DATASHEET

XIM Audio catalogue: licensed music production data for AI training

XIM Audio Inc., Canada. Catalogue measured on 2026-11-13. Document version 2026-11-13. DOI 10.5281/zenodo.23883162. Contact mikas@ximaudio.com, catalogue at ximaudio.com/data.

This datasheet follows the questions of Gebru et al., *Datasheets for Datasets* (2021), for the whole XIM Audio catalogue: 11 sets of modern music made in the DAW, pop to trance, as working producers made and sold it. The numbers are re-measured on the live catalogue every night; this document is the state on 2026-11-13 and is versioned on Zenodo when it changes. Audio is not attached: the catalogue is licensed, per set, on request.

4,408	9,463	3,637	289,034
songs as stems, 117,622 lanes	DAW sessions, 11 DAWs	paired sets: session + stems + masters	MIDI files, 378 filmed tutorial sessions

1. Motivation

For what purpose was the dataset created?
To license music production data for training, fine-tuning and evaluating audio and music models under terms the creators agreed to and are paid on. Every open multitrack set of real songs is non-commercial or has no licence (see the licence table); this catalogue exists so that a product or a lab can train on multitrack music legally, at a scale of thousands of songs.

Who created it?
The music was made by working producers who sell on a music production marketplace online since 2012. XIM Audio Inc. (Canada) licenses that catalogue as training data under the licence the marketplace grants it, and measures, documents and delivers it.

Who funded it?
Self-funded by XIM Audio Inc. and its sister marketplace at the time of this version.

2. Composition

Eleven sets. One item is one song, one session, one episode, one file or one table row, as the set says. The paired sets are the songs of the stems set that also ship their session and masters, so their hours are not added to the total above.

SET	ITEMS	FILES	HOURS OF AUDIO	DAW SPLIT
Stems	4,408	117,622 files	about 9,669	Logic 2,258, Ableton 1,485, Fl studio 117, Cubase 14, FL Studio 6
Paired production sets	3,637	137,485 files	about 7,790	Logic 2,147, Ableton 1,373, Fl studio 117
Vocal and instrumental pairs	2,006	4,012 files	about 222	
DAW sessions	9,463	n/a	about 691	Ableton 4,340, Logic 3,110, FL Studio 1,531, Cubase 365, Studio One 50, GarageBand 42, Bitwig 15, Reason 4, OMF / AAF 3, Maschine 2, Reaper 1
Tutorial sessions	378	n/a	about 224	
MIDI files	1,633	289,034 MIDI files		
Synth presets	445	167,387 preset files		
Loops and one-shots	1,695	458,371 audio files		
Instrumentals	3,351	n/a	about 184	
Stock music	1,927	n/a	about 81	
Marketplace data	54,446			

Stems hours: About 9,669 h: every lane's length added up, read from the file headers of 4,399 songs and packs on 2026-11-13, plus 9 not yet measured at the average. Sessions hours: About 691 h: the demo of 7,396 of the 7,872 sessions on the public store runs 4:23 on average, measured 2026-10-17, applied to all 9,463 project files inside them. Hours of finished songs; the lanes inside a session multiply it.

What one item holds

Stems

One folder per song, one 24-bit WAV per lane as the producer printed it from the session, the unmastered mix as the unity-gain sum of the lanes, MIDI where the session had it, and one manifest row per song with one instrument label per lane.

Paired sets

The project file zipped, every lane printed from it, the MIDI folder where the session had notes, the flat master and the mastered master where it exists (1,500 of 3,637), and the manifest row. 3,534 songs ship the session, the stems and the MIDI of the same track.

Acapella and instrumental pairs

Two sample-aligned files per song, the vocal and the rest of the song, printed from the same session as two buses, so the two files share one clock and one gain.

DAW sessions

The project file and its media as sold, parsed into one JSON schema (tracks, plugin chains, volume, pan, sends, buses, sidechain, clips, notes, automation, sections), and the manifest row. Finished songs run 3:52 at the median (p10 1:59, p90 6:55).

Tutorial sessions

One folder per episode with the video, the word-timed transcript, the finished project, its stems and its masters: 378 episodes recorded 2019 to 2026.

MIDI, presets, loops and one-shots, instrumentals, stock music

Files as sold, with a manifest row per file: BPM, key, instrument, synthesizer or category where known.

Marketplace data

De-identified tables of the marketplace: products, categories, orders by month and country, amounts as sold, reviews, with a data dictionary.

Labels and metadata

Every lane of the stems set carries one of 53 instrument labels: Counted 30 October 2026 over 112,088 lane files in 3,867 stems packs on the store, printed songs and producer packs together: 71% labelled from the producer's own lane name, 26% by an audio classifier, 4% unlabelled. All 53 labels are listed. Lane counts are from the 3,366 songs printed from sessions. Songs printed from sessions hold 24 lanes at the median (p10 12, p90 45, mean 29.7). The 24 most frequent labels:

LABEL	LANES	LABEL	LANES	LABEL	LANES
Bass	10,397	Lead	8,783	Kick	7,161
Snare	6,062	Vocal	5,641	FX	4,237
Clap	4,139	Hi-hat	4,139	Pad	3,806
Crash	3,169	Percussion	3,031	Drum bus	2,911
Hi-hat, open	2,897	Riser	2,797	Rim	2,784
Sweep	2,617	Synth	2,369	Strings	2,357
Reverse	2,297	Piano	2,249	Sub	1,977
Tom	1,590	Shaker	1,583	Pluck	1,401

Share of items with BPM, key and genre present, per set:

SET	ITEMS	BPM PRESENT	KEY PRESENT	GENRE PRESENT
Stems	4,408	100%	94%	100%
Paired production sets	3,637	100%	95%	100%
Vocal and instrumental pairs	2,006	90%	100%	100%
DAW sessions	9,463	99%	95%	98%
Tutorial sessions	378	100%	100%	97%
MIDI files	1,633	90%	84%	100%
Loops and one-shots	1,695	89%	87%	100%
Instrumentals	3,351	99%	100%	100%
Stock music	1,927	99%	90%	100%

Stems set spread: BPM 60 to 200, median 125; 2,177 songs in minor keys and 1,973 in major; sample rates 44100 Hz 3,498, 48000 Hz 755, 96000 Hz 146. Largest genre tags: Pop - Rock 937, Deep House - Garage 563, House Music 513, Electro House - Big Room 331, Progressive House 270, Trance 241, Ambient - Chill 167, Breakbeat - Beats 144, Techno (Peak Time Driving) 144, Hip Hop - Urban 142, Tech House 114, Melodic Techno 110.

Does it contain all possible instances or a sample?

All items live on the store on the measured date. Items leave the catalogue when a creator withdraws them; the nightly measurement reflects it.

Is there a recommended split?

Yes. Each delivery carries a creator-level train/validation/test split that holds across deliveries, so no creator sits on both sides.

Errors, noise, redundancies?
Lanes are as printed; silent or near-silent lanes exist where the producer printed them. Duplicates are removed by content hash per delivery. 4% of lanes are unlabelled.

Is it self-contained?
Yes. Nothing links to external resources; nothing is streamed or scraped.

Confidential data, personal data, offensive content?
No song titles, artist names, producer names or lyrics. Creators and songs appear as pseudonymous IDs. Vocals are the sound of a recorded performance, not a named person. The music is what was sold on a public marketplace; nothing was selected for or against subject matter.

---

3. Collection process

How was the data acquired?
Stems were printed by the producer, lane by lane, from the DAW project the song was made in, with its plugins loaded. Sessions are the producer's own project files as sold. Tutorials were recorded on camera by the presenter. No separation model and no generative model touched any file; nothing was crawled.

Who was involved, and how were they compensated?
Producers who sell on the marketplace. Each keeps ownership and is paid a share of every licence XIM Audio grants on their work.

Over what timeframe?
Files as sold on the marketplace since 2012; the tutorial episodes since 2019. The catalogue is measured on the date in the header.

Was consent obtained?
Every creator accepted terms that name "developing, training and improving software, audio tools and other technology, including machine learning". Where a written consent is still being collected for part of a set, the catalogue page says so and the quote says when it clears. A creator can withdraw: the licensee stops using a withdrawn item in new training within 30 days; trained models are not affected.

---

4. Preprocessing, cleaning, labelling

What was done to the raw data?
Nothing to the audio beyond printing it from the session. De-duplication by content hash per delivery. BPM, key, genre, DAW, sample rate and duration are read from the listing and from the file headers; the share present is in the tables above. Instrument labels: from the producer's own lane name where it was clear, else an audio classifier, else unlabelled. Sessions are parsed into JSON; the original project file ships beside it.

Is the raw data available?
Yes, every delivery is the files as sold plus the manifest.

---

5. Uses

What is it used for?
Source separation (stems are the ground truth, no separation model in the loop); mixing, mastering and gain-staging models (lanes plus flat and mastered masters); music generation that works in parts, lane by lane, rather than on finished mixes; models that read, write or act on DAW projects (sessions in one JSON schema); transcription and MIDI models (notes next to the printed audio); vocal and accompaniment models (the pairs set); and benchmark construction under a commercial licence.

What should it not be used for?

Reproducing or imitating a specific creator's recordings. Trying to re-identify creators or songs behind the pseudonymous IDs. Redistributing the files or any part of them.

Anything a user should know?

Compositions and lyrics are not licensed; the sound of the recording is. Models and outputs belong to the licensee. The catalogue is modern music made in the DAW; acoustic and classical recordings are a small share.

---

6. Distribution

How is it distributed?

Private S3 prefix or the licensee's own bucket, one scoped key per deal, access logs, with a manifest, checksums, the split and this datasheet. Never streamed, never public. An evaluation set of a few songs and the full index come first.

Under what licence?

Per set, non-exclusive, named use and term; three years from delivery unless the order form says otherwise. Research licences for universities and labs, commercial licences for products. The standard terms are public at ximaudio.com/rights; requests at ximaudio.com/request.

Public cards

Each set has a public card with the same numbers on Hugging Face (huggingface.co/ximaudio) and on ximaudio.com/data.

---

7. Maintenance

Who maintains it?

XIM Audio Inc. Contact mikas@ximaudio.com.

How often is it updated?

The catalogue is re-measured every night on the live store; the site, the cards and this datasheet are generated from the same numbers. A new version of this document is published on Zenodo when the figures move materially; older versions stay citable.

Can others contribute?

Producers join through the marketplace; their work enters the catalogue once they accept the terms.

Citation

```
@misc{ximaudio_catalogue_2026,
  title  = {XIM Audio catalogue: licensed music production data for AI training (datasheet)},
  author = {XIM Audio Inc.},
  year   = {2026},
  doi    = {10.5281/zenodo.23883162},
  note   = {Catalogue measured 2026-11-13; audio licensed on request}
}
```

Reference: Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., Crawford, K. Datasheets for Datasets. Communications of the ACM 64(12), 2021.