for productive comparative work. The Southern Chilean focus is motivated by the broader research program on music, territory, and identity in Southern Chile to which the present dataset contributes [6]. Because each band’s genre string is recorded as it appeared on the retrieval dates, the corpus is a synchronic snapshot: any analysis that treats it as representing a particular time window inherits this dating, an issue discussed further as a concept-drift limitation in the Section Known Limitations.

Inclusion criteria were applied at the retrieval stage. Bands were included if Encyclopaedia Metallum’s “Country” or “Location” field listed at least one city within the five-region perimeter, regardless of status (active, on hold, split-up, or unknown). Side-projects and one-person projects with at least one official release were retained. Excluded were entries flagged in the database as duplicates or as non-metal bands erroneously listed under metal-adjacent tags. Retrieval was performed via the public web interface using respectful scraping practices: an eight-second base delay with stochastic jitter between requests, a transparent User-Agent string identifying the research project and contact email, and full compliance with the site’s robots.txt directives.

The classification scheme released here is designed to be portable to other regional metal-scene corpora without modification of the core priority logic. Researchers applying the scheme to corpora with substantially different stylistic profiles may need to adjust priority orderings (see Section 3.2) or augment the keyword tables, but the deterministic, keyword-matching architecture itself transfers without change.

3.2. Eleven-Bucket Primary Scheme

Each band’s genre string is normalized by lower-casing and stripping diacritics, then matched against the keyword table (Table 2) in the listed priority order. The first matching keyword determines the band’s primary bucket; subsequent matches in the string are ignored. Bands not matching any keyword are assigned to a residual “Other” bucket.

Table 2. Keyword-priority table for the eleven-bucket scheme. Priority decreases from top to bottom: keywords for buckets listed earlier take precedence over keywords for buckets listed later when both match a band’s genre string.

Bucket	Primary Keywords (in Priority Order)
Grindcore/Gore	grindcore, gore, goregrind, mincore
Black	black metal, blackened, raw black, atmospheric black, symphonic black, ambient black
Death	death metal, brutal death, technical death, tech death, melodic death, slam
Doom/Sludge	doom metal, sludge metal, sludge, stoner, funeral doom, doom
Thrash/Speed	thrash metal, thrash, speed metal, crossover thrash
Heavy/Trad	heavy metal, traditional metal, NWOBHM
Power	power metal, symphonic power
Folk/Pagan/Viking	folk metal, pagan metal, pagan, viking metal, viking, celtic metal, celtic
Progressive	progressive metal, prog metal, avant-garde progressive
Metalcore/Deathcore	metalcore, deathcore, post-hardcore
Industrial/Avant	industrial metal, industrial, avant-garde metal, avant-garde, experimental metal, experimental, drone

Priority ordering is consequential. Grindcore/Gore is placed first because grindcore-tagged bands are often also death-metal-tagged; without priority on grindcore, a larger downstream bucket (Death) would absorb the grindcore subset. Similar reasoning applies to Folk/Pagan/Viking, which often co-occurs with Black; placing Folk after Black preserves the canonical “folk-tinged black metal” classification under Black, which is consistent with the genre’s reception history [10]. Researchers adapting the scheme to regional corpora with different stylistic profiles (e.g., scenes with heavy Folk–Pagan–Viking representation) may need to adjust priority ordering accordingly; the residual bucket distribution provides a diagnostic for such adjustments.

3.3. Seventeen-Category Refinement

For finer-grained analyses, four of the eleven buckets are decomposed into a total of ten sub-categories, while the remaining seven buckets are preserved unchanged. Decomposition is restricted to buckets that satisfy two conjoint conditions: substantial internal stylistic heterogeneity, in the sense of multiple sub-scenes with distinct stylistic, lyrical, and community boundaries; and a corpus n sufficient to support sub-category analysis under typical statistical assumptions ($n \geq 20$ in the development corpus). The four decomposed buckets and their internal divisions are as follows.

Black is decomposed into Atmospheric/Symphonic Black and Black. Atmospheric and symphonic variants of black metal constitute a stylistically distinct sub-scene whose orchestration choices and lyrical themes diverge substantially from raw and second-wave black metal, and whose audience and label networks are partially disjoint from those of the parent scene. The default “Black” sub-category absorbs raw, second-wave, and unmarked black metal.

