

PodcastRE Analytics: Using RSS to Study the Cultures and Norms of Podcasting

Eric Hoyt <ehoyt_at_wisc_dot_edu>, University of Wisconsin-Madison
J.J. Bersch <jbersch_at_wisc_dot_edu>, University of Wisconsin-Madison
Susan Noh <snoh8_at_wisc_dot_edu>, University of Wisconsin-Madison
Samuel Hansen <hansensm_at_umich_dot_edu>, University of Michigan and University of Wisconsin-Madison
Jacob Mertens <jmertens2_at_wisc_dot_edu>, University of Wisconsin-Madison
Jeremy Wade Morris <jwmorris2_at_wisc_dot_edu>, University of Wisconsin-Madison

Abstract

Over the past decade, podcasting has grown into one of the most important media forms in the world. This article argues that podcasting's unique technical affordances — particularly RSS feeds and user-entered metadata — open up productive methods for exploring the cultural practices and meanings of the medium. We share three different methods for studying RSS feeds and podcast metadata: 1) visualizing how topics and keywords trend over time; 2) visualizing networks of commonly associated keywords entered by podcasters; and 3) analyzing norms and common practices for the duration of podcasts (as a time-based media format, podcasting is unusual in that it is not bound by the programming schedules and technical limitations that provide strict parameters for most audiovisual forms). The methods and preliminary results reveal how metadata can function as a surrogate for studying large collections of time-based media objects. Yet our study also shows that, when it comes to born digital media, the metadata are never fully separate from the objects they describe. We argue that future work in AV in DH needs to delineate between methods best suited for digitized media collections compared to those most appropriate for born digital media collections.

Introduction

As a cultural form, podcasting resists easy definition. It is a highly porous medium, traveling with us over earbuds, phone speakers, and car stereos, accompanying us on commutes, jogs, errands, and road trips. It's a sound-based media that we also experience visually through live shows, thumbnail icons, and t-shirts that say "Friend of the Pod" or "Night Vale Community College." Despite these definitional challenges, the medium, by most measures, is booming — with the quantity of podcasts, listeners, advertising revenue, and non-profit funding increasing sharply year after year, including an "explosive" 2018, which saw the number of U.S. people over the age of twelve who have ever listened to a podcast climb above 50% for the first time [Edison 2019] [PodNews 2019].

If we consider podcasts from a purely technical standpoint, it is possible to narrow the definition slightly. As we have elaborated elsewhere, a podcast can be defined as a collection of downloadable files, of any format, served, with accompanying metadata, via an open updatable internet feed, primarily RSS [Hansen 2020]. An XML-based protocol, RSS allows for podcasters to easily publish their completed work and distribute it to audiences, who can opt to subscribe to particular feeds. In many ways, the metadata and the open feed are what separate a podcast from other media files on the internet, including other forms of on-demand audio (for example, music streaming platforms and audio book companies). Because RSS feeds are open, podcasting is platform-independent. Listeners can subscribe to feeds through a number of different podcatching apps and a variety of platforms. At least for now.

Both the expansive cultural meanings of podcasting and the rigid technical definition have animated our work over the past three years on the PodcastRE database and our desire to study and preserve this emerging format. Based at the

1

2

3

University of Wisconsin-Madison and supported by grants from the university and the NEH, PodcastRE (short for Podcast Research and accessible at <http://podcastre.org>) is a data preservation and research initiative. As we write in April 2020, the PodcastRE database has grown to over 2.5 million podcast episodes from over 16,000 unique RSS feeds which occupy 99 terabytes of space within our RAID storage array. The collection has expanded beyond what any individual could listen to within a lifetime, and it only keeps growing.

What can researchers do with millions of podcast episodes and their associated metadata? This article seeks to contribute to the body of digital humanities scholarship invested in harnessing the affordances of digital technology to investigate cultural data at a large scale [Jockers 2013] [Underwood 2019] [Clement 2016a] [Clement 2016b]. We argue that podcasting's unique technical affordances (e.g. RSS and metadata) open up productive methods for exploring the cultural practices and meanings of the medium. These methods, in turn, hold broader relevance for scholars seeking to integrate media studies with computational analysis (or, as the theme of this special issue nicely puts it, "AV in DH"). Our study of podcasting shows the ways that metadata can function as a surrogate for studying large collections of time-based media objects. To put it simply, it's far easier to query 2 million metadata records than it is to query 2 million media files of movies, TV episodes, or audio programs. Yet our study also shows that, when it comes to born digital media, the metadata are never fully separate from the objects they describe, nor can they fully describe, or replace the need for, returning to the media themselves during the final analysis. As a result, future work in AV in DH needs to thoughtfully delineate between methods best suited for *digitized media collections* compared to those most appropriate for *born digital media collections*.

In this article, we share three different methods for analyzing the metadata of PodcastRE's born digital corpus, assessing the strengths and weaknesses of each method and sharing preliminary results. First, as we will share, PodcastRE's Term Frequency Line Graph (<http://podcastre.org/lineGraph>) allows researchers to create visualizations of trending topics and keywords. Interpreting the results of the line graphs can be challenging, however, due to the messiness of the underlying metadata and the problem of "normalizing" a rapidly growing corpus and medium. Second, PodcastRE's Associated Keyword Cloud visualization tool (available at <http://podcastre.org/wordCloud>) enables researchers to query a keyword and generate a word cloud that displays the other keywords that appear alongside that keyword in podcasts. We argue that this data visualization harnesses the potential of the medium's inconsistent and messy metadata and allows for open-ended explorations, serendipitous discovery, and new questions about the agency of podcasters in self-defining their cultural output and connecting it with particular communities and conversations. Third and finally, we share approaches for studying the duration of podcasts. As a time-based media format, podcasting is unusual in that it is not bound by the programming schedules and technical limitations that provide strict parameters for most audiovisual forms, such as movies, television, and radio. If a podcast can run anywhere from a couple of seconds to several hours in length, how do norms and common practices develop that establish optimal models for a podcast's duration? To investigate these questions, we exported CSVs from the database (using a "mediaFileDuration" field generated by the individual episode files), sorted them into meaningful sample groups, and investigated the data for patterns.

Ultimately, our goal for this article is to share research and methods for studying the explosion of audio culture taking place in podcasting and through the sonic communities and conversations podcasting draws together. These methods are especially well suited for studying audio, but they would also be valuable for exploring online video collections and other digital media objects. As in any research study, though, it's important to address the specific before offering broader generalizations. With that in mind, we would like to now turn to a consideration of the history and design of RSS feeds, how they inform the underlying dataset (the PodcastRE collection), and the affordances and challenges of these structures.

RSS Feeds, the PodcastRE Collection, and Working with Messy Metadata

As the protocol that has enabled inconsistent and idiosyncratic podcast metadata to proliferate across the internet, it is fitting that there is no singular consensus on what the initials "RSS" should stand for. *Real Simple Syndication* is the most commonly cited meaning. But "Rich Site Summary" and "Resource Description Framework (RDF) Site Summary" have also been cited as the basis of the name. There is no question, however, that the technology has played a pivotal

role in the growth of podcasting and infrastructure of PodcastRE.

