<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of Report</strong></td>
<td>Final White Paper</td>
</tr>
<tr>
<td><strong>Grant Number</strong></td>
<td>HAA-255979-17</td>
</tr>
<tr>
<td><strong>Title of Project</strong></td>
<td>Investigating the Golden Age of Podcasting through Metadata and Sound</td>
</tr>
</tbody>
</table>
| **Name of Project Directors** | Dr. Jeremy Morris, Communication Arts, UW-Madison  
Dr. Eric Hoyt, Communication Arts, UW-Madison |
| **Name of Grantee Institution** | University of Wisconsin-Madison |
| **Report Submission Date** | Nov. 1, 2019 |
The Problem of Podcasting

Podcasting is only roughly 15 years old as a media form and practice, but it has already ushered in an explosion of amateur and professional cultural production. There are now over 525,000 podcast feeds and close to 18.5 million individual episodes in over 100 languages (Locker, 2018). There’s a podcast on almost every subject imaginable, from popular shows like Serial, WTF with Marc Maron, and Radiolab to lighter fare like the wrestling podcast Wrestlespective or shows that cover social issues like sexuality, identity, race, or politics, (e.g., Strange Fruit, 2 Dope Queens, Pod Save America, etc.). This diversity of new audio is part of what makes podcasts significant as a sonic, cultural form: podcasting offers both the potential to bring new voices into the mediascapes of everyday users and the possibility for sound workers to learn new skills and techniques for expressing and sharing ideas. This boom of content and soundwork has prompted many critics to claim we are in the midst of a “Golden Age of Podcasts” (Blattberg, 2014; Roose, 2014; Sillesen, 2014); a moment where the choice for quality digital audio abounds, and where new voices and new listeners connect daily through earbuds, car stereos, or office computers. In 2018, an estimated 73 million Americans will listen to podcasts every month, and the average weekly listener spends 6 hours and 37 minutes listening (Edison Research, 2018). Apple, the leading distributor of podcasts, recently passed 50 billion all-time podcast streams and downloads, with close to 30 billion of those coming in 2018 alone (Locker, 2018).

These statistics matter industrially, of course, but they also matter culturally. Podcasts represent a challenge or alternate approach to the corporate broadcasting model that governed much of 20th century radio, but they also represent a continuation of historical radio practices, aesthetics and institutions. They are a vital new space of experimentation for audio production—what Hilmes (2013) calls “soundwork”—and thus merit recognition from Humanities scholars as a significant new medium of cultural and artistic expression. Podcasting fosters and builds intense and intimate yet remote relationships between and among its listeners and hosts. Like the connections forged between 19th century novels and their readers, early cinema and its moviegoers, and 1930s radio stars and their fans, podcasting connects users to an imagined sense of community and public life. The key difference though is that podcasting’s entry barriers for producers are even lower, and we have the benefit of studying this remarkable moment as it plays out—the equivalent of a film scholar visiting a nickelodeon theatre in 1907. Podcasting also offers media scholars lessons about emerging industries, particularly ones working to figure out the economic and cultural models that will support the circulation of largely “free” content. Yet despite the lessons podcasting holds for scholars and the popular excitement this vital media form has generated, the sounds of podcasting’s nascent history remain mystifyingly difficult to analyze. There are few resources for anyone interested in researching the form, content, or history of podcasts and even fewer tools for preserving and analyzing the sonic artifacts being produced during this “Golden Age” of audio. What today’s podcasters are producing will have value in the future, not just for its content, but for what it tells us about audio’s longer history, about who has the right to communicate, and by what means (Sterne, Morris, Baker, & Freire, 2008). We may be in a “Golden Age” of podcasts but if we’re not making efforts to preserve and analyze these resources now, we’ll find ourselves in the same conundrum many radio, film or television historians find themselves: writing, researching, and thinking about a past they can’t fully see or hear.
To assist in efforts for saving and studying contemporary sound culture, we are developing PodcastRE (short for Podcasting Research)—a digital archive of podcasts and a suite of tools for searching and studying them. In the four years since the project’s inception in 2015, we are happy to report that we now have a robust and sustainable database/archive that currently preserves over 2.2 million audio files from close to 15,000 RSS feeds (and almost 80 terabytes of data). The site provides new, and previously unrealized, possibilities for researchers to visualize and analyze trends in keywords, through timeline graphs and metadata word clouds. We’ve also conducted preliminary research on emerging software and technologies for performing sonic analysis on some of the audio files in the database. The site and its tools have allowed us to create a host of publications from this project, including several published articles in media and cultural studies journals and edited collections, as well as a forthcoming journal article in a digital humanities journal and an entire edited collection dedicated to the PodcastRE project and the challenges and opportunities presented by studying podcasts from a media historical perspective.