Death is decomposed into Brutal Death/Slam, Technical Death, and Death. Brutal death and slamming death constitute a high-intensity sub-scene with distinct vocal techniques and rhythmic conventions. Technical death is a separate sub-style characterized by progressive-rock-derived structural complexity and instrumental virtuosity. The default “Death” sub-category absorbs traditional, melodic, and unmarked death metal.

Doom/Sludge is decomposed into Funeral Doom/Sludge, Stoner, and Doom. Funeral doom and sludge are heavy, slow, and often dissonant sub-styles with distinct historical roots—funeral doom in 1990s European death-doom, and sludge in early-1990s American hardcore-doom crossovers. Stoner is a 1960s-rock-influenced sub-style with its own audience and label network [11]. The default “Doom” sub-category absorbs traditional and unmarked doom metal.

Thrash/Speed is decomposed into Speed and Thrash. Speed metal is the older umbrella term associated with the late 1970s and early 1980s, while thrash is the more specific later refinement that crystallized in the mid-1980s American underground. In current usage, thrash is the larger and more actively recorded sub-scene, hence its assignment as default within the bucket.

The remaining seven buckets are not decomposed. For Folk/Pagan/Viking, Power, Progressive, Metalcore/Deathcore, and Industrial/Avant, the development corpus n is below the threshold for productive sub-category analysis ($n \leq 6$ each in this corpus). For Grindcore/Gore ($n = 22$), although in principle decomposable into grindcore, goregrind, and mincore, Encyclopaedia Metallum’s tagging conventions for these distinctions are inconsistent enough that automated keyword matching would produce more noise than signal; researchers requiring such distinctions should re-code manually. For Heavy/Trad ($n = 51$), the bucket is internally homogeneous in its Encyclopaedia Metallum tagging, with no widely recognized sub-scene boundaries that the database consistently encodes.

Decomposition is implemented by augmenting the keyword table with secondary keywords that operate only within the parent bucket. For example, within the Death bucket, the secondary keyword “brutal death” (priority 1) identifies the Brutal Death/Slam sub-category, the secondary keyword “technical death” (priority 2) identifies the Technical Death sub-category, and the default “death metal” (priority 3) absorbs the remaining bands into the Death sub-category. The full keyword table for the seventeen-category scheme is provided in `keywords_17category.csv`.

3.4. Validation

Validation of the classification scheme proceeded along three axes: deterministic consistency, residual coverage, and agreement with independent human coding. First, deterministic consistency: The keyword-priority logic produces identical bucket assignments across runs given the same source string and keyword table, by construction. The priority decisions documented in Sections 3.2 and 3.3 are themselves the auditable component of the procedure and can be reviewed and modified by users adapting the scheme to other corpora.

Second, residual coverage: The residual “Other” bucket contained 5 bands (0.9% of the development corpus) under the eleven-bucket scheme. Inspection of the five residual bands shows that they carry idiosyncratic genre tags—“Southern/Groove Metal”, “Progressive Rock/Metal”, “Groove Metal” (two bands), and “Death ‘n’ Roll”—that fall outside the priority table by design rather than by oversight. Absorbing “groove” into a Thrash- or Heavy-adjacent bucket would expand the priority table into idiosyncratic local territory whose cross-scene comparability is uncertain; absorbing “progressive rock” into Progressive would compromise the bucket’s specificity to metal styles. Residual rates below approximately 2% in corpus studies generally indicate adequate fit between a categorical scheme and the underlying tagging conventions of the source database. Researchers adapting the scheme to new regional corpora should monitor this indicator and intervene only when the residual rate exceeds that range or when residuals cluster on a particular local style not represented in the development corpus.

Third, agreement with independent human coding: To assess whether the scheme’s automatic assignments correspond to expert human judgement rather than merely being internally consistent, a stratified random subset of 60 bands was drawn from the development corpus and independently re-coded to the eleven-bucket level by a second coder with domain expertise in metal genres. The subset was stratified to guarantee coverage of every bucket, including the small ones, while preserving the approximate prevalence of the common categories; the coder worked from the raw genre strings alone, blind to the automatic output. Agreement between the automatic classification and the human coder was 81.7% (49 of 60 bands). Cohen’s κ was 0.78 (95% bootstrap confidence interval [0.66, 0.88], 2000 resamples), indicating substantial agreement under conventional benchmarks. The coded subset is released as `validation_subset.csv`, so that the statistic is fully reproducible from the deposited files.