On March 15, 1999, Netscape published the first specification for RSS [RSS Advisory Board 2019]. Based on XML and developed by Ramanathan Guha and Dan Libby, RSS was created so that the Netscape home page “My Netscape” could be refreshed with new content from webpages which used the specification [Hines 1999, para. 9]. Over the next couple of years, RSS went through multiple iterations, and then on December 25, 2000, Dave Winer and UserLand software released RSS 0.92 [RSS Advisory Board 2019]. It was this version of RSS which is most important in the history of podcasting as it was the first version which included the <enclosure> tag, which allowed for the attachment of media files. Concocted by Winer, with strong prompting from Adam Curry, as a way to deliver high quality multimedia files over the internet without the quality and wait time issues which plagued early streaming, the first use of an <enclosure> was to distribute a set of Grateful Dead MP3 files [Winer 2001, para. 44], presaging its dominant use in the years to come.

8

RSS continued to develop for the next decade and mostly stabilized as a specification with version 2.0.11 on March 30, 2009 [RSS Advisory Board 2019]. Since then the only updates to the format have come in the form of XML Namespaces, which are ways of adding outside-of-specification elements to XML documents that are commonly used by Apple, Google and other podcast distributors to expand metadata options for commercial purposes [Bray et al. 2009, para. 1] [Bergen 2015, para. 2]. Even as podcasting apps, playback technologies, and the on-demand sound industry has changed throughout the 2010s, the basic structure and syntax of RSS has remained constant, keeping the circulation of most podcasts relatively open and freely downloadable.

9

The open infrastructure of RSS also became foundational for our work on PodcastRE [Morris 2019]. Interested in studying podcasts, but worried about the vulnerability of digital audio files, we realized in early 2014 that there were few searchable databases of podcasts for studying and analyzing the booming audio culture taking place in podcasting. We began rather humbly by logging RSS feeds manually in iTunes and downloading audio files to a local hard drive, tracking as best we could podcasts that were being cited in the press as part of the renewed interest in podcasting, like *Serial* or *Welcome to Nightvale* [Adams 2015]. As the project grew, we implemented a more coherent collection process, and since 2018, we have been saving podcasts included in discussions of podcasting’s “golden age” as well as interrogating what podcasts are being left out of that discussion. We’ve navigated the need to preserve the “popular” by automating the collection of a particular index of what’s popular: the Apple Podcasts top 100 lists for the U.S., Great Britain, France, and Australia every 24 hours. This automated approach toward collecting embraces both the affordances of the digital media and the MPLP (More Product, Less Process) model proposed by Mark A. Greene and Dennis Meissner [Green and Meissner 2005]. Our efforts to identify and collect significant podcasts beyond the Apple Podcasts top 100 have been driven by collaborations with scholars who are researching independent podcasts produced by women, indigenous peoples, and people of color [Wang 2020] [Florini 2017] and by following the work of other networks, directories and databases devoted to highlighting marginalized/less visible podcasts (Podcasts in Color, Women In Podcasting, PotLuck Podcast Network, etc).

10

PodcastRE’s collection of 2 million podcast episodes has thus been built by a combination of algorithmic methods and informed hand-selections. There’s also a “submit a podcast” feature on the project’s website that allows individuals to add the RSS feeds for podcasts they’d like preserved. The 16,000 archived podcast feeds are a fraction of the over 1,000,000 podcast feeds that, according to estimates, are currently being distributed [Podcast 2020]. But the PodcastRE collection does offer a valuable and diverse cross-section of English-language podcasts from the past several years.

11

The common thread through all of this work has been RSS. To put it simply, if a podcast doesn’t have an RSS feed, then we cannot yet preserve it within our system. This is one of the reasons why the technical definition of a podcast — an open feed of downloadable files and associated metadata — has been so important to our work on PodcastRE. To achieve our goals of scale, not only did we need to be able to download podcast episode files automatically, we also needed to gather the metadata we could store automatically. For PodcastRE, the elements available through the RSS specification, and its associated namespaces, are as important as the podcast episode files themselves. RSS, in other words, defines the possible universe of metadata for the podcasts archived in PodcastRE.

12

What we did not immediately appreciate was how messy, idiosyncratic, and incomplete the world of podcasting metadata would prove to be. Podcast RSS metadata is a world away from the familiar and relatively consistent metadata fields of TEI and Dublin Core. One reason is the relatively sparse number of elements which are required for a feed to be valid. In fact, an RSS feed only needs four elements to be present in order to be valid: the <channel> parent element with associated <title>, <link>, and <description> elements. This would be a feed without content though as it would contain no <item>s [Winer 2003, para. 13]. Because authors fully manage their own RSS feeds, and the entry of the metadata into them, they are directly responsible for the depth and quality of the metadata. This aspect of podcast metadata cannot be stressed too highly. With the exception of a few elements like <googleplay:category> and <itunes:type>, there are almost no constraints on what podcast authors put into the various elements. Even fixed format elements like <pubDate>, which seems rather self explanatory to mean the date on which a podcast episode was published into a feed, can end up being used by authors to mean something very different. For example, there are many <pubDate>s before 1950 in the metadata for *The Reith Lectures* podcast from the BBC, long before the term podcast was ever coined. Instead, the series uses <pubDate> to mean the day the lecture was originally given. RSS authors continue to have the authority to change anything they wish—including something as fundamental as the title of an episode, or even their whole podcast, at any time. Looking in PodcastRE, we see examples related to branding, as when *Bookworm* added their network and became *KCRW's Bookworm*; or to SEO, as when *Highest Self Podcast* added some terms and turned into *Highest Self Podcast: Modern Spirituality, Ayurveda, Conscious Entrepreneurship, Mind-Body Balance*.

13

The inconsistent and incomplete metadata records created major challenges for our efforts to systematically preserve podcasts and make them easily searchable. We found it especially unfortunate that metadata fields that could have been revelatory for search faceting and social network analysis (fields such as <network>, <host>, and <contributor>) are not a part of any current podcast RSS specifications. Yet it was equally clear that authoritative approaches to metadata had their own problems and major blind spots. The inadequacies and biases of Library of Congress subject headings have received considerable attention within the discipline of information studies. For example, Juliet L. Hardesty [Hardesty 2019] has argued that the subject headings generally take the primacy of white men as a default; “Robert Frost” is cataloged under “Poets, American” without reference to gender or race, whereas Maya Angelou is listed under subjects including “African American women authors” and “African American authors.” The catalogers, in these cases, are applying a schema that upholds a white patriarchal worldview and minimizes both the needs of users and the ways in which creators and subjects would choose to define themselves.

