Open Analytics Platform for Podcasts

Abstract

In this proposal we describe a new open platform aiming to support the **open and independent podcast ecosystem**. It solves one of the biggest challenges for podcast hosts to become independent of big platforms such as Spotify and Apple: analytics and listening behavior. The platform consists of three main components. The **Open Podcast Analytics API** is an open standard for podcast players. It is complemented by a **reverse-proxy** component to transparently collect data until the standard has been widely adopted. Furthermore, an **importer** allows the **consolidation of analytical data provided by big platforms** such as Spotify. This triangle builds a solid analytical foundation for independent podcast hosts to compete with big players on the market.

Introduction and current podcast market situation

There are over 2 million podcasts[1] and the podcast industry will generate $2 billion in revenue this year and $4 billion by 2024[2]. The podcast ecosystem was designed using Open Source technologies like RSS and HTTP. Unfortunately the market is currently dominated by Spotify, Apple Music and Amazon Music who are working towards a closed podcast ecosystem with the goal to monetize their products and exclusive shows. Also, recent acquisitions of Spotify[3] show a strong push for market dominance.

Due to the tooling and other benefits in Spotify's closed system, content creators become more and more dependent on this closed environment. Because of this, the availability of free and open tools in this space has become more important than ever and is regularly discussed by media representatives (e.g. https://detektor.fm/wirtschaft/brand-eins-podcast-live-republica-podcasts-offenes-oekosystem)

Problem definition

Due to the huge investment of big players into their podcast platforms, the equality between these platforms and independent podcast hosts is no longer given as open tools are limited in comparison. Especially the RSS standard, which was a crucial factor for the success of the open podcast system, is limited in providing state of the art analytical data about the listening behavior of listeners.

A product like Google analytics with detailed information on listeners, doesn't exist in the realm of podcasts. Details like the age group, interests, location data and listening habits are important for advertisers but not available through an RSS feed.[4]

Having detailed information on users helps drive revenue — this is also true for podcasts. We want the podcast creators to receive most of the profits from their work and not the big platforms. What is even more important is independence to protect freedom of speech: podcast creators run the risk of getting locked out of a platform if they refuse to accept the terms of service or the publisher simply doesn't agree with the content.

[1] https://influencermarketinghub.com/podcast-statistics/
[2] https://www.theverge.com/2022/5/10/23065056/podcasting-industry-iab-report-audacy-earnings-patreon-pulitzer
[3] https://www.crunchbase.com/search/acquisitions/field/organizations/num_acquisitions/spotify
[4] https://detektor.fm/wirtschaft/brand-eins-podcast-live-republica-podcasts-offenes-oekosystem

Proposed Solution

To tackle the analytics challenges in the podcast ecosystem as described above, we propose an Open Analytics Platform which consists of three main parts:

1. Open Analytics Podcast API to collect fine grained usage data
2. Reverse proxy to collect basic analytics data and extended listener attribution
3. Analytics importer to consolidate analytics data provided by closed platforms such as Spotify

In the following sections the implementation and challenges of the three main columns of the platform are described in detail. The three main components can also be identified in Figure 1.

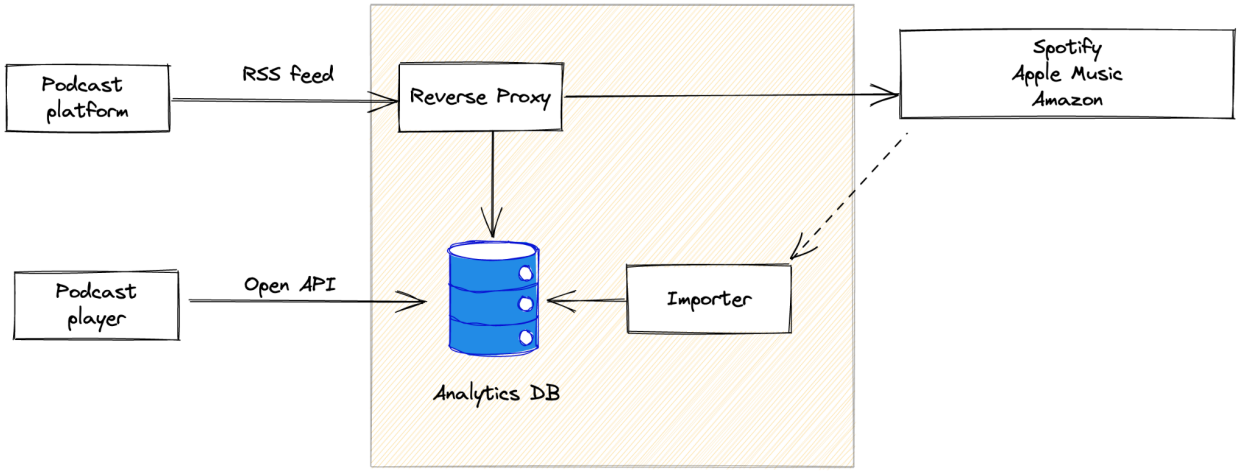

Figure 1: Architecture of the proposed Open Analytics Platform for Podcasts

1. Open Analytics Podcast API

For content creators it is crucial to get feedback from their audience to improve their content.

State of the art content development is based on fine grained metrics which are even available
in the very traditional world of TV (in Germany by AGF[5]). For website owners it is common to get a very detailed view on their users by e.g. Google Analytics or the Open Source alternative Matomo[6]. In contrast, podcast hosts are limited to very basic analytics such as the number of downloads which are not very reliable as an indication for listeners.