The eleven disagreements are themselves informative, since they do not scatter randomly but concentrate on the hybrid boundaries that the priority order is designed to resolve. Three fall on the Black/Death boundary (for example, strings of the form “Death/Black Metal”, which the scheme assigns to Black because Black outranks Death, whereas the coder read Death as primary); three fall on the Grindcore/Death boundary (for example, “Brutal Death Metal/Goregrind”, assigned to Grindcore/Gore by its top priority rank but read as Death by the coder); three involve groove- or roll-suffixed strings (“Groove Metal” and “Death ‘n’ Roll”) that the scheme leaves residual or routes by a secondary keyword; and the remaining two involve adjectival modifiers at the Black/Speed and

Heavy/Speed margins. In every case, the divergence from human judgement follows directly from an explicit, documented priority decision (Section 3.2) rather than from a classification error. The validation therefore confirms not only that the scheme agrees substantially with expert coding, but that its residual disagreements are concentrated where the deterministic rule deliberately overrides the ambiguity of multi-label tags—precisely the cases a user adapting the scheme would re-examine first.

All three validation axes are reproducible from the deposited files: deterministic consistency and residual coverage from `bands_classified.csv`, together with the keyword tables and `classify.py`; and the inter-coder statistic from `validation_subset.csv`, which records both the automatic and the independent manual assignment for each of the 60 bands.

4. User Notes

The classification scheme is intended as a starting point for four classes of analysis.

Cross-scene comparative work. Researchers conducting parallel analyses of other regional or national metal scenes [12]—Brazilian, Mexican, Argentinian, Polish, Greek, Indonesian, and others—can apply the same keyword-priority logic to local Encyclopaedia Metallum corpora to produce bucket distributions directly comparable to the Southern Chilean reference distribution reported here. Such comparability is particularly relevant for territorialized analyses of music scenes [6], where exogenous factors—regional histories, local language ecologies, and infrastructural conditions of music production and circulation—are hypothesized to shape the stylistic composition of scenes. Holding the classification scheme constant across scenes allows differences in bucket composition to become substantively interpretable, rather than confounded with differences in bucketting methodology.

Brokerage typology and cross-genre mobility studies. The eleven-bucket scheme provides the categorical alphabet over which musician-level multi-genre indicators are defined [13]. Per-musician counts of distinct buckets across participating bands enable the construction of brokerage typologies that distinguish aesthetic from geographic forms of brokerage, identify musicians who specialize locally versus those who circulate across territorial and stylistic boundaries, and support investigations of how musical mobility relates to scene-level structural properties. The present encoding scheme is the necessary first step for any such operationalization.

Comparative adoption workflow. A typical adoption workflow for an investigator working on a different regional scene proceeds in four steps. First, retrieve the local corpus from Encyclopaedia Metallum using the same inclusion criteria documented in Section 3.1. Second, run `classify.py` with the default keyword tables and inspect the printed bucket distribution alongside the residual rate. Third, compare the local distribution with the Southern Chilean reference reported in Table 1: substantive differences in bucket composition are informative when interpreted territorially. Fourth, if the local residual rate is elevated, inspect the bands in the residual bucket; where local styles concentrate in a recognizable category not covered by the default priority table, the keyword tables can be augmented with the relevant local keyword at an appropriate priority rank, and the adapted tables archived alongside any reported results to enable cross-scene replicability. To illustrate this workflow on real out-of-sample data, the scheme was applied, with the keyword tables left entirely unchanged, to a random sample of 60 Norwegian bands retrieved under the same inclusion criteria (deposited as `pilot_norway_classified.csv`). The scheme's core logic transferred: the distribution was dominated by Black (46.7%), consistent with the documented profile of the Norwegian scene, and most hybrid strings resolved through the priority order. The residual rate rose to 5.0% (three of 60), and the residuals were diagnostic rather than random—two of the three carried primary styles