14

In contrast, podcast creators have a tremendous amount of agency in how they define themselves and attempt to connect with users (i.e. listeners and audiences). When the creators of the PHX podcast entered the keywords “podsincolor” and “women of color” within their RSS feed, they actively chose to present themselves this way and place their work within a larger network of podcasts produced by people of color. The flexibility that characterizes metadata practices prove to be critical for marginalized podcasters in forming community, as they seek to carve out space for themselves within media production practices and platforms that consistently privilege hegemonic whiteness, accepted paradigms of masculinity, and heteronormativity. While this does not necessarily mean that self-policing within metadata production does not happen as a result of the asymmetrical power dynamics between platform and creator, it does still provide yet another avenue in which marginalized communities can stand in opposition to the individualistic neoliberal ideologies that undergird contemporary user-driven media production [Hogan 2008] [Florini 2017]. It is critical to note, however, that the non-uniformity of metadata production yields ambivalent practices, both where innovative podcasters can resist the influence of various dominant ideologies, while others use this space to reinforce their centrality simultaneously.

15

For example, there are also many instances of podcasters stuffing their RSS with keywords in order to make them prominent within content aggregators and “podcatchers.” The internet abounds with advice and speculations for search engine optimization and strategies that can be utilized in order to gain attention to one’s content, such as the optimal number of keywords, the kinds of thumbnail images that should be connected to content, and more [Crowe 2019] [Podcast 2015]. The opaque nature of how Apple Podcasts organizes its search results impacts the manner in which metadata is written, and this influences the ways that podcast creators self-define their own content. The dominating

16

influence of Apple Podcasts categories can be observed by the fact that within the entirety of the PodcastRE database, the most used keywords lists are dominated by terms that are outlined either fully or in part by Apple Podcasts genre specifications. For example, with the exception of the words, “podcast” and “radio,” the top fifteen keywords for the podcast classification (the entire podcast series) terms all cohere to various genre classifications within Apple Podcasts. Similar patterns can be seen for the episodic classification where, with the exception of “Talk Radio,” “Podcast” and a blank space/uncategorized, the top ten keywords reflect Apple Podcasts categories. The large amount of uncategorized keyword terms may gesture towards the fact that after 2013, the keywords metadata field became deprecated, meaning that it no longer affected the output of Apple Podcasts’ search engine algorithms (<https://support.libsyn.com/kb/the-rss-feed/>). After this discovery, many podcasters may have forgone the labor of adding keywords, as the fields that most influence search engine optimization are now the title, author, and description tags.

Even though Apple Podcasts deprecated keywords within its search algorithm, we became excited about the role keywords could play for our work on PodcastRE. What sort of data visualizations and discoveries might be possible by harnessing RSS metadata at scale? Ultimately, we built two data visualizations for the site. Perhaps not surprisingly, the more successful of the two was the one that most embraced the idiosyncratic, messy, and user-created nature of RSS.

17

Graphing Metadata Term Frequency Across Time

How do keywords and other fields used to describe podcasts change over time? Could tracking these changes prove useful for spotting trending topics within the podcasting ecosystem? To explore these questions, we created PodcastRE’s Term Frequency Line Graph (publicly available at <https://podcastre.org/lineGraph>), which tracks the frequency across time that any word or phrase within the metadata appears. The fields searched include the title, creator, synopses, and keywords. A visualization graphing the term “money” within PodcastRE is displayed below. If a user clicks on any point within the graph, their browser opens up a new tab displaying all of the matching podcast episodes from that month or year that contain a matching search term.

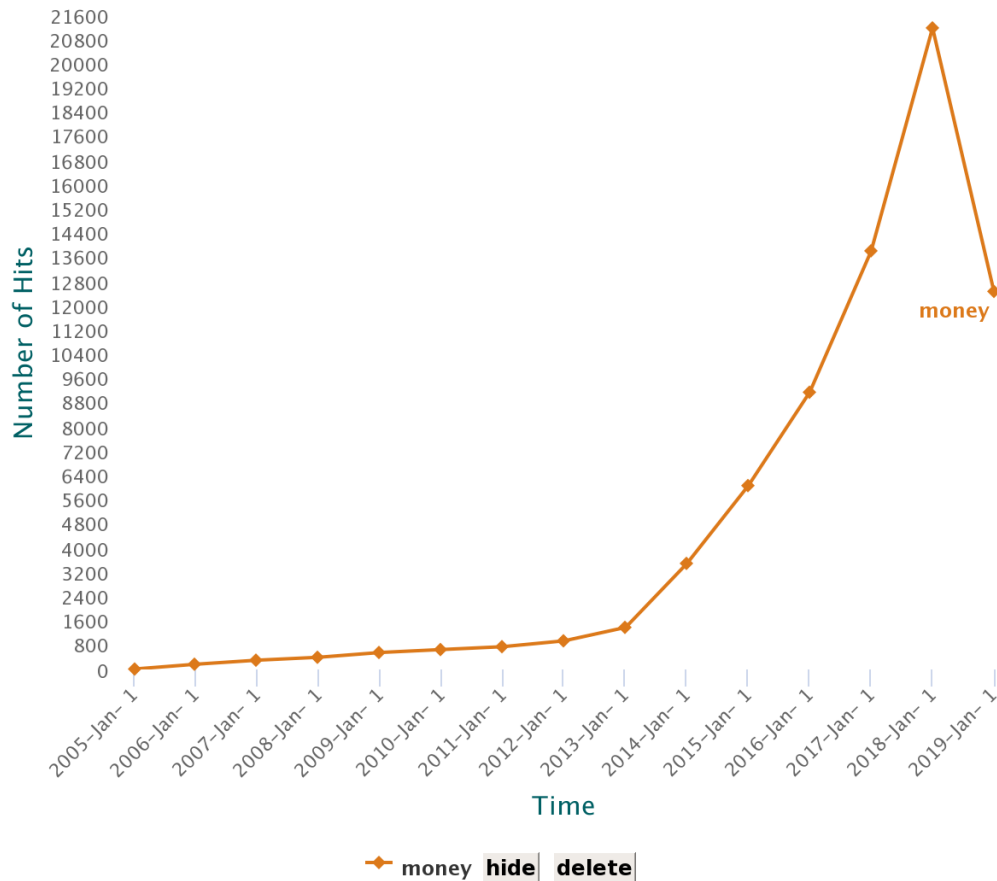
18

The Term Frequency Line Graph searches metadata included within individual podcast episodes and across the entire feeds (for example, while “NBA” may be a keyword that describes a podcast feed as a whole, “China” may be a keyword that describes a topic discussed within one episode of the podcast). By default, the X-axis of the graph is divided by years; however, users can toggle to a monthly scale. This allows for researchers to see when certain topics or keyword phrases spike on a seasonal cycle (for example, “baseball” consistently has an uptick during the playoffs every October) versus more macro-scale trends that rise and fall over a period of years.

19

Keywords Over Time

Source: PodcastRE



PodcastRE.org

Figure 1. Term Frequency Line graph of “money”, tracked over time, within PodcastRE’s corpus. Tool is available at <http://podcastre.org/lineGraph>.