In building the PodcastRE database, we had to confront the significant challenges inherent in podcast preservation. We had to reflect on what the opportunities and issues tell us about contemporary podcasting, since we knew that the solutions that emerge for collecting, preserving, and analyzing born digital objects will also offer insight for scholars digitizing older media objects. Inconsistent metadata, for example, present an obstacle for creating a streamlined search engine, with standardized fields and faceting capabilities. When the same metadata are viewed in a different light, however, they can be incredibly revealing about the different approaches between professional and amateur podcast producers, both of whom contribute to the broader media ecosystem. Therefore, we found ourselves in the position of trying to develop a robust, functional, and user-friendly infrastructure while simultaneously attending to the noise, texture, and messiness of digital cultural production. And while the need to balance these different goals is hardly unique in the digital humanities, the forms with which we are working (hundreds of thousands of digital sound files) provided challenges that differed from text and literary-based DH projects (Drucker, 2012; Ramsay & Rockwell, 2012).

First, we faced challenges in terms of collection and selection. How do we decide which podcasts to save? How do we develop tools to capture both the most popular and pervasive shows as well as the more margin-alized and less immediately discoverable ones? Given how mediated podcasts are by the platforms that serve them, we find that different strategies are required for approaching the variety of shows podcasting offers. Next, we had challenges with the curation of data and metadata. The curatorial decision to save a podcasting feed results in a multi-step data preservation process that saves the original data and metadata files while transforming the data in other ways. Finally, we were forced to reflect on what it meant to “save” a podcast. Are podcasts just the audio files and RSS feeds? Or does saving podcasts also mean saving the companion websites, session files, and other materials that audio production generates? We consider how the very ways we define and limit the boundaries of the term “podcast” shape the future directions for the database and research potential for scholars and users looking to study and save the new sounds podcasting produces.

While we believe we have made progress in answering these questions and addressing these challenges, much work remains to be done in terms of developing sound preservation practices for the collection, analysis, and archiving of podcasting’s emerging sonic cultures.
The rest of this white paper details the activities and outputs of the grant-funded project. For further discussion of the above challenges, please see the *Journal of Radio and Audio Media* article (Morris, Hansen and Hoyt 2019) as well as the forthcoming edited collection *Saving New Sounds: Dispatches from the PodcastRE Project*.

**ACTIVITIES:**

1) **DEVELOPED A ROBUST, USEABLE, PUBLIC DATABASE:** During the grant period, we were able to completely overhaul the project’s main output, the website (http://podcastRE.org) and its accompanying database. Thanks to the work of our lead computer specialist and our graduate research assistant (funded partly through NEH funds and through an additional internal UW-Madison UW2020 grant), the site’s capacity has increased significantly. Before the grant, the database had approximately 200,000 audio files from 1,000 RSS feeds. We now house 2,207,759 audio files from 14,918 RSS feeds. In terms of data, the collection has grown from 6 terabytes (TB) at the start of the grant cycle to close to 80 TB. We have a robust RAID system to support this infrastructure, that is capable of storing up to 200 TB and expanding much further though the acquisition of more drives. The entire system is regularly backing up to an LTO8 tape library, providing long-term preservation. The robust back end supports a significantly-improved advanced search function and powers two completely new visualization features that were developed by the NEH-funded graduate research assistant. The associated keyword cloud feature and the term frequency line graph (both explained further below) provide new research possibilities for navigating the database’s contents. The site is now, as far as we know, the largest publicly-oriented research collection of podcasts of its kind.