We propose a GDPR-compliant protocol, which could become the open standard for podcast analytics. The protocol would provide the following advantages:

- Play, pause, contextual information (for example if the listener is currently traveling), and listener profile (age, etc.).
- Protocol based on API specification and defined as metadata inside RSS feed. This would allow every host to use their own analytics server.
- Players can send data to the API (defined by podcast).
- Centralized API to collect data.

2. Reverse proxy to collect basic analytics data

The analytics data about episode downloads is usually very limited on many podcast hosting platforms. Instead, we want creators to have full access to this information and own it. In order to do that, we propose a reverse proxy, which sits between a podcast platform (e.g. Anchor) and the client (e.g. Spotify or any podcast player).
It serves the RSS feed from the hosting platform but stores tracking information in real time. For example, if Spotify requests the RSS feed, a dynamically adapted RSS feed attributed to Spotify is delivered. This allows the reverse proxy to attribute all future requests to Spotify, independent of the actual client used by the user (e.g. Spotify web client, Spotify mobile client, etc).
Furthermore, all incoming requests for episodes are tracked and detailed statistics are collected. The tracking itself will be based on well-established standards such as Matomo Open Source tracking. The reverse-proxy can be self-hosted and will be fully Open Source.

[5] https://www.agf.de/
[6] https://matomo.org/

As the reverse proxy transparently sits between the hosting platform and the client and is based on the RSS standard, it is fully compatible to all hosting platforms. It can be easily installed and doesn't need any changes on the hosting platform or user side.

3. Analytics importer

Internal statistics and detailed usage data, such as the number of people who were listening to the whole episode, are provided by big platforms such as Spotify or Apple. Alas, the data is only accessible within their dashboard. To the best of our knowledge, there is no way to aggregate and consolidate the data to have a unified view on listening behavior.

Our solution is a data importer, which runs as a background job and connects to the APIs of each provider. It would store the data in a unified format which can be visualized with tools like Matomo or Grafana. For that, the user has to provide their login credentials to the importer. The importer can be self-hosted and will be open source.

Benefits for the media industry

The podcasts scene has taken an important role in the media realm and is evolving into a major revenue stream for news and information publishing. Especially the open podcast scene plays an important role in the podcast ecosystem as emphasized by Marc Krueger (Podcasts, Audio, Strategie rbb / ARD[7]).
The proposed Open Source solution including the Open Analytics API standard provides a powerful toolset to independent podcast hosts and content creators and empowers them to compete with big players such as Spotify or Apple. Thus, the competition will increase, foster diversity, cultivate an open podcast ecosystem, and broaden the spectrum of podcast formats.

[7] https://detektor.fm/wirtschaft/brand-eins-podcast-live-republica-podcasts-offenes-oekosystem

Example use cases

Assuming the availability of all collected data as described in the previous section, the data can be used to realize the following two example use cases which are not possible without a unified access to podcast listener analytics:

State of the art content development

Podcast hosts can drill down in the aggregated and consolidated analytics data and derive crucial information (e.g. listening preferences, listening context such as car or at home) for their content development or advertisement partner (e.g. listener profiles) to monetize their podcast. This professionalization of the podcast market was also emphasized by Christian Schalt (Chief Digital Content Officer RTL Radio Deutschland GmbH) and Marc Krueger (Podcasts, Audio, Strategie rbb / ARD)[8].

Podcast Recommendation Engine based on artificial intelligence

Due to the lack of proper data, there is no open recommender engine, which provides the functionality for users to explore the humongous space of podcasts. For example the very popular machine learning approach of collaborative filtering relies on usage data to find users with similar listening preferences. For such an approach it is crucial to have access to fine grained listener analytics to recommend suitable podcasts to other listeners. This data would be provided by the proposed platform.

[8] https://detektor.fm/wirtschaft/brand-eins-podcast-live-republica-podcasts-offenes-oekosystem

Conclusion

There is a massive opportunity to build an open podcast platform for the next 2 million podcasts.
We want content creators to own their data and provide them tools to be able to compete with big platforms such as Spotify or Apple. This includes public service radio stations, which should be independent of big corporations.

There is strong potential to create a leading open standard for podcast analytics. We want to build the foundation for a better podcast ecosystem where creators own their data and information is not locked in behind the walled gardens of big corporations.

About the Team

Matthias Endler (core team)

Matthias is a backend engineer who worked at trivago where he met Wolfgang and Andy. He is passionate about the open web and writes services in Python, Go, and mostly Rust. He cares deeply about the engineering community. His blog at endler.dev and his YouTube channel “Hello Rust” helped him gain valuable experience with managing community. He is an Open Source developer for more than ten years. He has an M.Sc. in Computer Science.

Wolfgang Gassler (core team)

Wolfgang is an Open Source enthusiast and started programming in the late 90s. He has worked as developer, project manager, and engineering manager in various industries for more than 25 years and obtained a computer science PhD in the field of databases and information systems with a focus on recommender systems and semantic knowledge bases. Wolfgang experimented with podcasts already in 2011, when he published his first podcast about databases. Since 2022, he is one of the hosts of the weekly German podcast “Engineering Kiosk” covering tech culture, software development, and Open Source topics.

Andy Grunwald (advisor)

Andy is a professional Software Engineer and Engineering Manager focusing on Backend, Infrastructure, and Reliability Engineering. For more than 15 years, he has been focusing on Open Source, Engineering Culture, Community, and Knowledge Sharing. For 10 years, Andy has been organizing one of the biggest Meetups in Germany called Web Engineering Düsseldorf, and a non-profit conference named localhost.engineering in 2019.
Since the beginning of 2022, Andy has been one of the hosts of a growing German tech podcast Engineering Kiosk and has built up his increasing experience in podcast publishing since then. Throughout all these projects, there is one common theme: Acting as a developer evangelist for technology, innovation, knowledge sharing, openness, and people first.