When researchers use the Term Frequency Line Graph to look for trends across a span of years, however, they quickly encounter an interpretive challenge: almost any term they search will appear to dramatically increase in 2017 and 2018. This is because the PodcastRE collection grew exponentially over those two years, a result of the growth in the podcasting ecosystem as a whole and our own curatorial decision to automatically preserve any feed that appears on the Apple Podcasts Top 100 chart in the U.S., U.K., Australia, or France. While we give users the ability to “normalize” the graph results (which employs an equation to account for the larger number of podcasts from some years compared than others), we know this feature has its limits. What does it mean to “normalize” the number of podcasts during a period in which the medium is rapidly evolving?

20

We have tried to address this interpretive challenge through a “Rate of Episodes Added” button, which provides contextualization in regard to the database itself. By showing how many episodes are added per year, users can see how the rate of growth in the database can affect the numbers that are being shown for any query’s term frequency. Additionally, the “Area Graph” button transforms the data into a stacked graph, which allows for comparisons across multiple queries at particular moments in time and reminds users that the graphs are malleable and dynamic. Finally, the user can move to a more granular level at any point by clicking on a point in the graph, allowing them to investigate the actual podcast feeds and episodes that appear as abstractions within the graph. Users can save the data to a CSV file, a JPG, PNG or SVG vector image, so that this data can be applied to a variety of presentational contexts.

21

In many ways, PodcastRE’s Term Frequency Line Graph exemplifies the limitations digital humanists are likely to encounter when applying data visualizations built for *digitized text collections* to *born-digital media collections*. We

22

modeled the user-experience and technological framework of PodcastRE's Term Frequency Line Graph on that of the Arclight app (<http://projectarclight.org>), which searches the 2.5 million page corpus of the Media History Digital Library (MHDL) [Hoyt 2016]. The MHDL is composed of books and magazines pertaining to the histories of film, broadcasting, and recorded sound from 1915 to 1960, which is an especially robust period for the searching of named entities (such as people, film titles, or radio station call letters). Additionally, the normalization function for Arclight graphs works quite well (the most represented year of 1915-1960 is only double in size of the least represented year, avoiding PodcastRE's challenge of grappling with exponential growth). Normalized searches for the names of movie stars, for example, generally map onto the arcs of their popularity and/or notoriety, sometimes, though not always, with surprising results. Data visualizations built for searching entities within large corpora of digitized texts are less adept at producing immediately legible results for searching the metadata keywords of a rapidly growing born-digital medium. What would it mean to design a data visualization tool that embraced the messiness of born-digital objects and their metadata, rather than trying to smooth them out?

Associated Keyword Word Cloud

In developing PodcastRE's Associated Keyword Word Cloud, we sought to harness and foreground the specificities and idiosyncrasies of born digital media collections. This data visualization takes the keywords that podcasters entered to describe their work and puts them into conversation with other podcasters' keywords. A specific example is helpful for understanding how it works.

23

Using the keyword "money", in a search conducted in the fall of 2019, we found the term appeared in the metadata of 68,619 podcast episodes saved within PodcastRE, collected from 587 discrete RSS feeds. The other keywords that appear most frequently along with "money" in podcast metadata are visualized below (see Figure 2). This visualization includes predictable matches within the popular financial self-help genre (e.g. "wealth," "business," "entrepreneur"), as well as meaningful intersections that lay outside financially-oriented podcasts (e.g. "spirituality," "Relationships & Sex," "Fear"). When a user clicks on the keyword value in the cloud, the user is immediately transferred to the PodcastRE database interface, where it shows all of the podcasts that used these paired keyword values. Figure 3 reveals the results of the podcasts that contain both the keywords "money" and "spirituality." The process promotes serendipitous discovery and may lead the researcher toward encounters they hadn't anticipated. For example, modern witchcraft is better represented in the podcasts with "money" and "spirituality" as keywords than most traditional forms of organized religion.


24

Search PodcastRE for Related Keywords

☒ All Keywords ☐ Podcast Keywords ☐ Episode Keywords

Search

Associated Keywords



Podcast

Full Time Hustler Art & Commerce Interview Lifestyle HermanJoe Business/Automotive Fear Training Purpose & Mindset

economics Finance Foroohar education Mindset Ireland Investing Amazon FBA work Entertainment

Current Events Spirituality national_news money wealth explicit News trump An Irishman Abroad Money radio finance

Recreation/Theme_Parks/Disney/Walt_Disney_World/ publicpodcast Episodes Season 1 Technology

Startups Get Started politics Jarlath Regan history Show sex Podcasts Money & Business

Business wnyc_app_local local_wnyc technology success business wnyc episode news 1. Everyone

Charlie Build Wealth york new FIRE Drill Podcast Music blockchain

Scavenger Life Money & Career health npr Education Podcast Episodes wnyw episode news 1. Everyone

Weekend Warrior Rana Success Uncategorized science Entrepreneur News & Politics Career

bitcoin tv life Talk Radio Irish storytelling entrepreneur News & Politics Career

investing Weekly Trading Updates Irish storytelling entrepreneur News & Politics Career

Creativity Personal Growth Enjoy Your Money Comedian cryptocurrency Comedy

Blog Relationships Relationships & Sex

PodcastRE.org

Figure 2. Word cloud for the query, “money,” on the “All Keywords” category. Taken using the Associated Keyword Cloud visualization tool at <http://podcastre.org/wordCloud>.

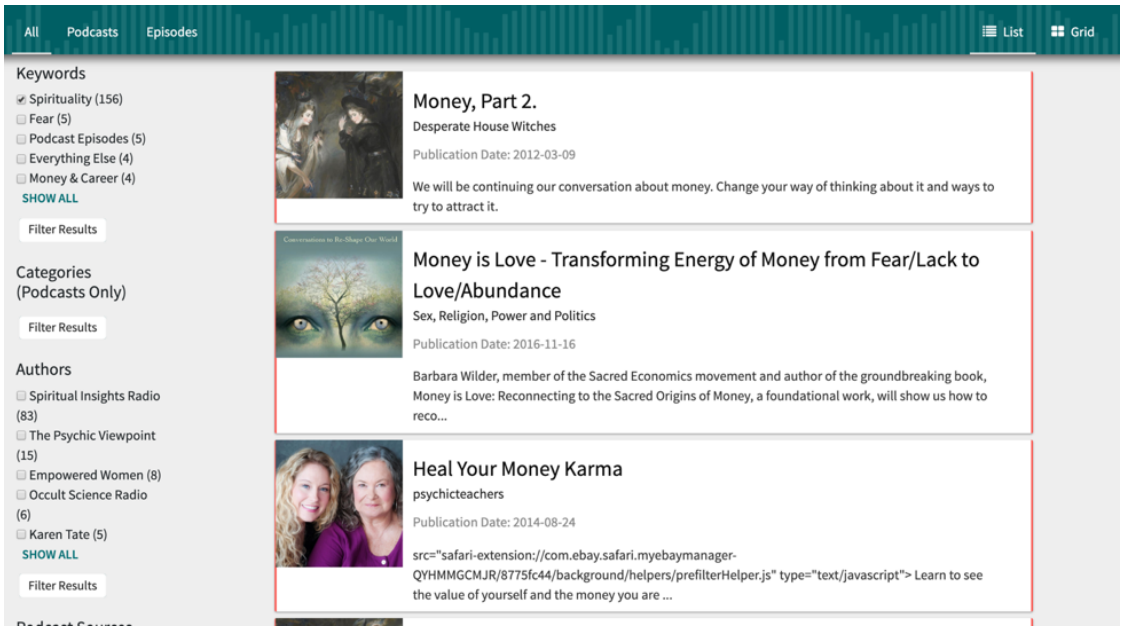


Figure 3. Screen shot of the results page for podcast episodes containing the keywords “money” and “spirituality.”

