

Ethical & Legal Considerations for Podcast-Based Research

David Jurgens and Dallas Card
July 28, 2026



Why These Matter for Podcasts

- Podcasts are public speech, but produced by real, identifiable people, often speaking candidly in an informal, conversational register
- "Publicly available" does not mean "ethically or legally unrestricted for research use"
- Three main areas of concern: copyright/redistribution, privacy (identifiability, sensitive disclosures), consent/human subjects
- If working with platforms like Spotify, platform Terms of Service may also be relevant!

Copyright & Redistribution Basics

- Best to assume podcast audio is a copyrighted work owned by the creator
- Technical access via open RSS feeds does not equal a legal right to redistribute
- Exact rules (fair use, fair dealing, text-and-data-mining exceptions) vary by jurisdiction and are genuinely unsettled in places
- As when working with social media data, analyzing data in-house is generally lower-risk than publishing/redistributing
- Caveat: This is not legal advice!

Principles of Fair Use (US)

1. **Purpose and character of the use** — is it transformative? Nonprofit/educational/research use is favored, but not automatically exempt
2. **Nature of the copyrighted work** — factual/informational content gets more latitude than highly creative work; podcasts are often a mix of both
3. **Amount and substantiality used** — a short excerpt is much safer than redistributing whole episodes or full transcripts
4. **Effect on the market** — does the use substitute for, or harm demand for, the original?
 - No bright-line rule: fair use is a case-by-case defense, typically only tested if challenged — not something to self-certify with confidence
 - Related but narrower doctrines exist elsewhere (e.g., "fair dealing" in the UK/Commonwealth) — usually an enumerated list of specific permitted purposes rather than a flexible four-factor balancing test

<https://www.copyright.gov/fair-use/>

Privacy Risks

- Many hosts and guests are public figures, intentionally speaking to the public, but this does not necessarily apply to everyone!
- **Data transformation:** transcription makes previously "hard to search" audio trivially searchable, fundamentally changing its exposure
- **Identifiability:** hosts/guests are typically named, but not always; combining podcast data with other public sources increases re-identification risk, especially for smaller/independent shows with limited existing public presence
- **Public ≠ fair game (contextual integrity):** sharing a story with listeners is not the same as consenting to have it mined as research data

Consent & Human Subjects Considerations

- Podcasters and guests did not consent to research use when they recorded
- Guests especially may have even less awareness/control than the host over how the episode is used later
- An IRB exemption is likely to apply in most cases, but whether this counts as human subjects research will vary by institution
- see Fiesler and Proferes, "'Participant' Perceptions of Twitter Research Ethics": some hosts and guests may be surprised and/or displeased if their output is being used for research (but it's complicated)

Fiesler and Proferes, "'Participant' Perceptions of Twitter Research Ethics", In *Social Media + Society* (2018).

Case Study: How SPoRC Handles These Issues

- **Only redistributes post-processed data** (transcripts/derived features), never raw audio — not a substitute for the original product and eliminates involvement in potential voice cloning, etc.
- **Data provided in bulk** — not replacing podcasts available as an on-demand audio medium
- **Restricts usage to research purposes** — license/terms narrow the scope of redistribution rather than releasing openly for any use
- **Only older episodes (from 2020)** — creators who wanted content removed have had more opportunity to do so already, lowering the odds of redistributing something someone actively wants taken down
- **Offers an opt-out mechanism** — lets podcasters request removal of their content from the dataset
- None of these eliminate risk entirely, but together they meaningfully reduce it — a practical template, not a legal guarantee

Recap & Key Takeaways

- Podcast data involves real, identifiable people, often disclosing more than they'd expect to have "mined"
- Privacy and consent/IRB questions deserve real thought, even for data that's technically public
- Copyright/redistribution rules are jurisdiction-specific and not something to self-certify — consult your institution
- Practical mitigations like SPoRC's (limit redistribution, use older data, restrict to research use, offer opt-out) can meaningfully reduce risk
- Treat "it's public" as a starting point for judgment, not a conclusion