

Data Sourcing & Collection

Building a Podcast Corpus

David Jurgens and Dallas Card
July 28, 2026



Podcasts vs. Other Sources of Social Data

- Social media/Reddit: centralized, (historically) API-accessible, arrives pre-structured with built-in engagement metrics (likes, upvotes, reply trees)
- Podcasts, by comparison:
 - more one-directional (producer to consumer)
 - limited direct engagement metrics
 - require significant pre-processing (transcription, etc.)
- Unique affordances of podcasts
 - decentralized distribution (RSS)
 - long-running, long-form conversation
 - multimodal data
- Net takeaway: richer, identifiable discourse, with higher processing cost and less visible metrics

Roles in the Podcast Distribution Ecosystem

- **Hosting providers** (e.g., Libsyn, Buzzsprout, Acast, Megaphone, Simplecast):
 - where the audio actually lives; generate the RSS feed on the creator's behalf; often also sell dynamic ad insertion
- **Directories/listening apps** (Apple Podcasts, Spotify, Overcast, Pocket Casts):
 - consumer-facing; read RSS feeds and present them to listeners; generally don't store audio themselves
- **Aggregator/index services** (PodcastIndex, Podchaser):
 - crawl many feeds to build searchable databases; not for listening, but for discovery/analytics — this is the layer researchers typically tap into
- Some distributors blur these roles (e.g., Spotify is simultaneously a directory, a host for some shows, and a gatekeeper for exclusive content)

How Podcast Distribution Works

- Podcasts are not hosted by a single platform — they're syndicated via RSS
- Creator → hosting provider → RSS feed → directories/apps (Apple, Spotify, PodcastIndex, etc.)
- If you can locate the RSS feed (as an XML file), that provides an index into individual mp3 files and metadata for each episode

Introducing PodcastIndex

- Open-source, free-to-use podcast index (built as a community/decentralized alternative to closed platform indexes)
- Why we use it as a starting point: open API, bulk data access, no platform gatekeeping
- Scale: more than 4.7 million feeds (some now offline)
- Downsides: not fully comprehensive; does not provide access to paywalled shows
- Still one of the best starting points to find RSS feeds

Filters▼

Supporting Apps, Directories and Hosting Companies

App



TrueFans

app, website, podcast player, rss proxy, video, hosting



Up Next

podcast player, app



Podcast Guru

app, podcast player

Supported Elements

Alternate Enclosure, Block, Boostagrams, Chapters, Chat, Content Link, Episode, Funding, GUID, Image, License, Live Boosts, Live Item, Location, Locked, Medium, OP3, Person, Podping, Podroll, Publisher, Remote Item, Sat Streaming, Search, Season, Social Interact, Soundbite, Trailer, Transcript, Txt, Update Frequency, Value, Video, Wallet Switching (VTS)

Alternate Enclosure, Block, Chapters, Chat, Content Link, Episode, Funding, GUID, Image, License, Live Item, Location, Locked, Medium, Person, Podping, Podroll, Publisher, Remote Item, Season, Social Interact, Soundbite, Trailer, Transcript, Txt, Update Frequency, Value, Wallet Switching (VTS)

Alternate Enclosure, Boostagrams, Chapters, Funding, GUID, Live Boosts, Live Item, Location, Person, Podping, Podroll, Remote Item, Sat Streaming, Search, Season, Transcript, Value, Video, Wallet Switching (VTS)

Platforms

Android, Fediverse, Linux, Mobile, Web, Windows, iOS, macOS

iOS

Android, Mobile, iOS, macOS

<https://podcastindex.org/>



Platform Solutions Resources Open Source Enterprise Pricing

Search or jump to... / Sign in Sign up

Podcastindex-org / database Public

Notifications Fork 8 Star 30

<> Code Issues 18 Pull requests Actions Projects Security and quality Insights

master 1 Branch 0 Tags Go to file Code

 **daveajones** Database schema for MySQL c10aaa7 · 3 years ago 9 Commits

README.md	mit license	4 years ago
create_table_statement.sql	example requested - closes #4	6 years ago
db_newsfeeds_1000rows.csv	Sample of newsfeeds table	6 years ago
pcapi_schema.sql	Database schema for MySQL	3 years ago

README

Database

About

No description, website, or topics provided.