Like the Term Frequency Line Graph, the Associated Keyword Word Cloud uses the Highcharts Javascript library to animate the visualization. To retrieve the information it needs, we query the keyword metadata facet within PodcastRE's Solr search index, and we return and store them as key value pairs, with the number of podcasts that maintain both the queried keyword and the additional keyword (the hit count) next to the particular word. For example, if a user queries the word "love" within the database, a potential key value pair that would appear would be ["relationships", 163], where "relationships" would be the associated keyword for "love," and the "163" stands for how many times this keyword was added alongside the word "love." The results are sorted through keywords that have the most hits down to the associated keywords that have the least hits. By targeting this metadata keyword field and assigning the "weight" of a word to be the number that is assigned to the hit count of the key value pair, we were able to visually represent which keywords were paired most often with the queried word, by making the word with the heaviest "weight," the largest in the word cloud. Because certain topics have a range of associated keywords that spanned hundreds of words, we limited the number of keywords that can be shown on the word cloud to a maximum of 200 words. While this decision may hinder researchers from getting the full range of associated keywords, this limitation was imposed to ensure readability on the visualization. Two hundred words seemed like a reasonable count in order for researchers to gain a sense of the wide range of relational topics that podcasters were dealing with, and simultaneously allow the visualizations to be effective in showing which keywords were the most actively engaged with.

There are two options on the Associated Keyword Word Cloud interface that aid in isolating whether the keywords shown are related to podcasts in their entirety or exclusive to certain episodes. Additionally, if users want a merging of these two levels of metadata, they can search across both podcast and episode keywords by using the "All Keywords" option. In this manner, for podcasts that may deal with a wide range of topics, such as news or current events podcasts, there can be a closer examination on a micro episodic level of what kinds of keywords are used to define certain topical content. Often, the keywords that are used to describe podcasts are not uniformly applied to define episodes, so providing these two levels of analytical range gives researchers more flexibility in the kinds of questions they can ask using PodcastRE.

26

All attempts to interpret the Associated Keyword Word Clouds ultimately lead back to reflecting on the practices, norms, aspirations, and communities of the podcasters themselves.

27

As discussed earlier, keywords allow content creators to define their work to listeners and podcatcher applications. They are a space of creator agency, where podcast producers deploy keywords to create networks of ambient affiliation with other podcasts and subject matter. By making one of PodcastRE's database visualization tools intimately connected to these creator-defined keywords and their relationships to other keywords, we provide an alternative mode of discoverability apart from the algorithms that govern commercial aggregators such as Apple Podcasts. In this manner, PodcastRE hopes to provide a different approach that foregrounds creator agency and their interactions with their own metadata through the digital archive's organization, particularly with these metadata visualizations.

28

Studying the Durations of Podcasts

The Term Frequency Line Graph and Associated Keyword Word Cloud can both be effectively applied toward exploratory research and achieving serendipitous discoveries. But we also wanted to use PodcastRE and the "mediaDuration" field to examine a more focused question. What patterns can we notice about the duration of podcasts, and what can they tell us about practitioner norms and assumptions of what makes for a good length of a podcast? Unlike most other AV forms — movies, television, and radio — podcasts are a time-based medium that are not constrained by programming schedules (broadcast schedules, movie theater showtimes) and technical limitations (reels of film and tape). If a podcast could run anywhere from a couple of seconds to several hours in length, how do norms and common practices develop around perceived ideas of a podcast's optimal duration? We realized that metadata could help us answer this question.

29

In this section, then, we propose and share two approaches to studying podcast duration. First, we consider how duration analysis might clarify the differences between two programs of the same specification classification, in this case two popular daily programs from *The New York Times* and *NPR*, using data gathered from episodes ranging from

30

the former's launch in early 2017 to an end point of April 2018. Second, we conduct an investigation of a much larger scale, analyzing large rosters of programs to juxtapose duration across networks and genres. Our case studies here are the comedic programs of Earwolf and the comparably more serious fare of Gimlet Media, using data gathered from episodes ranging from 2009 until early 2018. In both of these cases, the statistics were gathered by first running an SQL query on the PodcastRE database, then exporting metadata for all of the episodes into a .csv file, and finally finding averages, medians, and other numbers using Microsoft Excel. All of these approaches required us to assemble subsets of data from within the PodcastRE collection (and the .csv files), rather than treating the entire collection as a dataset.^[1] The genre and network categories that we ourselves added to the spreadsheets opened the data up for more meaningful analysis, especially when paired with the duration metadata provided by the RSS feeds.

Our first approach to studying duration explored what has become one of the most popular contemporary podcast formats: the daily news program. How long should a daily news podcast take to consume? When *The New York Times* launched *The Daily* in February of 2017, host Michael Barbaro described the fledgling program thusly: "This is how the news should sound. Fifteen minutes a day, five days a week. It isn't quite a podcast — although you can listen wherever you listen to podcasts. It isn't quite the radio — although the mechanics are largely the same. It isn't quite the newspaper — although we'll be drawing heavily on the journalism that powers The New York Times" [Barbaro 2017]. Though Barbaro pegged the program as difficult to explain, it was a nearly immediate hit, gaining over five million monthly listeners by July of 2018 [Jerde 2018]. As Barbaro told *Vanity Fair* that same month, "When we started the show, we had many goals... We didn't realize we were going to make money that was actually going to get pumped back into the company" [Pompeo 2018]. Yet as is often the case, success breeds imitators and competitors, and *The Daily* witnessed the rise of its biggest challenger in June of 2017 when NPR launched *Up First*, a daily "10-minute morning news podcast" that is "designed with digital listeners in mind but will also serve as a preview of the news stories that will be treated in depth on public radio stations across the country throughout the day" [NPR 2017]. That program was also a swift triumph, and as of October 2018, both *The Daily* and *Up First* sat comfortably in the top five most popular podcasts according to Podtrac's rankings: the former tailed behind only *Serial*, while the latter occupies the fifth spot [Podtrac 2018].