2) **BUILT METADATA ANALYTICS AND VISUALIZATION TOOLS:** The NEH grant was specifically focused on developing new analytic and visualization tools for the site. After building an Apache Solr search index, we developed and launched three key tools as part of our metadata analytics platform at http://podcastre.org/analytics. This is the only tool of its type for searching and researching podcasts and these tools are already being used to produce new research publications (see publications below). Specifically, we developed the Advanced Search (http://podcastre.org/advancedSearch), the Associated Keyword Cloud (http://podcastre.org/wordCloud), and the Term Frequency Line Graph (http://podcastre.org/lineGraph) tools. The Advanced Search tool allows users the ability to filter results by date, by category, by keyword, by description, and by podcast source. The faceted search ability allows users to further hone any search, making it more specific and useful for a user’s needs. The Term Frequency Line Graph lets users graph the frequency of any queried term within the database over time. This includes words found in titles, descriptions, and keywords over a specific date range (i.e. rise in instances of the phrase #MeToo), either by year (e.g. from 2014-2018) or by month (Jun-Oct., 2019). The graph can be visualized as a line graph or area chart, and each node on the graph also brings the user to the list of podcasts that include the queried term. The Associated Keywords Cloud allows users to find which words are most often grouped together in relation to a particular query. The larger the word within the Word Cloud, the more often these terms are grouped together. As with the frequency graph, clicking on a node (in this case a word in the cloud) will allow you to access the database and display the list of podcasts that refer to both the keyword that you have queried and the chosen associated keyword. Both these visualizations offer a macro level view of the database and its contents, and the Keyword Cloud offers a novel approach to keyword searches, by allowing users to recognize the context around their particular area of research/keyword. We also
developed help documentation for each of these features, and the site now provides both in-text instructions and video tutorials for users.

3) **TESTED SONIC ANALYSIS TOOLS:** The other major effort of our project was to test and use new open source tools to explore what kinds of arguments might be made from an analysis of the sonic components of sound files (rather than their visual representations). Through our collaborations with another NEH-funded project, *Tools for Analyzing Text-in-Performance*, we became beta-testers of the programs *Gentle* and *Drift*. The former allows for the creation of aligned transcripts, with details about pacing, words per minute, and similar features, while the latter allows for a sophisticated analysis of pitch and frequency of the voices in a given audio file. Along with a team of graduate students (some supported through NEH funds, others through UW2020 funds), spent the second year of the project testing these programs, applying them to particular subsets and collections of podcasts in our archive and conducting viability tests for incorporating the tools into our website. We have not yet found a way to integrate these tools into the database in a way that would be beneficial to all of PodcastRE’s users; at this point, they still require a fairly specialized level of knowledge which makes it harder to translate and create useable public tools, as we were able to do with the visual analytic tools. However, the experiments did allow us to create new publications with novel modes of sonic analysis, contributing new methods for the study of podcasts.

4) **PUBLIC PRESERVATION TOOLS:** The site is now a highly useful and public search and research tool. We have also implemented a “submit a podcast” feature that allows public users to add RSS feeds to our database for podcasts they wish to study and preserve. This feature (which is accessible at [http://podcastre.org/feedSubmission](http://podcastre.org/feedSubmission)) vastly increases the reach of the project and allows for more community involvement. We are also still developing user roles for the website, which will allow us to assign specific permissions to certain users (i.e. researchers) who might need greater access to preserved shows than average users.

5) **PUBLICIZING OUR ACTIVITIES:** Finally, we spent a good portion of the grant period publicizing our activities through numerous publications, conference talks and panels, and through guest lectures in appropriate classes and venues. We conducted several interviews for local and national media outlets, and we have produced two episodes of a podcast that documents the process of building the database (with one or two more episodes still to be released). We also communicated new feature updates via sound and media listservs, Twitter and Facebook. We summarize the publications and other outputs in the “Award Products” section, but we wanted to make note of a special publication that exceeds anything we promised in our original proposal. Given the excitement around the site and the new potentials it opens up, we recognized the need in the field for an edited collection that shared our research, along with contributions from leading scholars of sound and media studies and from UW-Madison graduate students who worked on the project. We are pleased to report that our book, *Saving New Sounds: Dispatches from the PodcastRE Project*, is under contract with the University of Michigan Press and will be completed in 2020. It will be published on an open access basis and freely available online, extending the reach of this project even further.