absent from the Southern Chilean development corpus (Symphonic Metal and Gothic Metal) and so were not represented as primary buckets, indicating precisely where a Norwegian adaptation would augment the tables, while the third was a Progressive/post-metal border case. The pilot also surfaced a normalization edge case for multi-era genre strings of the form "... (early); ... (later)", which the classifier treats as co-present tags; investigators applying the scheme to corpora that use such temporal annotations should pre-process them according to the time window of interest, consistent with the concept-drift limitation noted in Section 3.1 and Section Known Limitations. One priority decision warrants explicit attention by adopters. Because Black is ranked above Folk/Pagan/Viking, a band tagged primarily as folk metal with a secondary black-metal affiliation is assigned to Black, just as a folk-tinged black-metal band is; the ordering preserves the canonical reading of folk-tinged black metal but does so at the cost of folk-metal precision. Researchers for whom folk-metal precision is analytically central—for instance, in scenes with a large pagan-folk sub-scene—should demote Black below Folk/Pagan/Viking in their adapted priority table, and document the change.

Sensitivity analyses. The seventeen-category refinement supports robustness checks: any substantive finding under the eleven-bucket scheme can be re-tested under the seventeen-category scheme without re-coding the source data. Where findings diverge between the two granularities, the divergence becomes diagnostic of where the bucketing choice does substantive work.

Known Limitations

Three limitations bear acknowledgement. First, the keyword-priority logic imposes a hierarchical assignment on what is, in practice, a multi-membership classification; bands whose genre strings span buckets of comparable salience are assigned to the bucket whose keyword appears first in priority order. Second, the development corpus is a single-region (Southern Chile) sample; the priority table may need adjustment for scenes with different sub-genre profiles (e.g., scenes with heavy Folk–Pagan–Viking representation may require demotion of Black in priority order). Third, the eleven-bucket scheme inherits any biases present in Encyclopaedia Metallum's curatorial coverage—the database underrepresents bands without official releases and DIY tape-and-digital projects [14], an issue analogous to coverage gaps discussed in popular-music corpus studies more generally.

A fourth limitation concerns temporal coverage. Encyclopaedia Metallum was founded in 2002, and although the database includes substantial retrospective documentation, bands active prior to 2002 depend on fan-driven historical reconstruction, whose completeness varies by region and sub-style; pre-2002 bands in peripheral scenes are particularly likely to be under-documented. Researchers conducting longitudinal analyses of scene evolution should therefore consider temporally stratified sub-samples rather than treating the full corpus as a uniform observational window. A fifth and related limitation concerns concept drift: a band's classified bucket reflects the genre string visible in Encyclopaedia Metallum at the time of retrieval, which in turn reflects curatorial decisions accumulated over the band's documented lifetime. When a band's substantive style has shifted over time but the database tag has not been updated to reflect the shift, the classification will report the historical rather than the current style. This is generally a minor concern for synchronic analyses, but it is not negligible for them: because the corpus is a synchronic snapshot fixed at the retrieval date (Section 3.1), any synchronic analysis that treats the data as representing a particular time window inherits the retrieval-date constraint, and conclusions are properly read as describing the scene as documented on 28–29 April 2026 rather than at an arbitrary present. The concern is most acute for studies tracking stylistic evolution within scenes, where outdated tags directly bias the trajectory.

Funding: This research was funded by ANID—FONDECYT Regular Project No. 1241458 and by the ANID Millennium Science Initiative Program—Núcleo Milenio en Culturas Musicales y Sonoras (CMUS), grant number NCS 2025-65.

Institutional Review Board Statement: The development of the classification scheme described in this data descriptor was conducted within the framework of the ANID—FONDECYT Regular Project “Representaciones territoriales emergentes en la música popular del sur de Chile” (Project No. 1241458), which received ethical clearance from the Comité Ético Científico of Universidad de Los Lagos (CEC-ULagos; internal project code H008/2024, approval Oficio Ordinario No. 025/2024, dated 23 April 2024). CEC-ULagos is accredited by the SEREMI de Salud de la Región de Los Lagos in accordance with Chilean Law No. 20.120 on scientific research involving human subjects. This research conforms to the principles of the WMA Declaration of Helsinki. The deposited dataset consists of band-level metadata retrieved from the publicly available, fan-curated Encyclopaedia Metallum database, where artists list their bands on a voluntary basis. Band identifiers in the deposited files are pseudonymized through replacement with non-sequential numerical IDs; raw genre strings are preserved to enable replication and reuse, which entails a residual risk of re-identification for bands with sufficiently idiosyncratic genre tags. No personal data beyond what is publicly visible on Encyclopaedia Metallum is processed or released.

Data Availability Statement: All files described in this data descriptor are openly available at Zenodo, under a CC-BY 4.0 license (<https://doi.org/10.5281/zenodo.20222458>). The deposited dataset comprises nine files: keywords_11bucket.csv (eleven-bucket keyword-priority table); keywords_17category.csv (seventeen-category keyword-priority table); classify.py (Python implementation of the keyword-priority logic); bands_classified.csv (classified development corpus, 560 bands); validation_subset.csv (60-band inter-coder validation subset); pilot_norway_classified.csv (60-band Norwegian out-of-sample portability demonstration); bands_input_template.csv and bands_classified_template.csv (worked example of the input and output formats); and README.md (usage documentation and rationale for priority ordering). The file list and counts here are consistent with those given in Section 2.

Acknowledgments: The author thanks the Encyclopaedia Metallum for maintaining the underlying database.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Pino, M.C.; Giancola, M.; D’Amico, S. The association between music and language in children: A state-of-the-art review. *Children* **2023**, *10*, 801. [CrossRef] [PubMed]
2. Nguyen, T.; Flaten, E.; Trainor, L.J.; Novembre, G. Early social communication through music: State of the art and future perspectives. *Dev. Cogn. Neurosci.* **2023**, *63*, 101279. [CrossRef] [PubMed]
3. Encyclopaedia Metallum. The Metal Archives. Available online: <https://www.metal-archives.com> (accessed on 29 April 2026).
4. Fricconet, G. Establishing a taxonomy of metal subgenres based on quantitative musical and thematic features. *Met. Music. Stud.* **2022**, *8*, 183–203. [CrossRef]
5. Czedik-Eysenberg, I.; Wieczorek, O.; Flexer, A.; Reuter, C. Charting the universe of metal music lyrics and analyzing their relation to perceived audio hardness. *Trans. Int. Soc. Music Inf. Retr.* **2024**, *7*, 129–143. [CrossRef]
6. Soto-Silva, I.; Zamora, K.; Silva-Zurita, J. Espacialidad en la experiencia musical. Reflexiones indisciplinares y una propuesta a partir del territorio. *El Oído Pensante* **2024**, *12*, 85–107. [CrossRef]
7. Holt, F. *Genre in Popular Music*; The University of Chicago Press: Chicago, IL, USA, 2007.
8. González, J.P. *Pensar la Música Desde América Latina: Problemas e Interrogantes*; Ediciones Universidad Alberto Hurtado: Santiago, Chile, 2013.
9. Varas-Díaz, N.; Nevárez Araújo, D.; Rivera-Segarra, E. (Eds.) *Heavy Metal Music in Latin America: Perspectives from the Distorted South*; Lexington Books: Lanham, MD, USA, 2020.
10. Kahn-Harris, K. *Extreme Metal: Music and Culture on the Edge*; Berg: Oxford, UK, 2007.
11. Weinstein, D. *Heavy Metal: The Music and Its Culture*, rev. ed.; Da Capo Press: New York, NY, USA, 2000.
12. Bennett, A.; Peterson, R.A. (Eds.) *Music Scenes: Local, Translocal, and Virtual*; Vanderbilt University Press: Nashville, TN, USA, 2004.

13. Crossley, N. *Networks of Sound, Style and Subversion: The Punk and Post-Punk Worlds of Manchester, London, Liverpool and Sheffield, 1975–80*; Manchester University Press: Manchester, UK, 2015.
14. Shea, N.J. A demographic sampling model and database for addressing racial, ethnic, and gender bias in popular-music empirical research. *Empir. Musicol. Rev.* **2023**, *17*, 49–58. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.