31

Episode duration has been a central selling point for each of the two podcasts. As seen above, both of the series' launch press releases mention episode length. Descriptions of the programs on their official websites also focus on duration. *Up First* has remained consistent in its advertised average runtime: "NPR's *Up First* is the news you need to start your day. The biggest stories and ideas — from politics to pop culture — in 10 minutes" [NPR 2018]. *The Daily*, meanwhile, has added five minutes to its initial announcement: "This is how the news should sound. Twenty minutes a day, five days a week, hosted by Michael Barbaro and powered by New York Times journalism" [New York Times 2018]. The programs are, essentially, two different approaches to the morning commute: *Up First*'s proposed shorter length seems guaranteed to slot into almost any daily trip to work, while *The Daily*'s longer runtime requires either a lengthy commute, multiple listening sessions, or even perhaps the utilization of 1.5x or 2x speed playback options. Such duration decisions are complimented by storytelling approaches: *Up First*'s short length is matched with a "greatest hits" style compilation of short stories, while *The Daily*'s relatively lengthier duration is primarily spent on the discussion of a single story. In theory, then, the former aims to quickly provide its listeners with headline-style blurbs about the day's biggest stories, while the latter seeks to exhaustively cover a single topic.

32

Such temporal differences are roughly borne out by the metadata found in PodcastRE's database, although the story is more complicated than the descriptions of the series imply. *The Daily* (mean duration of 22:51, median duration of 22:12) runs nearly ten minutes longer than *Up First* (mean duration of 13:33, median duration of 13:17), with both programs on average running a few minutes longer than their advertised lengths. The differences between the two series is much starker when considering the range in podcast durations, as *Up First* is relatively consistent in episode duration (shortest episode of 11:01 and longest episode of 17:46 for a range of 6:45) while *The Daily* varies widely between episodes (shortest episode of 13:00 and longest episode of 41:23 for a range of 28:23). These durational differences align neatly with the programs' content choices (i.e. multiple headlines vs. single story focus), though they provide critical additional clarifications. While both *Up First* and *The Daily* release episodes each weekday morning, the former's tight range and shorter length ties it more closely to its proposed function as morning commute listening, while

33

the latter’s wider range and extended runtime emphasizes delivering a full story adequately. Since podcasts do not have the same durational constraints of broadcast media, these choices in runtime are clear aesthetic and storytelling decisions – yet given the evolutionary radio approach of NPR’s daily podcast and *The New York Times*’s commitment to the news story, these decisions are not completely detached from their companies’ original mediums.

On a larger scale, podcast duration analysis can point towards divergent approaches by podcast networks and in certain genres. As an example, we conducted an analysis of thirty Earwolf programs^[2] and nineteen Gimlet Media programs.^[3] The former describes itself as “the leading comedy podcast network devoted to creating the best, funniest, and most entertaining podcast shows in existence” [Earwolf 2018]. Gimlet Media specializes in more “serious” fare, characterizing itself as “the award-winning narrative podcasting company that aims to help listeners better understand the world and each other” [Gimlet 2018]. Though both companies employ personnel who have worked or continue to work in other mediums, Earwolf and Gimlet distinguish themselves from other major podcast networks such as *NPR*, *iHeartRadio*, and *WNYC Studios* through their podcast nativism: both companies began as strictly podcast-focused networks rather than emerging within older media companies.

Perhaps as a result, the two networks have markedly different approaches towards podcast episode length. Of the 30 surveyed Earwolf programs, 3 have average runtimes between 0-20 minutes, 4 have average runtimes between 20-40 minutes, 6 have average runtimes between 40-60 minutes, 10 have average runtimes between 60-80 minutes, 3 have average runtimes between 80-100 minutes, 3 have average runtimes between 100-120 minutes, and 1 has an average runtime between 120-140 minutes (See Figure 4). This means that over half of the surveyed programs have average episode durations over an hour, with programs ranging from *Eardrop*’s average runtime of 3:17 and *Never Not Funny*’s average runtime of 2:05:33. The shortest single episode was a 38-second *Eardrop* episode, while the longest individual episode was a *Comedy Bang! Bang!* that lasted 3:19:02. Earwolf’s individual shows also frequently feature drastic ranges in shortest and longest duration: *Hollywood Handbook*, for instance, has a range of 1:53:27 between its shortest and longest episodes, while *Comedy Bang! Bang!*’s range is 2:39:06.

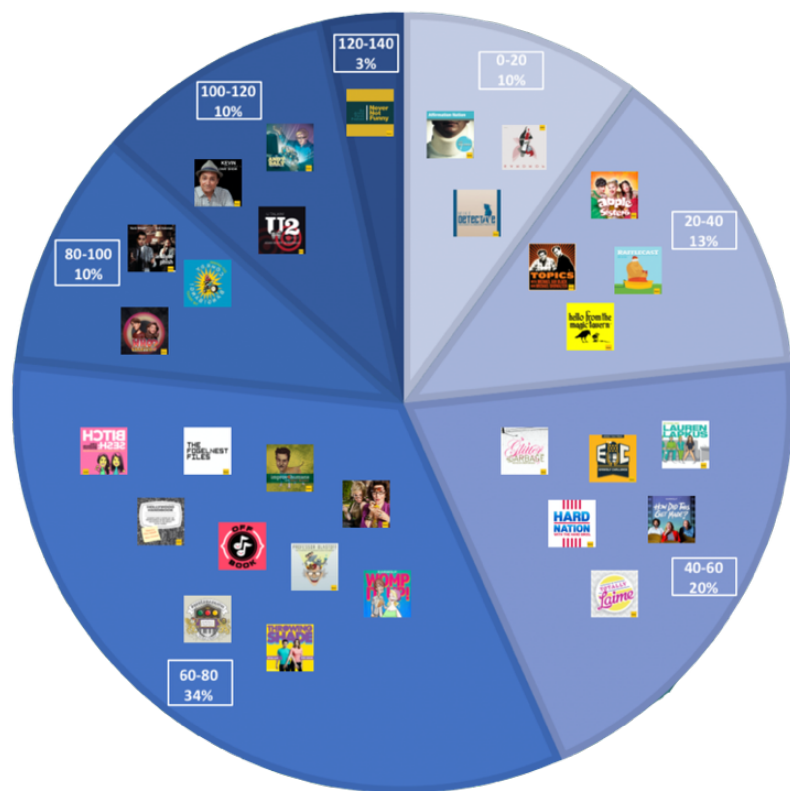


Figure 4. Average durations of 30 surveyed Earwolf programs.

Gimlet, on the other hand, is much more consistent in its runtimes across series, though there is still variation between

individual episodes. Of the 19 surveyed programs, only 1 had an average runtime between 0-20 minutes, and that program (*Chompers*) serves a specific and brief function: children are meant to listen to the series as they brush their teeth. 2 of the series had an average runtime between 40-60 minutes, though both of those shows (*Twice Removed* and *Mystery Show*) are no longer producing episodes. The other 17 series, then, had average run times between 20-40 minutes, aligning Gimlet's roster with conventional advice on podcast episode length.^[4] While individual episodes still varied quite a bit (*Reply All*, for instance, had a range of 1:35:29 between its shortest and longest episodes, while *Mogul* had a range of 1:16:17), these ranges were still much smaller than the largest Earwolf ranges.