Readme

Activity

Custom properties

30 stars

3 watching

8 forks

Report repository

Releases

No releases published

https://github.com/podcastindex-org/database

Anatomy of an RSS Feed

- **Feed-level metadata:** title, author, description, category tags, artwork, owner/contact info
- **Episode-level metadata:** title, description, publish date, duration, **enclosure** (audio file URL), episode/season numbers, GUID
- **Optional / inconsistent / commonly misused:**
 - GUID — meant to be a stable unique identifier, but sometimes reused or changed across re-uploads
 - Episode/season numbers — frequently absent, reset, or reused
 - `itunes:episodeType` (full/trailer/bonus) — inconsistently set, so trailers/bonus content often can't be reliably filtered out
 - Newer Podcasting 2.0 namespace tags (`podcast:transcript`, `podcast:person`, `podcast:chapters`) — rich when present, but only a minority of feeds include them

Feed XML — Podcast metadata

```
<channel>
  <title>Data Stories</title>
  <link>https://datastori.es</link>
  <description><![CDATA[A podcast on data and how it affects our lives – with Enrico Bertini and Moritz Stefaner]]></description>
  <lastBuildDate>Thu, 12 Dec 2024 16:55:36 +0000</lastBuildDate>
<image><url>https://datastori.es/wp-content/uploads/2017/12/ds-1400-facelift.png</url><title>Data Stories</title><link>https://datastori.es</link></image>
<atom:contributor><atom:name>Enrico Bertini</atom:name></atom:contributor>
<atom:contributor><atom:name>Moritz Stefaner</atom:name></atom:contributor>
<podcast:person img="https://datastori.es/wp-content/uploads/2019/08/enrico-150x150.jpg">Enrico Bertini</podcast:person>
<podcast:person img="https://www.gravatar.com/avatar/4eb788c4efbab6f3088e6ae16a85f09f.jpg?s=512&amp;d=mm&amp;r=g">Moritz Stefaner</podcast:person>
<generator>Podlove Podcast Publisher v4.5.3</generator>
  <copyright>© 2026 Enrico Bertini and Moritz Stefaner</copyright>
  <itunes:author>Enrico Bertini and Moritz Stefaner</itunes:author>
  <itunes:type>episodic</itunes:type>
  <itunes:summary><![CDATA[Enrico Bertini and Moritz Stefaner discuss the latest developments in data analytics, visualization and related topics.]]></itunes:summary>
<itunes:category text="Arts"><itunes:category text="Visual Arts"></itunes:category></itunes:category>
<itunes:category text="Technology" />
<itunes:category text="Education" />
  <itunes:owner>
    <itunes:name>Enrico Bertini and Moritz Stefaner</itunes:name>
    <itunes:email>mail@datastori.es</itunes:email>
  </itunes:owner>
  <itunes:image href="https://datastori.es/wp-content/uploads/2017/12/ds-1400-facelift.png"/>
  <itunes:subtitle>A podcast on data and how it affects our lives – with Enrico Bertini and Moritz Stefaner</itunes:subtitle>
  <itunes:block>no</itunes:block>
  <itunes:explicit>>false</itunes:explicit>
```

Feed XML — Episode metadata

```
<item>
  <title>169 | Data Conversations with Vidya Setlur</title>
  <link>https://datastori.es/169-data-conversations-with-vidya-setlur/</link>
  <pubDate>Mon, 12 Dec 2022 14:07:01 +0000</pubDate>
  <guid isPermaLink="false">podlove-2022-12-02t16:23:24+00:00-d17d2a416b437a4</guid>
  <description><![CDATA[169 | Data Conversations with Vidya Setlur]]></description><post-id xmlns="com-wordpress:feed-additions:1">4790</post-id>
  <atom:link rel="http://podlove.org/deep-link" href="https://datastori.es/169-data-conversations-with-vidya-setlur/#" />
</item>

<enclosure url="https://datastori.es/podlove/file/6988/s/feed/c/mp3/169-data-conversations-with-vidya-setlur.mp3"
length="65865461" type="audio/mpeg"/>

  <itunes:duration>01:08:26</itunes:duration>
  <itunes:author>Enrico Bertini and Moritz Stefaner</itunes:author>
  <itunes:subtitle></itunes:subtitle>
  <itunes:title>Data Conversations with Vidya Setlur</itunes:title>
  <itunes:episode>169</itunes:episode>
  <itunes:episodeType>full</itunes:episodeType>

  <itunes:image href="https://datastori.es/wp-content/uploads/2022/12/ds169-vidya.jpg"/>
  <content:encoded><![CDATA[<p></p>
<p><span style="font-weight: 400;">We have <a href="https://research.tableau.com/user/vidya-setlur">Vidya Setlur</a> on
the show to talk about the role language, and natural language processing (NLP) play in data visualization and
analytics.</span></p>]]></content:encoded>
```

Feed-Level Issues

- Dead feeds (podcast discontinued, host stopped paying, URL 404s)
- Feed moves (creator switches hosting provider → new URL; old one may or may not redirect)
- Feeds are dynamic! New episodes are added, and old ones may be removed
- Duplicate feeds (same show indexed under multiple URLs/IDs)
- Private/dynamic feeds (paywalled subscriber content — e.g., Patreon-gated feeds — not publicly crawlable)
- Practical implication: any static "snapshot" of a feed list decays quickly

Episode-Level Issues

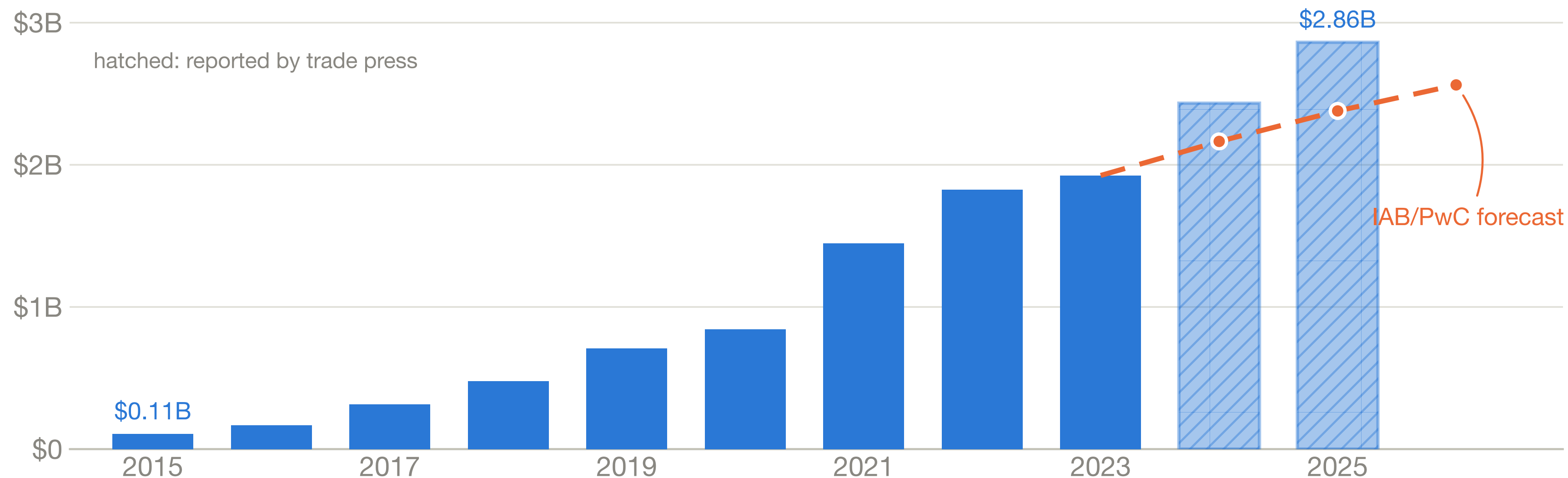
- Missing or broken enclosures (audio file no longer resolves)
- Inconsistent metadata (episode numbers reused, missing dates, malformed titles)
- Retroactive edits (episodes re-uploaded, re-titled, or removed after publication)
- Trailers/bonus content mixed in with regular episodes
- No guaranteed way to distinguish "main feed" content from ads/promos at the metadata level
- Not a true static archive of the past

Podcast Categories

- Most podcasts use a hierarchical category ontology (19 top-level categories) introduced by Apple
 - e.g., Arts, Business, Comedy, Education, etc.
- Many have secondary categories
 - e.g., Astronomy, Chemistry, Earth Science, etc. under "Science"
- Categories are self-selected by creators at submission time, entered as a free text field — not independently verified
 - Notable absences: no standalone top-level "Politics" category (likely folded into "News" or "Society & Culture" — verify current taxonomy before presenting)
 - Useful as a coarse stratification variable, not a precise content label
- <https://podcasters.apple.com/support/1691-apple-podcasts-categories>

It Pays to Podcast

- US podcast ad revenue:
\$105.7M (2015) → \$1.93B (2023) → \$2.86B (2025)
- The industry's own May-2024 forecast said \$2.56B for 2026



Interactive Advertising Bureau and PricewaterhouseCoopers LLP. "U.S. Podcast Advertising Revenue Study: 2023 Revenue & 2024–2026 Growth Projections." May 2024 <https://www.iab.com/insights/us-podcast-advertising-revenue-study-2024/>

IAB, April 2025, via *Barrett Media*: 2024 <https://barrettmedia.com/2025/04/17/podcast-advertising-revenue-grew-more-than-25-in-2024-new-iab-data-shows/>

IAB, April 2026, via *Radio Ink*: 2025. <https://radioink.com/2026/04/16/iab-digital-audio-grew-10-in-2025-as-podcasts-near-3b/>

Podcast Ads

- Two main types: baked-in (recorded permanently into the file) vs. dynamically inserted (DAI — different listeners/downloads may get different ads, or none)
- Host-read/organic-sounding ads vs. programmatic/produced ads
- Why this matters for research:
 - The "same" episode file can differ across downloads/time
 - Ads are a confound for content analysis (topic models, sentiment, etc.)
 - Ads add transcription noise and are rarely delimited in metadata
- Open problem: no completely automated way to strip ad segments from most feeds

Popularity & Engagement Data

- Platforms do not expose true listen/download counts publicly
- Common proxies (all imperfect):
 - Chart rankings (Apple/Spotify charts — algorithmic, not raw counts)
 - Ratings & review counts
 - Social media mentions/shares
- Some individual apps do expose numbers directly — e.g., Castbox publishes subscriber counts on-platform — but this only reflects that one app's user base, not overall popularity
- Third-party aggregators like Podchaser attempt to combine signals across multiple sources into broader popularity/ranking metrics, but offer little transparency into their underlying methodology
- Chart-based sampling skews toward already-popular, English-language, US-centric shows

Beyond RSS: Alternative Platforms

- YouTube: increasingly a primary distribution channel for video podcasts, some with no RSS feed equivalent
 - YouTube is now the **most** popular platform for consuming podcasts (but unclear if people are actually watching them)
- Spotify-exclusive shows: not discoverable via open RSS/PodcastIndex at all
- Implication: an RSS/PodcastIndex-based corpus is systematically incomplete — it under-represents video-first and platform-exclusive content

Joseph Bernstein. "Who Is Watching All These Podcasts?" The New York Times, July 20, 2025. <https://www.nytimes.com/2025/07/20/style/podcast-video-audience.html>

Recap & Bridge to Next Section

- Podcasts trade social media's structure and transparency for long-form, identifiable, less-curated discourse
- Podcasts are decentralized; PodcastIndex is a practical, open entry point
- Real corpora require handling dead/duplicate feeds, ad contamination, and incomplete metadata
- Popularity and category data are useful but biased/opaque proxies
- Next: processing audio data into something useable