In summary, we believe we have met and surpassed the project’s original goals. We made judicious use of NEH funding (and returned $10.48!) and have created a useful and valuable resource. There were no major changes or omissions in project activities, and no major changes
in key project personnel. The co-PIs, Jeremy Morris and Eric Hoyt, were part of the project for the duration, as was the key computer specialist, Peter Sengstock. Sam Hansen was our lead Project Assistant, who developed the search index and site infrastructure, and Susan Noh was the lead Visualization and Analytics project assistant. We also received significant assistance from JJ Bersch and Jacob Mertens who created visualization tool tutorials and tested the sonic analysis tools, respectively. We will continue to stay active in the preservation of PodcastRE’s data. As Trevor Owens reflects in *The Theory and Craft of Digital Preservation*, “Nothing has been preserved, there are only things being preserved…. The work is never finished.”

**ACCOMPLISHMENTS:**
Beyond the site itself (http://podcastre.org), and the new search, visualization and analytics tools it provides (http://podcastre.org/analytics), our primary accomplishments have been in the realm of conference presentations and publications documenting the database building process, as well as the exploring some of the possibilities that research with (and on) podcasts provides. We list those publications, presentations, and media appearances in full in the “Award Products Section”. For now, we will simply note that our original proposal promised, in addition to the analytics platform, the following output: 2 conference presentations (1 at a media studies conference, 1 at a digital humanities conference) and 2 publications (again, 1 in a media studies journal, 1 in a digital humanities journal) and a podcast documenting our process. We have surpassed these objectives significantly, and we are happy to claim we have created:

1) an open-access edited collection featuring over 17 scholars in the field discussing podcasting’s pasts, present and futures and the potentials of tools like PodcastRE.org
2) Over a dozen chapters and journal articles across media studies and digital humanities publications
3) Almost a dozen conference presentations and keynote lectures on the database and the challenges of podcasting preservation and research
4) Numerous media appearances
5) 2 episodes of our 3-part podcast series
6) A digital humanities crowd-sourced award
7) A number of related external grants.

**AUDIENCES**
The site continues to be available and open to both UW community members and the general public. It is the largest publicly oriented research database of its kind and can be used by anyone looking to study or explore the contemporary and historical podcasting ecosystem. We provide metadata and tools for searching through millions of podcasts, and we have visual analytic tools with tutorials for users of all kinds. We are still building the researcher dashboard for the site, which will allow researchers at UW-Madison and other such institutions to have even greater access to the collection for research purposes, and based on the significant interest we’ve received from emails from professors wanting to teach and research with the database, we believe this functionality will significantly increase the site’s usefulness for its community. The use of the database continues to grow as we publicize it through our networks, at conferences, and through more public media. We have received emails from professors who use the database in their classes and teaching and, lead-PI Morris has been asked to guest lecture in classes at other universities where students are using the database. The database was selected as a “user case” for another NEH-funded project, the *Tools for Analyzing Text-in-*
Performance. Our project is also part of a much larger audio preservation project run out of the Library of Congress, The Radio Preservation Taskforce. Recently, we have also been collaborating with The Metropolitan New York Library Council (METRO)’s $142,000 AW Mellon Foundation grant “Preserve this Podcast: A Podcast Tutorial and Outreach Project.” The project is designed as an education and awareness campaign aimed at amateur and professional podcasters to promote affordable, easy-to-implement archival techniques for digital audio preservation. When they host and run their workshops, the Preserve This Podcast Team notes to individual podcasters the utility of submitting their shows to the PodcastRE.org database as a preservation solution.