Figure 5. Average durations of 19 surveyed Gimlet programs.

The relative homogenization of Gimlet Media podcast duration, then, stands in stark contrast to the diverse podcast lengths of Earwolf. Every Gimlet Media podcast had an average run time under an hour, and 84.2% of the shows surveyed had average runtimes between 20-40 minutes. 56.7% of the Earwolf podcasts analyzed, meanwhile, had average runtimes over an hour long, and 76.7% of Earwolf's podcasts had average runtimes over 40 minutes long — in other words, over three-quarters of Earwolf's shows ran longer on average than Gimlet's "sweet spot." Individual episode lengths varied in each of the networks' programs, but Gimlet's programs featured smaller ranges than the large variation found in many of Earwolf's programs. Such differences may be the result of institutional decisions, generic divergences, or series lengths. Whatever the cause, however, Earwolf and Gimlet serve as evidence that podcast networks can have wildly divergent approaches towards episode duration, and that studying duration can lead us to insights about genre conventions, production values and more.

On a recent episode of the *Start Up* podcast, the show's host, and Gimlet CEO, Alex Blumberg was reflecting on his decision to sell Gimlet media to Spotify. He noted that Gimlet's gambit to standardize the production of highly edited and tightly produced 'quality' podcasts (that often followed very specific duration and other editorial decisions) had turned out to be a financially unfeasible strategy that was losing ground to cheaper and more popular chat cast style podcasts (where duration and other attributes are more flexible given the lower costs involved for editing and polishing the finished piece). His comments are a reminder that, despite the format's substantial growth in the last two decades, there are still many lingering questions about the forms, conventions and economics of podcasting. We believe it is especially crucial during this time of flux, before podcasting stabilizes like so many other media have, to study the different

approaches podcasters of all types are taking as they experiment with this emerging sonic format. Although duration numbers seem like relatively innocuous or descriptive metadata, the research from PodcastRE suggests they reveal historical relationships between new and old media formats, industrial and economic assumptions about “ideal” formats, and generic conventions that shape both amateur and professional podcasts.

Conclusion

Our work on PodcastRE has aimed to provide tools and data that account for podcasting’s complexity as a cultural form while simultaneously taking advantage of its unique technical affordances. The centering of RSS metadata and what can be mined from it through advanced search, graphing keywords over time, or visualizing word clouds of associated keywords has helped us facilitate the automated collection of a significant corpus of podcasts from a crucial period in the format’s emergence. It has also facilitated novel, fine-grained exploration of specific file characteristics as well, like duration metrics, across a variety of genres and shows.

39

The reliance on RSS, however, has also forced us to confront the messiness and intricacies of a born digital object whose metadata and descriptive features are dynamic and podcaster generated. Podcasting’s relatively open and accessible origins have helped create a vibrant environment for web-based audio – one that includes the scores of podcasts available and the multiplicity of voices behind them, but also the numerous apps, aggregator sites and distribution technologies that have emerged to support podcasting’s rise. RSS and XML have not only been important to our work on PodcastRE, but to podcasting more broadly, and to the agency and control it has provided podcasters for defining their work on their own terms as well as for listeners in terms of defining their listening practices. Recently, there have been a number of attempts to move away from the more open and accessible versions of podcasting, to more closed and profitable models (e.g. exclusive shows tied to one platform, like Spotify, or subscription-based services like Luminary). While these options may make podcasting more user-friendly and convenient, or may offer podcasters more options for monetizing their work, they also make podcasts more platform-dependent, less analyzable, and less open to research.

40

The centrality of RSS to both podcasting and PodcastRE has been a theme throughout this article. We believe our methods and findings, however, hold relevance for beyond scholars researching other topics at the intersection of media studies and DH. As this study has shown, metadata records can serve as surrogates for studying large collections of time-based media objects, allowing researchers to query the durations of millions of media objects in a fraction of the time it would take to ingest and analyze transcoded media files. Yet our work has also shown that, when it comes to born digital media, the metadata are never fully separate from the objects they describe, nor are they fully capable of replacing close listening and other media studies methods. There is a need to delineate between methods best suited for *digitized media collections* compared to those most appropriate for *born digital media collections* and for devising strategies to blend AV and DH methods. By making these distinctions, we can better apply DH to AV and identify change and continuity, at a large scale, across media history.

41

Notes

[1] It should be noted that the ability to obtain duration data is not currently available to front-end users of the PodcastRE site.

[2] *Affirmation Nation*, *Analyze Phish*, *Andy Daly Podcast Pilot Project*, *The Apple Sisters*, *Bitch Sesh*, *Comedy Bang! Bang!*, *Eardrop*, *Earwolf Challenge*, *Fogelnest Files*, *Glitter in the Garbage*, *Hard Nation*, *Hello From The Magic Tavern*, *Hollywood Handbook*, *How Did This Get Made*, *Improv4Humans*, *Kevin Pollak’s Chat Show*, *Mike Detective*, *Never Not Funny*, *Off Book*, *Professor Blastoff*, *Rafflecast*, *Ronna and Beverly*, *Spontaneanation*, *Throwing Shade*, *Topics*, *Totally Laime*, *U Talkin’ U2 2 Me*, *Who Charted*, *With Special Guest Lauren Lapkus*, and *Womp It Up!*

[3] *Chompers*, *Crimetown*, *Every Little Thing*, *The Habitat*, *Heavyweight*, *Homecoming*, *Mogul*, *Mystery Show*, *The Nod*, *The Pitch*, *Reply All*, *Sampler*, *Sandra*, *Science Vs.*, *StartUp*, *Surprisingly Awesome*, *Twice Removed*, *Uncivil*, and *Undone*.

[4] Though most blogs on the subject recommend tying duration to whatever length your content demands, they also routinely recommend shorter average durations, with *We Edit Podcasts*, for instance, writing, “it is possible to become successful with a longer show, but in general, the 22 minute rule trumps all” [We Edit Podcasts 2016].

Works Cited

- Adams 2015** Adams, D. "After 'Serial,' Sponsors Pour Money into Podcasts," *The Boston Globe* (2015): <https://www.bostonglobe.com/business/2015/02/13/after-serial-sponsors-pour-money-into-podcasts/OKAzhUWtqCHQbl3luEliBN/story.html>, accessed November 26, 2019.
- Barbaro 2017** Barbaro, M. "Get Ready for The Daily, Your Audio News Report," *The New York Times* (2017): <https://www.nytimes.com/2017/01/30/podcasts/the-daily-get-ready-for-the-daily-your-audio-news-report.html>, accessed November 30, 2018.
- Bergen 2015** Bergen, M. "Google Brings Podcasting to Play Music, Swinging at Apple's Dominance," *Recode* (2015).
- Bray et al. 2009** Bray, T., Hollander, D., Layman, A., Tobin, R., & Thompson, H. S. "Namespaces in XML 1.0 (Third Edition)," W3 (2009): <https://www.w3.org/TR/xml-names/>, accessed February 23, 2019.
- Clement 2016a** Clement, T. E. "Towards a Rationale of Audio-Text" *Digital Humanities Quarterly*, 10.2 (2016).
- Clement 2016b** Clement, T. E. "When Texts of Study Are Audio Files: Digital Tools for Sound Studies in DH." In S. Schreibman, R. Siemens, and J. Unsworth (eds), *A New Companion to Digital Humanities*, Chichester ; Malden, MA: John Wiley & Sons, Ltd., Chichester (2016): 348-57.
- Crowe 2019** Crowe, Anne. "101 Quick & Actionable SEO Tips That Are HUGE." *Search Engine Journal*. October 21, 2017. <https://www.searchenginejournal.com/101-quick-seo-tips/180563/>.
- Earwolf 2018** Earwolf, "About Earwolf," Earwolf (2018): <https://www.earwolf.com/about/>, accessed November 30, 2018.
- Edison 2019** Edison Research, "The Infinite Dial 2019," Edison Research (2019): <https://www.edisonresearch.com/infinite-dial-2019/>.
- Florini 2017** Florini, S. "This Week in Blackness, the George Zimmerman acquittal, and the production of a networked collective identity." *New Media & Society* 19.3 (2017): 439-454.
- Florini 2020** Florini, S. and Barner, B. "'I'm Trying to Be the Rap Oprah': Combat Jack and the History of the Loudspeaker Network." In J.W. Morris and E. Hoyt (eds), *Saving New Sounds: Dispatches from the PodcastRE Project*, University of Michigan Press, Ann Arbor (forthcoming 2020).
- Gimlet 2018** Gimlet Media, "About," Gimlet Media (2018): <https://www.gimletmedia.com/about>, accessed November 30, 2018.
- Green and Meissner 2005** Greene, M. and Meissner, D. "More Product, Less Process: Revamping Traditional Archival Processing," *The American Archivist*, 68.2 (2005): 208-63.
- Hansen 2020** Hansen, S. "The Feed is the Thing: How RSS Defined PodcastRE and Why Podcasts May Need to Move On." In J.W. Morris and E. Hoyt (eds), *Saving New Sounds: Dispatches from the PodcastRE Project*, University of Michigan Press, Ann Arbor (forthcoming 2020).
- Hardesty 2019** Hardesty, J. "Bias and Inclusivity in Metadata: Awareness and Approaches". Indiana University Digital Collection Services.
- Hines 1999** Hines, M. "Netscape Broadens Portal Content Strategy," *Newsbytes*. (1999): <http://link.galegroup.com/apps/doc/A54120248/ITOF?u=umuser&sid=ITOF&xid=377f45>.
- Hogan 2008** Hogan, M. "Dykes on Mykes: Podcasting and the Activist Archive." *TOPIA: Canadian Journal of Cultural Studies* 20 (2008): 199-215.
- Hoyt 2016** Hoyt, E., Hughes, K., and Acland, C.R. "A Guide to the Arclight Guidebook." In C.R. Acland and E. Hoyt (eds), *The Arclight Guidebook to Media History and the Digital Humanities*, REFRAME/Project, Falmer (2016): pp. 1-29.
- Jerde 2018** Jerde, S. "How NYT's The Daily Grew to 5 Million Monthly Listeners and Became a Breakout Star," *Ad Week* (2018): <https://www.adweek.com/digital/how-nyts-the-daily-grew-to-5-million-monthly-listeners-and-became-a-breakout-star/>, accessed November 30, 2018.
- Jockers 2013** Jockers, M. L. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, Champaign (2013).
- Morris 2019** Morris, J. W., Hansen, S., & Hoyt, E. "The PodcastRE Project: Curating and Preserving Podcasts (and Their Data)" *Journal of Radio & Audio Media*, 26.1 (2019).

NPR 2017 "Up First: The Essential Morning News Podcast From NPR," *NPR* (2017): <https://www.npr.org/about-npr/522211062/up-first-the-essential-morning-news-podcast-from-npr>, accessed November 30, 2018.

NPR 2018 "Up First," *NPR* (2018): <https://www.npr.org/podcasts/510318/up-first>, accessed November 30, 2018.

New York Times 2018 "The Daily," *The New York Times* (2018): <https://www.nytimes.com/column/the-daily>, accessed November 30, 2018.

PodNews 2019 PodNews. "The Total Number of Available Podcasts Is Now 700,000," PodNews (2019): <https://podnews.net/update/700000>.

Podcast 2015 . "10 SEO Tips For Your Podcast." Podcast Motor. September 22, 2015. <https://www.podcastmotor.com/seo-tips-podcast/>.

Podcast 2020 Podcast Insights. "2020 Podcast Stats & Facts (New Research From Apr 2020)," Podcast Insights: <https://www.podcastinsights.com/podcast-statistics/>

Podtrac 2018 Podtrac, "Podcast Industry Audience Rankings," Podtrac (2018): <http://analytics.podtrac.com/industry-rankings/>, accessed November 30, 2018.

Pompeo 2018 Pompeo, J. "'We Didn't Expect to Make Money': How *The Daily's* Michael Barbaro Unexpectedly Became the Ira Glass of *The New York Times*," *Vanity Fair* (2018): <https://www.vanityfair.com/news/2018/07/how-the-daily-michael-barbaro-became-the-ira-glass-of-new-york-times>, accessed November 30, 2018.

RSS Advisory Board 2019 RSS Advisory Board. "RSS History." RSS Board (n.d.): <http://www.rssboard.org/rss-history>, accessed February 23, 2019.

Recode.net 2015 Recode.net, <https://www.recode.net/2015/10/27/11620066/google-brings-podcasting-to-playmusic-swinging-at-apples-dominance>, accessed February 23, 2019.

Underwood 2019 Underwood, T. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, Chicago (2019).

Wang 2020 Wang, J. H. "The Perils of Ladycasting: Podcasting, Gender, and Alternative Production Cultures." In J.W. Morris and E. Hoyt (eds), *Saving New Sounds: Dispatches from the PodcastRE Project*, University of Michigan Press, Ann Arbor (forthcoming 2020).

We Edit Podcasts 2016 We Edit Podcasts, "What Is the Optimal Length for a Podcast?" We Edit Podcasts (2016): <https://www.weeditpodcasts.com/what-is-the-optimal-length-for-a-podcast/n>, accessed November 30, 2018.

Winer 2001 Winer, D. "Payloads for RSS." (2001): <https://web.archive.org/web/20080214205403/http://www.thetwowayweb.com/payloadsforrss>, accessed February 23, 2019.

Winer 2003 Winer, D. "RSS 2.0 Specification," RSS 2.0 at Harvard Law (2015): <https://cyber.harvard.edu/rss/rss.html>, accessed February 23, 2019.