CHALLENGES
We are thrilled with the new analytics and visualization interface, and the research possibilities it provides. Our biggest challenge in developing them was working through how to deal with the inconsistent metadata that exists in the podcasting ecosystem, which thus made it difficult to accurately graph or create a WordCloud for some of the datasets. These metadata inconsistencies have also posed other challenges: sometimes the erratic categories, tags and other data that users encode in their RSS feeds can disrupt our ingestion script, causing a lag in our ability to import new podcasts. While the problem is relatively easy to resolve, as the database grows these challenges increase. We also realize that we have some basic user statistics but, given that our focus was on building the site’s infrastructure, search index and visualization tools, we did not invest in comprehensive audience measurement tools. We have largely anecdotal stories around use of the database so in moving forward, we will be looking at different ways to quantify and measure the impact of the project’s reach, especially given the success we know the project has had for audiences we’ve been in contact with. Another key challenge with this project has been the long-term preservation of the data the project is generating (as discussed further below). We had hoped an institutional solution might be available to us for storage of these wonderful resources though our campus-wide digital collections solution is still in progress. That said, we believe our project has served as an excellent test case for our campus’ efforts to strengthen its digital collection capabilities; our project has been at the forefront as an example of the kinds of storage and technical needs that researchers from many disciplines are having as they build their own digital collections (and seek support from the university for storing, preserving and managing these collections).

CONTINUATION
The project will continue indefinitely. The robust storage system we have built, along with the backend information model we have developed to support the database, means that the site will continue working with minor maintenance to be provided by the lead-PI and the lead computer specialist. Although our 200TB of storage space will undoubtedly run out as the database grows, our system’s capacity for growth is significant with the addition of new drives and storage devices. We had hoped to partner with our campus libraries for long-term preservation of the database and its contents, but the current lack of a campus-wide digital collection solution has meant that our best solution for long term preservation remains the local backups and archive we have created in house in the department of Communication Arts (with multiple redundancies in multiple locations). The grant funding allowed us to significantly boost our infrastructure and capacity, and combined with UW2020 funding and external grants, co-PI Eric Hoyt has been able to create what we’ve come to call the AV Data Core (http://avdata.commarts.wisc.edu/).
This core provides the infrastructure for our project as well as for the Media History Digital Library and the Wisconsin Center for Film and Theater Research’s digital holdings. It has already helped us secure new grant funding for other digital projects and it has been instrumental in helping us back up and preserve PodcastRE’s current holdings. The AV Data Core will continue to serve in this capacity until we migrate to larger digital collection solution that emerges from the UW’s central library system (or a more systematic/automated shared solution with WCFT, like Archivematica). That said, as we start to develop more useful and useable user statistics and analytics, we will use these to apply for further grant funding, such as a further internal UW-Madison grants, or external grants, like an NEH Digital Humanities Advancement grant (Level III) or an ACLS Digital Extension Grant.

FURTHER READING

Edited Collection

Articles and Chapters


Conference presentations, performances and exhibitions


Media Appearances:


Appearance on WPR’s Central Time radio broadcast, with Rob Ferrett. 31 May 2017

Appearance on NPR’s The Big Listen Podcast, with Lauren Ober. 15 Jun. 2017

Appearance on local podcast The Natecast with Nate Chapelle. 19 Jun. 2017
Project highlighted in field-specific podcast Aca-Media, with Chris Becker and Michael Kackman. 5 Jun. 2017

**Media Products:**
Episode 1 and 2 of “The PodcastRE Project”, two episodes of our 3-part podcast detailing the goals and aims behind the project. Created by our graduate research assistants, Samuel Hansen and Tom Welch. Available on the homepage of http://podcastre.org.


**Awards**
Digital Humanities Award: Highlighting Resources in Digital Humanities. The PodcastRE Database was voted winner of a DH 2018 Award by a community of digital humanities scholars and users. http://dhawards.org/dhawards2018/results/

**Related Grants**
Utilizing the digital infrastructure developed through funding from the NEH and the UW2020, PI Jeremy Wade Morris and Co-PI Eric Hoyt have successfully obtained roughly half a million dollars in external grants. These grants include: