LIBRARY OF CONGRESS

+ + + + +

RADIO PRESERVATION TASK FORCE

+ + + + +

SAVING AMERICA'S RADIO HERITAGE:
RADIO PRESERVATION, ACCESS, AND EDUCATION

+ + + + +

COMMITTEE ON METADATA AND DIGITAL ARCHIVING

+ + + + +

SATURDAY,
FEBRUARY 27, 2016

+ + + + +

The Committee met in Room 0302, Hornbake Library, University of Maryland, College Park, 4130 Campus Drive, College Park, Maryland, at 11:00 a.m., Cynthia Meyers, Chair, presiding.

COMMITTEE MEMBERS:

CYNTHIA MEYERS, PhD, College of Mount St. Vincent, Chair
ANDREW BOTTOMLEY, University of Wisconsin-Madison
RACHEL CURTIS, American Archive of Public Broadcasting
CASEY DAVIS, American Archive of Public Broadcasting
KEN FREEDMAN, WFMU
MARY KIDD, New York Public Radio/WNYC
MARIT MACARTHUR, University of California, Bakersfield
JEREMY MORRIS, University of Wisconsin-Madison
JOHN PASSMORE, New York Public Radio/WNYC

DISCUSSANTS:

JACK BRIGHTON, PBCore
ERIC HOYT, PhD, University of Wisconsin-Madison
STEPHANIE SAPIENZA, Maryland Institute for Technology in the Humanities
WILLIAM VANDEN DRIES, Indiana University
MARK J. WILLIAMS, PhD, Dartmouth College

P-R-O-C-E-E-D-I-N-G-S

(11:05 a.m.)

CHAIR MEYERS: Hello. We don’t seem to have mics up here. Hello? We only have 90 minutes. Actually, now we have 85 minutes. We have an amazing group of people here today to talk to us about metadata. The RPTF, oh, I’m sorry. I’m Cynthia Meyers. I’m a college professor, I’m a radio historian. I’ve used archives.

PARTICIPANT: We cannot hear you.

CHAIR MEYERS: Yes, okay.

PARTICIPANT: I think everybody is going to have to come closer.

CHAIR MEYERS: Okay. I’m Cynthia Meyers. I’m a college professor and a media historian. I use archives, I don’t make them. I’m here to learn a lot today. And I’m also here to try to help coordinate where the RPTF is going to go with the data that we have collected and how we’re going to make it publicly accessible.

So we have a lot of great people here

today to talk to us about how to do that. And we've already learned a lot this morning about how people are doing that.

We have a number of presenters. And I've asked them to present very briefly, like that five minute round table format. And I'm going to ask them to introduce themselves.

And then we have a number of discussants as well. And I sent to the discussants and the presenters a kind of list of questions. And I know that we won't get to all of these questions today.

But just to structure the discussion today, the questions that I'm asking us to think about are what can we learn from other metadata aggregators. There are a number of other places that people are aggregating information about archives and collections.

What information is important for the RPTF to describe and why? We're working on a collection level basis, but really what's important for the public researchers and educators to know about those collections, what formats, PBCore probably?

And then how should we handle different institutions' records? We are trying to interact with a lot of different institutions with a lot of different concerns and issues. Are we going to be reformatting information? Are we just going to be taking what they have and using it?

And then there are the interface issues. When we build this site, this portal, you know, really what are we trying to design for public access? How are we going to try to connect with a general public and not just fellow researchers and archivists?

And then finally, who's going to be doing this? That's part of what we need to start sorting out. And then what are our priorities?

So that's the general structure. And I know that we won't be able to get to all of it. But I'd like to start now by asking Jeremy and Andrew to tell us a little bit about what they're doing with podcast archiving.

MR. MORRIS: Sure. Hi, I'm Jeremy Morris. And this is Andrew Bottomley. I'm going to try and talk in an abnormally loud voice. Does that work?

PARTICIPANT: Yes.

MR. MORRIS: Okay. Yes. I felt a bit weird, I mean, we're talking about preserving old radio, and here I am talking about podcasts. But in Sam's talk yesterday he mentioned just how vulnerable that format is, even though we think it's something that's going to be around all the time.

So started just as an idea of wanting to save some podcasts as an iTunes database on my computer. And then we tried to really think about how we could start collecting these in a kind of massive way.

This is, over here on this screen, it's called the Audio Culture Archive Database. Because I'm not entirely sure what to call it just yet. But we have about 75,000 audio files now with close to

90,000 metadata records on over 907 podcast feeds.

So we’re pulling metadata mostly from the XML files. And it’s allowed us to create a very basic search interface here. I’ll try and show just a little bit of it. We have kind of, every time new podcast feeds get added, they get put into a kind of graphical interface here. And there are pages upon pages of that.

You can also search through different basic, basic categories right now, description, author, title. This is really the work of an RA who spent, you know, 30 hours coding something for me this summer. And I basically need to just pay a lot more people to do a lot more work on it if I want to do stuff.

Andrew and I have been kind of curating which feeds go into it. But really, we’re just trying to make a huge collection at the moment.

It’s also Version 1 of this. I’ve kind of applied for a grant to allow for more, not only more podcast to be stored but a kind of more robust back end. What we’re looking at doing right now at the University of Wisconsin-Madison which is where I’m based, is integrating this with the library, the campus library’s infrastructure.

Because obviously they have sort of a lot more expertise than our department does. This is really just being housed kind of on a server in our department at the moment.

So we’re trying to find a way to integrate it with the library’s collection so that it can be kind of a bit more of an error-free process.

Right now it’s through open source software called gPodder which goes out and collects things every time we tell it to. It’s basically, I think, scheduled on a weekly basis bringing in new podcasts. What we’d like is just to make sure those feeds are always, and predictably, and reliably bringing in podcasts.

Yes, I didn’t know how much time to spend doing it. I know we’re sort of quick for time. So just to let you know, there is this initiative going on. We’re trying to bring in a whole bunch of podcasts from both, like, professionally produced podcasts to the, you know, folks in basements who are recording things.

And I guess we wanted to end the session, or at least our part of it, by just asking a few questions about, you know, things that we’ve run into, challenges we’ve run into in the process and how to think about how that aligns with some of the stuff that RPTF is doing.

MR. BOTTOMLEY: So, you know, the podcasts are, you know, deep, vulnerable and ephemeral. Even in the short time that we’ve been doing this, we’ve run across many expired web links, podcast feeds often end abruptly, cease to be curated or also, as we see podcasting professionalizing, things getting locked behind proprietary systems.

And so as more commercial attention turns towards podcasts and industry monetizes more, you know, there’s this question of is this as accessible as it may seem, right? And is it going to become less and less accessible moving forward?

Also, the objects themselves can be quite unstable. Objects that are in the database are changing or change over time. One thing is dynamic advertising which allows producers to change the ads or add ads to old podcasts and even potentially customize to the individual listener. So, you know, how do we see a content that is less like a recorded radio show and maybe more like a blog. What's the original, right, is the question here. I'm Jeremy, you want to --

MR. MORRIS: Yes. The case of Serial is an interesting one. So the database is structured so that we store the individual files on our server. But what is available on the Web is now just basically to us, because it's not public yet.

But when it is more public, it would go to the reference file that's online. So if you play the Serial file right now for the first episode of the first season, you'll hear what's online now which has an audible ad at the beginning.

And if you listen to Serial, you know how integral that, like, MailChimp ad was, right. So our database version has the MailChimp version. I’m not sure if that’s the same as the Glass Records that the guy had in the back of his trunk, you know, but I have the Serial with the MailChimp ad actually in it, right. But these objects are kind of changing.

MR. BOTTOMLEY: And this is important for, you know, cultural historians in broadcasting, right. It’s very important when you can get original radio broadcasts with the ads or with all of the sort of the paratextual material around just the program, right, to get a sense of how it was actually heard historically in context at the time of release.

And that’s a very fluid thing with podcasting. So that is, I think, an important question.

You know, and then there’s the question

of no standard template for metadata capture. You know, iTunes has its own preconditions as do other aggregators. But podcasters from amateurs to organizations often don’t fill it out completely.

So can we rely on just the information being provided by the producers or as people who are trying to make it searchable and accessible to researchers? You know, what is the metadata that we need to be adding to it?

MR. MORRIS: I think we have some more, but I want to make sure we get enough room for everybody else. The end question I think I would leave with is just, as researchers, thinking about using a database like this, what kinds of things you’d want to see in a database full of podcasts that you could be asking questions about for doing research on it. Thanks.

CHAIR MEYERS: Okay, thank you, Jeremy and Andrew. I’d like to move to Rachel and Casey, if you could introduce yourselves and speak briefly about what you’ve been up to.

MS. DAVIS: Hello, I’m Casey Davis. I’m the project manager for the American Archives of Public Broadcasting at WGBH. And speak up?

So I’m going to talk a little bit about WGBH’s side of the project. And then Rachel will talk about the Library’s side of the project.

So as you probably are aware, now that you’ve probably heard a couple of American Archives presentations over the past couple of days, the AAPB is a collaboration between WGBH and the Library of Congress to preserve and make accessible historic public radio and television content created by stations across the country.

And it started with an inventory project in 2011 funded by the Corporation for Public Broadcasting. And WGBH managed that project to work with 120 stations to aggregate inventory records.

So these stations went through their closets and created metadata records about every single tape that they had in their closets or stored under, you know, their desks.

And we worked with them by -- most of these stations don’t have a professional archivist or librarian on staff. So we really had to hand-hold them in gathering this information.

So we created templates and CSV files so that they would know what fields we wanted them to capture. We also asked them to follow specific controlled vocabularies.

We used PBCore as our data model. I’m not sure if you’re familiar with PBCore, but it’s a metadata schema funded initially by the Corporation for Public Broadcasting in 2004. It’s been through several iterations of further development. We just released PBCore 2.1. And we just launched a new website at PBCore.org.

The metadata standard includes descriptive technical, intellectual property and intellectual property metadata. And for the American Archive we’re using that in conjunction with PREMIS for our preservation metadata. But

Rachel can tell you more about that.

So after the inventory project, CPB funded the digitization of 40,000 hours. And in addition to the digitization, they funded the development of what we call the Archival Management System. And it’s an open source system developed by AVPreserve, developed specifically for this project.

And it’s developed for stations to login and access their own inventory records and view their proxy files as they’re being digitized. It is a PHP application with a MySQL database based on PBCore. It allows for import and export of PBCore XML and CSVs that are PBCore compliant.

We are continuing to enhance the metadata now that we have digitized the 40,000 hours. We’ve initiated a process called minimum viable cataloguing, so we’ve identified fields in the PBCore schema that we think are important to prioritize when we don’t have a full time staff member cataloguing all the time.

So we have graduate student interns spending 15 or 20 minutes per item doing minimum viable cataloguing adding as much metadata as necessary for minimum discoverability.

You know, that’s going to take probably six years at least to go through all the 40,000 hours that have been digitized. And if we were doing full cataloguing, like Library of Congress subject headings, I’ve estimated that would take 32 years. So we’re not going to do that right now.

And we also developed a public interface. So the Archival Management System is where we manage our metadata. The public interface is available at AmericanArchive.org where we provide online access to the public to the metadata records and the online reading room.

And it’s a Blacklight application using a Solr index. And we have a PBCore API that is available from the AMS. And then that data is ingested into the Solr index and exposed to the public.

And we’d love to talk with you all about how our, you know, work can inform the RPTF’s development of their database and website.

MS. CURTIS: So I’m Rachel Curtis. I’m the digital conversion specialist at the Library of Congress. I just started on this project about three months ago, so I’m fairly new. But it’s been very exciting.

So at the Library of Congress, I work on our side. So we took the preservation files and the associated metadata with that. And so we’re taking things that are in PBCore and putting them into the Library of Congress’ collection management database which is called MAVIS.

The issue we’re running into with that is that MAVIS does not support PBCore. So we’ve had to create mapping to take the metadata and ingest it into that system which has created some challenges for us, especially, as Casey mentioned, we don’t have 32 years to catalogue everything.

So I work closely with Library staff on

working on solutions on both the technical and descriptive metadata. And also another one of my projects is to document our process as well.

So with the issues with MAVIS, another issue with the PBCore to MAVIS not using that is that there are formatting requirements that MAVIS has that require authorities to already exist in that system.

So it’s not as simple as taking the information, putting it in the system. There needs to already be a list of -- it needs to see those same entries in MAVIS for it to go over. So we need to create authorities. Right now, a lot of that information is just going into a notes field which doesn’t really make searchability or discoverability easy.

But we’re taking information from AMS. We’re mapping it over to MAVIS. And we also, another thing I’m dealing with is duplication in the system.

There are things in the Library of

Congress that have already been digitized, already have a record, but they’re also being digitized as part of this project. So there needs to be some linking between those records in order for staff at Packard Campus to recognize that these two things exist.

And then we’re also investigating using open source tools like OpenRefine to deal with the huge amount of metadata that we have. And so, yes, just excited to talk with guys and get suggestions on how to deal with all the things we’re working with. Thank you.

CHAIR MEYERS: Thank you. All right. I’d like to turn to John Passmore at WNYC.

MR. PASSMORE: Mary’s going to say --

CHAIR MEYERS: Oh, Mary.

MR. PASSMORE: -- a few things first. And then I’m going to say a few things.

CHAIR MEYERS: Introduce yourself please.

MS. KIDD: Yes. Hi, everyone. I’m

Mary Kidd. And I am an NDSR resident at New York Public Radio. John is my mentor as well as Andy who you just saw speak.

So for those of you who don’t know what NDSR is, it stands for the National Digital Stewardship Residency. And it’s a Library of Congress project funded by the IMLS with new graduates of library and archives masters into archives throughout the country. And specifically in my program there’s four other residents at various institutions throughout New York City.

So what I’m working on is what NYPR is calling a Digital Preservation Roadmap. And so that consists of three main assessments, a collection assessment which is kind of like what is being made, what file formats, how much, how much are they growing?

Another data assessment which is based on the collection assessment which is, like, what descriptive metadata or administrative metadata is being created by these file formats. And then lastly is a storage assessment. So that’s really kind of understanding the holistic life cycle of these digital assets that NYPR is outputting.

So one of the ways that I’m collecting information is through interviews. And so I’m scheduling interviews with staff, sitting down with my phone, and pressing record, and interviewing them for about an hour, hour and a half, and then transcribing them by hand which has been painful.

But also through the transcriptions, I’m able to gather a lot of really important qualitative data that then kind of informs what is -- next slide -- what I’m creating which are these sort of visual work flows of how all the systems that NYPR uses, contextualizing them in these greater sort of production work flows.

And I call this One of Many, Many Flows, because there are many, many work flows going on. This is just what I put together for news. And you see there’s many ways for things to go into these systems. There’s also many ways for them to go out, either on air or through digital distribution through things like podcasts and streaming online.

Another thing that I did right at the very beginning was I had the Archive purchase a license to TreeSize which is like a disc analysis software.

And what that does is it scans the drives and kind of quantifies exactly how much is in each folder. But you can also do other things. Like, I used it to run a checksum analysis and then understand duplication behavior across all the borders.

So to conclude, one of the things that I've discovered as an NDSR resident is that if you're going to work as an archivist, a digital archivist, specifically in a radio archive, you're not necessarily always dealing with sound.

You are handling all kinds of things from pretty much whatever a content creator wants to experiment with. So recently they've been experimenting with Snapchat. So there are a lot of other sorts of media that you have to kind of understand and anticipate in terms of developing a sort of preservation roadmap like I am.

So I’m excited to be here to answer questions about that. I’ll hand it over to John.

MR. PASSMORE: Okay. So, yes. Like Mary -- I’m sorry, I’m John Passmore. I’m the archives manager at WNYC.

And like what Mary was talking about, I think in some respects we’re not just a radio station anymore. We’re sort of a multimedia company. We just reran one of our divisions, WNYC Studios. And, you know, there’s all sorts of content being created and being created in different places, in different ways, by different people, by different kinds of producers.

But usually behind that could, well, there’s some potential for there to be a single digital object, like a piece of audio. But it’s being created in different places, produced in different ways and distributed in different platforms.

And in terms of what that means for metadata is you’re getting all this information about an object, but it’s being created in different places. And information is very different but also very relevant.

Also these places tend to be siloed. So you may have a piece of audio that, you know, went out for a podcast. But there’s a reporter that has a transcript, there’s an intern that entered metadata into the CMS. There’s technical metadata that was in a repository. There’s rights management information in the archives, it goes on and on.

And this stuff is largely unconnected. So, you know, our goal is to sort of create a, you know, wrap that audio object, if it’s audio, in, like, a cozy metadata blanket or something --

(Laughter.)

MR. PASSMORE: -- that’s surrounded by all the information that is created about this object, you know, wherever it exists.

So for example, you know, we have a PBCore database in the archives. We have, you know, maybe in our RSS feed rates we have an object, if it’s a podcast or a news story. There could be technical information about that object in a repository, if it’s a NULL expression, it’s kind of old flavor. There’s also, like, a transcription software which has basically text files which doesn’t exist anywhere else.

So one of the things we’re trying to do is use PBCore to aggregate and normalize all this data and sort of, like, create this canonical digital object that’s wrapped up in a bunch of PBCore information that is basically being harvested by content creators. Because we don’t have the time to catalogue everything.

So it’s sort of a way, by running up on some sort of small microservices to harvest this data, aggregate it, normalize it in a way that we can understand it and in a standard that sort of opens, it sort off, it guarantees long term preservation, or hopefully it guarantees it. And that’s it.

CHAIR MEYERS: All right. Thank you very much.

(Applause.)

CHAIR MEYERS: I’d like to have Ken Freedman now from WFMU.

MR. FREEDMAN: My name is Ken Freedman. I’m general manager of WFMU. We are an independent free-form radio station. We are 100 percent listener sponsored and practically listener free.

(Laughter.)

MR. FREEDMAN: I come to you as a general manager and a practitioner of radio. I am not a scholar. I am not an archivist. And actually I think these facts have freed us up to take a pretty radically different approach than what I’ve heard in the last couple of days.

We began our archiving in 1997 of all of our programming with the goal of making all of our radio archives available to the public immediately without any hurdles, without any registration needed, period. That’s been our goal all along, to simply make everything available to the public.

We started in >97 and we went full time archiving January 1st, 2001. And since that time, we’ve archived everything, 24 hours a day. And we actually have six channels of programming in addition to other podcasts. So at this point, we have probably about 170,000 hours of programming available online for the public.

In order to protect our asses in terms of copyright, in 2001 we went on a campaign to get waivers from record labels and artists to free us from the restrictions of the Digital Millennium Copyright Act. And we’ve, at this point, received about 2,000 such waivers on the DMCA restrictions.

Those waivers do not free us completely. They do not put us in the clear completely. But it has lowered the risk sufficiently that I feel safe to be taking the approach that we are taking.

And this is not significantly different from the approaches of other types of giant archives such as Brewster Kahle’s Internet Archive. I felt vindicated when I asked Brewster what determines what he puts up on the Internet Archive. And he told me flat out stuff that we won’t get nailed for. It is our goal to not get nailed either.

And in fact, I have never had a problem with anything that we’ve put up on our archives. In fact, 99 percent of the feedback that I get on our archives are people who are incredibly grateful that we’ve put their material up or their family’s material up, et cetera, et cetera.

I did get sued once for a recording of people who claim to have been abducted by space aliens. Unfortunately, my insurance company paid 25 grand to the producer of the space alien album. But to me that’s a small cost of doing business for what we’ve managed to do with our archives.

The archives are available to the public. It’s MP3 and MP4. They’re available internally to the station as FLAC and WAV. And our approach to metadata has also been fairly unusual.

We have taken what I've learned this morning is known as descriptive metadata. We do not ascribe to PBCore or any other kind of metadata situation like that. Because we find it incredibly unwieldy.

We have a very small budget, so the approach that we've taken for metadata is to have every program go up with an accompanying playlist page, a web page that the public can comment on. So it becomes actually a very, very rich trove of key words and metadata that is supplied by the public and, in fact, that is ongoing.

The archive approach that we've taken, I think, is much more similar to UbuWeb, a vast and illegal archive of avant garde audio and video, or to the approach of the Internet Archive. And I think that these are other approaches that should at least be considered which don't seem to be considered here at this conference.

The numbers that we've heard here, while impressive, are nothing compared to what an institution like the Internet Archive is able to secure and archive. Brewster told me that, at this point, he is archiving 1 billion photos a day from the Internet, every single day.

I feel that a lot of the archiving approaches that we've been discussing here are very, very difficult, expensive, unwieldy. And because of this difficulty in preserving the past, we are letting the present slip away.

For example, yesterday I heard a lot of discussion of Father Coughlin, the American Fascist broadcaster. But now we are about to have a newfound Father Coughlin become President of the United States, perhaps the first talk radio President of the United States.

And yet all of the broadcasts of Michael Savage, Rush Limbaugh, Mark Levin and all these other right wing talkers have probably gone completely unarchived for history. And I think that’s a real shame.

And it’s incredibly easy to archive these things. They’re all being streamed on a daily basis. I wish the Internet Archive was archiving talk radio. Alas, they are not. But that’s at least a basis for some discussion. Thank you.

CHAIR MEYERS: Thank you, Ken.

(Applause.)

CHAIR MEYERS: I’m going to turn it over now to Marit MacArthur who will tell us a little bit about what she’s been doing.

MS. MACARTHUR: Hi. Can you hear me in the back? Okay, I’ll try to keep it at this volume. Thanks to Cynthia, and Josh Shepperd and Neil Verma for adding me to this committee, really at the last minute.

I’m a poetry scholar and a sound studies person, not an expert in metadata or digital archiving, but I’m a librarian’s daughter. I’ve worked in libraries and done a lot of research in traditional archives. So I feel like I understand their limitations.

Right now I’m an ACLS Digital Innovations fellow at the University of California - Davis. Where I’m adapting and developing linguistic analysis tools with speech recognition and machine learning methods for humanistic research on noisy, low quality audio archives.

And I’ve also been involved as a researcher in the NEH Digital Humanities Institute High Performance Sound Technologies for Access and Scholarship, HiPSTAS, which is directed by Tanya Clement, some of you will know.

And I want to mention that HiPSTAS has supported a machine learning audio analysis platform called ARLO, Adaptive Recognition with Layered Optimization, which may soon get a boost of funding from an IMLS grant which could make it useful for analyzing large audio archives.

But I’m here today not representing HiPSTAS or ARLO but to introduce three open source tools that could be quite useful to the Radio

Preservation Task Force.

One is a combined transcriber -- which, Mary, you could use, and you would have to stop transcribing by hand -- combined transcriber and forced aligner called Gentle. A pitch tracker called Drift which visualizes and quantifies paralinguistic features of vocal performance like pitch range intonation patterns, speaking rhythm and tempo which are crucial to nuanced analysis, given the highly subjective nature of auditory perception.

I’m happy to say more about that in conversation for anyone interested in analyzing non-musical vocal performance.

But the tool I will concentrate on here briefly is called the App and the Territory, after that Michel Houellebecq novel, the Map and the Territory, whose title implies the inevitable discrepancies between the map and the terrain it purports to represent.

We all know that know that big data is

transforming research in the humanities and social sciences. Yet many of the interfaces that we use to interact with big data still ask us to consider each data sample, or each recording in our case, one at a time and to know what we are looking for before we click on it.

If we want a specific radio play by Orson Welles, that’s one thing. But if we want to browse around in a collection of, like, obscure small town radio shows with no metadata or very little, we need to be able to explore without knowing exactly what we’re looking for.

To enable users to explore audio archives, someone or something has to generate metadata about them which is usually very expensive.

And this is just an example, kind of typical interface that I’ve dealt with online. This is PennSound, the large audio archive of poetry recordings at the University of Pennsylvania.

Sorry, this is a little bit small. So, you know, and it’s various ways to search and look for different recordings. But what I often do is, like, okay, well let's see. I want to open to John Ashbery. Okay, well, which column and which recording do I -- you know, it's endless decision making, right, rather than exploring.

The developer of the App and the Territory, which is where every dating is AATT, is an amazing creative developer trained in computer science and art at Cornell. His name is Robert Ochshorn. And he makes the obvious and important point that, AMetadata is only a God-given right for commercial mass media.@

And in his talk against metadata at the 2014 Chaos Communication Congress in Hamburg, Germany, he argued that we should, AStop designing software that assumes the prior existence of metadata.@

So AATT addresses the issues I've mentioned by enabling an audio archive to self-organize and then allowing the user to interact with many, before he said once, directly.

It sorts individual recordings according to fine grade similarities and then visualizes them with a method called t-SNE, which stands for t-Distributed Stochastic Neighbor Embedding, which embeds a high dimensional space audio similarity in a low dimensional one on the computer screen.

And I’ll show you a map. And I’m sorry, the projectors aren’t very focused. If this were in better focus, you would see that every point of light is very distinct. There’s also a prettier version that’s blue that kind of looks like, you know, the earth from space at night.

So this dynamic map of AATT, which I’ll play in a moment, represents an archive of 80,000 sound effect files which are much more difficult to cluster than music. And each sound is represented by a point of light. And it took AATT seven hours to sort these sound effect files.

And as you listen to it, you’ll see the red circle which is moved over the map with a mouse.

It plays each sound point as it glides over it. And we can hear groups of sounds together, like staccato sounds, buzzing machine noises, female phone operator voices, beeps, gunfire, applause. So this is about a minute. I'll just play a little bit of it. And the audio is off. Sorry, it was on when I tested it.

(Pause.)

MS. MACARTHUR: I'm sorry, I'm almost done. But I really would like you to hear this.

PARTICIPANT: Is there a URL that we can go to after --

MS. MACARTHUR: Well, we can also start discussing. And, like, as soon as it works I'll show you. Just, what I was trying to figure out, I'll tell you the rest of what I was going to say about this.

So the App and Territory allows the user to explore an audio archive, even one entirely lacking in metadata, by interacting directly with and listening to samples of the recordings intuitively discovering patterns and associations.

The master map for a given audio archive can be further reduced to smaller maps of similar sounds filtering -- thank you -- by filtering with a metadata query if you have sufficient metadata, or narrowing results based on maximum distance from a selected sound. Because it's organized by audio similarity. And also the data about audio similarity can form the basis for developing metadata.

The t-SNE method that AATT uses was developed in 2018. And it won the Merck Visualization Prize in 2012. It's been widely adopted for visualization and many types of research from bioinformatics to computer security to music clustering. And I think it's time to try it out on the kind of archives we work on. To me it would be like an ideal finding aid for the 21st century.

(Video playback.)

MS. MACARTHUR: That gives you the idea. There are more voices later. Thank you.

CHAIR MEYERS: Okay, thank you very much.

(Applause.)

CHAIR MEYERS: I'm going to hand this over to William. How do you pronounce your last name?

MR. VANDEN DRIES: That's okay. Vanden Dries.

CHAIR MEYERS: Vanden Dries, to talk a little bit about what the RPTF has collected so far and what he's done with it so far. And if you could introduce yourself.

MR. VANDEN DRIES: Hello. My name's William Vanden Dries. I work at Indiana University. And first of all, thank you, Cynthia and Josh, for giving me the chance to share what has happened so far.

I joined with the metadata team with RPTF several months ago. But a little background, before that I had been working on, kind of tackling one of the recommendations in the National Recording

Preservation Plan -- the one to create some sort of searchable directory of sound recording collections around the country.

And while working on that, Sam Brylawski put me in touch with Cynthia, and Josh and Neil. And I’m very happy he did. Because radio collections are certainly a great place to focus for the initial building out of something that will let us accomplish that recommendation.

And so they initially shared a subset of the data that they had been collecting for, I think, about a year or so. And I took that and reformatted it, kind of took a look at what they had, and then built up kind of a quick and dirty instance of Blacklight, the same app that the AAPB is using.

And a little while later, somewhat recently, they shared the overall set of data that they had collected. And they had basically split it up into three regions around the country and had different teams working to collect metadata about the institutions, broadcasting stations, private collectors and anybody they could find that had relevant collections to radio.

And then I eventually got all that metadata. It varied a little bit in the fields that they collected and the depth that each region went into. But they would bring it all together and, with a little bit of editing, a little bit of enhancing, I ended up pulling out about 432 collection holders and a little under 650 collections for representative data.

And this is really just a start, because there were a handful of institutions, and broadcasting stations and collectors that were listed but no specific collection information had been collected yet.

This was really, according to the team, really a first contact with all these places. And it was always expected to have more follow-up later on.

And that follow-up was expected to

happen after this conference when we could have a discussion about what metadata scheme, what fields would be most useful to request from those places moving forward.

So I won’t go into too much detail about Blacklight and all that, but I was really happy to have a great conversation with Casey yesterday where she went into depth about kind of the back end of their software and in what ways we might be able to share metadata with them in the future. Because we are going to be working with some of the same sources.

In addition to this, the Association for Recording Sound Collections has joined the project. They are committed to being the host for this website. So we’re exploring the feasibility of that and what we would need to do to make that happen. And we hope that moves forward in a positive way as well.

Let’s see, so I think that some challenges that lay ahead are the fact that while we would love to follow all the methodologies that AAPB has done and, you know, we can learn a lot from them, there are certain differences between the types of metadata that we’re collecting.

They are going down to -- so the AAPB is focusing on the item level primarily. And the goal of the metadata collected by the RPTF task force has been at a much higher level -- collection level.

And so we need to explore whether PBCore can represent that data well and if not how we can move forward with some sort of metadata fields that can in the future work with PBCore.

Casey informed me that there are a couple of people who are working to harmonize archival collection description with PBCore which would be fantastic if we can make some headway in that direction.

Because there is a lot of precedent in archival description that these fields have already been used for years, and years and years. And so we should definitely try to take advantage of that.

So one thing that I hope we can get out of today, if time permits, is coming up with maybe a fair minimum set of fields that would be useful to people that would be visiting this website and looking for collections and then maybe a couple of more extended sets of metadata fields that would be helpful if the information is available if the cost and time permits.

I think that is pretty much it. I just want to stress that in moving forward we're going to definitely be expanding this dataset. We're going to be a lot more consistent about it. But the discussion that's going to happen here today will help us move forward in those directions.

CHAIR MEYERS: Can you tell us anything about what we did collect? Because different researchers interacted with different institutions and got different categories of information.

MR. VANDEN DRIES: Sure. Let me pull it up on your system.

CHAIR MEYERS: So just to review, RPTF

started out with a group of college professors who contacted different institutions and collectors in different regions. And our mission was just to find out if there was stuff there. So we have a kind of a variety of information that we collected. We didn’t have a standard kind of collection, not that it’s so --

MR. VANDEN DRIES: So this represents the fields that are currently in the data. Some of it was collected by the RPTF team, some of them were additional fields that I actually enhanced the data with while I was working with it. And this is just for the broadcasting stations, institutions, and private collectors.

So not every record that was indexed into the Solr index in the Blacklight was actually -- had all these fields filled out. I did what I could up to this point, but it’s definitely -- a lot of follow-up is needed.

And then as far as the collections go, there were a lot fewer fields that were collected.

And I have not started trying to enhance the fields or expand the fields that were collected by the task force yet.

But, you know, I definitely have some ideas. I’m a digital archivist at Indiana University, so I work with archival collections a lot. And so I have some ideas on collection level description that we could integrate with the data that -- or enhance this data with.

One thing that I was really happy to see is that, when a collection was identified at an archive or a special collections department, a lot of times a finding aid URL was documented which will be incredibly helpful moving forward to go and retrieve some of that collection level information.

CHAIR MEYERS: Okay.

MR. VANDEN DRIES: And I won’t go into all my ideas about the fields that we need, because I don’t want to take up any more time. But I’ll be happy to discuss that with anybody in more detail.

CHAIR MEYERS: Okay. Thank you,

William. Okay. We have a number of discussants, but we also have a limited amount of time. I’d like our discussants to just stand up and identify themselves. And then I’d really like to start the discussion. So, Eric, where is Eric?

DR. HOYT: Eric Hoyt from University of Wisconsin - Madison. And I photograph the Media History Digital Library.

CHAIR MEYERS: Okay, Jack.

MR. BRIGHTON: I’m Jack Brighton. I’m apparently with PBCore. I’ve been involved in that project for about 10, 11 years. I’m also with the University of Illinois, Urbana-Champaign, WILL public television and radio, and also with the Institute for Nonprofit News. I’m a producer and an accidental archivist.

CHAIR MEYERS: Stephanie.

MS. SAPIENZA: I’m Stephanie Sapienza. I work here at the Maryland Institute for Technology in the Humanities, a digital humanities center at University of Maryland. And we develop projects where we work with professors, either at Maryland or outside, to develop tools, sometimes out of archives and sometimes not, for the study of various aspects of the humanities online. Sometimes that’s media related and sometimes not.

My background from before this, I worked as the project manager on the American Archive for three years in its incubation stage while it was at CPB before we handed it off to Casey and Karen. And then before that I was doing experimental film --

CHAIR MEYERS: David Walker. Is David Walker here?

PARTICIPANT: He’s in another session. He was double-booked.

CHAIR MEYERS: Oh, he was double-booked. Why don’t you --

DR. WILLIAMS: Hi, I’m Mark Williams from Dartmouth College. And I conduct an ambitious research project called the Media Ecology Project. If you could Google those three words, that would be wonderful.

What we’re trying to do is advocate for more and better online access to archival content but to put it in relationship to platforms and tools that scholars and academics can use to create both traditional and new kinds of 21st century scholarship.

And the metadata that’s created can be harvested back by the archives. So we’re trying to make it more than a one-way pipeline. And I would like to advocate for and have this great group help to advocate for the creation of a scholarly, secure tier of access which we brought up at the last session.

Not to, you know, make things only secure, only available to scholars, but I know from meeting with various archive content holders, various parts of their collection could really only be accessible through that kind of a tier of access. And we do want to try and open up more material.

CHAIR MEYERS: David Pierce? Did he

get double-booked too, David Pierce?

DR. HOYT: I'll try my best to represent David. David co-directs or, excuse me, founded and directs the Media History Digital Library which I work on with him.

CHAIR MEYERS: Okay. All right. So I'd like to start --

PARTICIPANT: What's your name?

CHAIR MEYERS: Oh.

DR. HOYT: What's that --

PARTICIPANT: What's your name?

DR. HOYT: My name's Eric Hoyt.

CHAIR MEYERS: Did I not, yes.

PARTICIPANT: Yes.

CHAIR MEYERS: I'd like to start with the discussants. And I would really like it if people jumped in with some specific suggestions or priorities for the RPTF to start working on. Like, in your experience, what do you think we should be worried about first.

MS. SAPIENZA: I was going to jump in,

because I keep harping on, I hear from William that you’re looking at data that’s in various formats from different providers, some of which is, a lot of which is collection level and some of which is item level.

And as I’m looking back, like, everyone who works on the American Archive knows for a long time we were under a lot of constraints, bureaucratic, funding related. And we didn’t always do everything the way that we would do it in a vacuum.

And obviously there were so many stakeholders involved that at some point or another, you know, decisions were made and we had to move forward.

But if I had to erase all that, and I was a student in the class and I was given just the task of, like, saying what would I do, Stephanie Sapienza, if I had a clean slate, I would say hybrid federated search or Linked Data.

And the reason I say that is because I

have always wondered, and I’ve never understood why more archives don’t do federated search.

The model for, I mean, many of you probably know what that means, but for those who don’t it’s when, instead of searching one central database where the data is harvested or pulled into one location, it’s querying through an API data at different, through different databases elsewhere. And that allows for a little bit more flexibility in terms of you don’t have to impose any kind of structure, like structured metadata, on the other institutions.

The work goes into the API so that you’re working on the way that it queries the data and brings it back into a form where it’s readable, but it doesn’t force you into any kind of structure.

And the reason I say hybrid is because there are, as we discovered at the American Archive, many, many institutions who need something. They actually have no system in place whatsoever. They’re just using CSVs.

And some people really need something like the Archival Management System to do something from scratch. And we have actually, you know, the Mint. We have a system within the AMS that mapped any kind of metadata that you had, either FileMaker or whatever. And you can transform it into a form that can be pulled in.

But that’s still going back to that, like, harvesting notion. So it’s, like, if someone’s actually going to do that, they should basically, in my opinion, they should agree that that’s going to be the way they manage their metadata for good from now on.

And that they should, actually, when you have this, it should ever happen once, otherwise, I would say instead of something which allows you to query different types of metadata at different locations.

And not only that but, you know, the project that I’m trying to get through, this is the second time we’ve submitted it for funding with the

NEH, Eric Hoyt is involved in it, Josh Shepperd. Casey Davis is an advisor on it. But we are trying to take the NAEB, the metadata and files for the NAEB collections, which are here at Maryland, which is one of the national content collections we got added to, all the stations lifted content.

There’s 3,500 hours. It’s pretty substantial. And it’s all pre-NPR material, 1950 to 1970. And there’s a large collection of paper materials that’s only at EAD. It’s at the collection level at Wisconsin Historical Society.

And so the project we’re trying to fund basically creates a Linked Data hub system similar, based on the model DPLA that allows you to take the collection level metadata from the paper and connect it via authority records to the item level metadata.

And so it’s that kind of thing. It’s not, I’m not saying my project is like the be all and end all of that, but that kind of thinking where it allows you to connect to different systems but not force them into one --

MR. VANDEN DRIES: Right. And I completely agree. And one of the initial stabs at working on the opening of the broader sense of the directory that the plan calls for included an effort to use Linked Data.

And I think in moving forward just in the library and archives community in general and then just overall the Web community and data communities, Linked Data is going to be essential. And so we might as well build it into the project from the beginning.

MS. SAPIENZA: Just one more thought on this, just a final thought. Footage.net, for those of you who don’t know it, I’m not saying it’s the be all, end all of, I’m just saying it’s a great example of federated search in action that works really well and very seamlessly. And I’ve been using it forever. And I --

PARTICIPANT: What was it again?

MS. SAPIENZA: Footage.net.

PARTICIPANT: I’m sorry?

MS. SAPIENZA: Footage.net. It’s like a stop footage search engine that queries, like, ABC News, NBC News, and major archives from all over via distributed search or federated search and pulls them into a central results interface. But they’re still kind of the original data, metadata is existing outside.

MR. VANDEN DRIES: So, yes. One thing I remember, Casey, you were saying you guys had to do a lot of hand holding with some of the sources. I would expect that would be the same if we went with the federated search Linked Data creation of this even if they were ultimately responsible for managing it.

I know we could probably learn a lot from your experience working directly with the collection holders and trying to make that as an efficient process as possible to help them create the data.

MR. BRIGHTON: Do we have a question over here? One over here?

PARTICIPANT: Is that me?

MR. BRIGHTON: Yes.

PARTICIPANT: Oh, thank you. I do have a question, because I guess as much as I'm interested in the metadata, I have no idea how librarians do it.

I'm worried about the taxonomic designation of genre, so genre such as social, flexible, and firm. And I'll distinctly note Ken's show, Seven Second Delay, is it comedy, is it documentary, is it experimental? It's all three. But typically in those fields I only get to choose one.

MR. VANDEN DRIES: Yes.

PARTICIPANT: Is there a way you can build into the system that kind of flexibility of multiple genre labels and --

MR. VANDEN DRIES: Sure.

PARTICIPANT: -- just, so it's nothing to worry about, but I typically only see one option when I see pull-down menus, you know. So like the liberation of genre, particular with music, I mean, that's one area.

And I'm just like, you know, that shifts from decade to decade. As I tell my students, I grew up in the >80s. I never, never once ever, ever, ever did I ever say this is the best 80s music I've ever heard.

(Laughter.)

PARTICIPANT: So that's all ex post-facto, right. But I need that ex post-facto aspect for them and for me.

MR. BOTTOMLEY: Finish your podcast, like, if you're using iTunes for instance, where -- and which, like, independent podcasters are identifying their own genre. I mean, one of the genres for podcasts is podcast. Like, you know --

PARTICIPANT: That's the best genre?

(Laughter.)

MR. BOTTOMLEY: Right. And there's a lot of these things where it's a hybrid. And they can only pick one. Is it comedy, is it talk, is it, you know, music?

CHAIR MEYERS: I had a, oh, go ahead.

PARTICIPANT: Since the podcasting came up, I had a question for the podcasting archive. Are you taking into consideration changes in how the creators want their podcast accessed?

And are you thinking in terms of creating paywalls and things of that sort, sort of after it is initially free to the public and teaching the students the hardcore history?

And I'm aware that the earlier podcast episodes that were originally free are now only available through paywalls. So if you're grabbing all of that content, are you keeping track of those that the creator wants to --

MR. MORRIS: Yes. There's a field that is like a payment URL metadata field that we're trying to keep, like, a track of whether or not that's happening.

But, I mean, right now it's not available to anybody other than me when I'm on UW campus. So integrating it with the wider library system is a bigger question about, like, how and in which ways do we make it public.

So like I said, the reference file is how people would see it. So you’d see, and if the file’s available online, you could link to it. If it’s behind a paywall, you wouldn’t be able to actually play it from the interface that I showed you.

I may have a copy on my server, and that’s where I was, who was it, Marc was mentioning, you know, like a researcher specific login, fair use, you know, like I need to look at the early Marc Maron episodes. Because I’m doing a dissertation on that, right.

But those things that would otherwise be paid for would not be like that, or commercially sort of available, would not be widely available. They’d be researcher access, unless we could reach out to the podcasters and say, hey, this is the thing that we have going on, you know, and get clearance, get permission.

CHAIR MEYERS: I'd like to actually ask a question of our committee and the audience about metadata aggregation. Are there other aggregators out there? There are a number of sites that are aggregating archive collections. I've seen one at danceheritage.org where they're aggregating archive collection information about dance.

DR. WILLIAMS: Aggravating?

CHAIR MEYERS: Aggravating -- aggregating and aggravating.

DR. WILLIAMS: Aggravating also, you know.

CHAIR MEYERS: So what can we learn from these other efforts? Does anybody think that there's one that we should particularly consider modeling ourselves on? And I'm also thinking about interface and user experience. If anybody has used one that they thought -- and I'm also looking at things like Europeana, I can't even speak, Archive Grid, DPLA, Europeana --

MR. BRIGHTON: Yes.

CHAIR MEYERS: There are a number of different --

MR. BRIGHTON: They're kicking our ass, basically.

CHAIR MEYERS: So would somebody like to, yes?

PARTICIPANT: I'd like to throw out a suggestion in terms of aggregation, you know, in terms of a model that DPLA is doing and hosting regional hubs, right.

And I think, you know, one of the underlying themes that is running through so many of the sessions the last two days has been, you know, you have the big institutions, the big archives. And then you have the smaller stations and the smaller archives that have a disparity in resources, disparity perhaps in staffing, metadata, et cetera.

And one of the solutions around that, of course, is collaboration, and particularly around something like metadata, right. So if you think about aggregating that, say at the regional level, you can imagine that metadata could be created that has a particular kind of geographic focus, whether it's by state, whether it's by region of the country.

And that could sort of disperse some of the painful resources around metadata creation and find some kind of adherence that could encourage broader access across stations, across archives. So that aspect with the DPLA is, I think, something to consider, as you consider moving forward with metadata --

PARTICIPANT: So the southwest would fund all the southwest material with the metadata list, the northeast, blah, blah, blah, blah, blah.

PARTICIPANT: Yes, exactly. Assuming that there are shared programming interests, there are shared topics, there are shared themes.

PARTICIPANT: So you want to do it regionally, not by one specific metadata interface that, like, federated search where anybody could get it. Each region would have their own selected things or, I’m just asking --

PARTICIPANT: No, I mean, I think that’s up to --

PARTICIPANT: It’ a question.

PARTICIPANT: -- up to the task force, up to the caucus to consider, but, you know, in terms of where -- there are some interesting inroads with metadata.

DR. HOYT: I think we also have to acknowledge that for most users Google will be that aggregated search platform.

So, I mean, I think thinking regionally and thinking about internal coordination, like, on the internal level that can be really valuable. But in terms of user experience, how are they going to find these things?

I think, you know, we can build really good intermediary platforms, but we have to acknowledge that the majority of users, they’re going to go to Google first. But the good news is that, if we create metadata and index it in a particular way, they'll find it through Google too.

I also think, and I want to hear what you have to say, Karen, but if I can just throw out sort of two concepts that maybe we can, like, print on T-shirts and bumper stickers. One is Casey's minimum viable cataloguing. I just, I love that term. And I think that's something we should absolutely embrace.

And we figure out what's essential, and then we ask, like Tim was saying, maybe users to weigh in on things like genre that are important to them. But then the second concept would be minimum viable API.

So, like, for the kind of Linked Data, federated search that Stephanie was describing to work, you also need to make sure that collections, websites can speak to each other which is not always the case. So if we're going use that rather than harvest and build an index, I think we need that. Karen?

MS. CARIANI: I was going to say the issue is that you’re making an assumption that all the people that have these collections have their collections online.

DR. HOYT: Yes.

MS. CARIANI: Because that’s the only way Google is going to find them, if they actually are online. So that’s kind of a big assumption. Because a lot of entities don’t have that capability or that ability. And that’s why they are sending data to the people or to the entities that know how to aggregate. Does that make sense?

DR. HOYT: Well, I guess, I mean, I was thinking more about the descriptive cataloguing kind of information. That becomes the discoverable on Google even if you can’t listen to the source file.

MS. CARIANI: But that’s assuming that they have their collection catalogue online. I mean, in a lot of the academic institutions it’s behind the firewall in academic institution. And it’s not going to be in --

DR. HOYT: Yes. No, that’s a good point.

MS. CARIANI: -- radio stations don’t even have a catalogue online, much less a catalogue at all.

DR. HOYT: Yes.

MS. CARIANI: So I think in terms of having the aggregations or having something that they can send their data to where it can be aggregated is a great thing.

DR. HOYT: Yes.

MS. CARIANI: I was actually going to speak to Stephanie’s point about Linked Data too which is, and I could be completely wrong and not understand Linked Data, but when we tried to explore Linked Data, you have to have a unique identifier for the term, right.

And that has to go into some major place. And that effort to build those terms in those unique identifiers is a huge effort which I think is what scares people away from getting involved in Linked Data.

And I don't if this group could perhaps be an entity that could start building that catalogue of, you know, unique identifiers for terms around radio collections. It would be a huge step for people to then be able to take advantage of it.

MS. SAPIENZA: And that actually, that was one of the reasons we added the EAC-CPF as authority files. That doesn't actually solve the problem fully, but for this particular project it was, we ran into that. And it was our way of linking, for example, the collection level EADs to the item level.

But, yes, I think that the barrier is, it's technology. People in our field don't know how to develop APIs generally. And so we do what we know how to do.

But I think Karen's very correct, there's a huge barrier in terms of that one particular point. But I think a bigger barrier is how much work it took for us to do a lot of the things we did on American Archives the way we did them is also a very, you know, it was very labor intensive too.

And I feel like either one has its pluses and minuses. But I think, what I kept running into is that a lot of stations, even WNYC and others, like, they have their own internal system that they don’t necessarily want to stop using.

So if there’s any way that the RPTF can think of a way so allow stations, or not even stations, just archives, holders of this content, to continue using their own systems and not use one centralized system, if there’s any way to explore that, I think it would be worthwhile at least to explore.

MR. FREEDMAN: I think there’s a lot of low hanging fruit, which are current existing streaming broadcasts, which are being sent out on the Internet with metadata.

It’s not the metadata that we’re talking about here. It’s the metadata that actual radio listeners and radio station managers care about. And that stuff is going out 24 hours a day on more radio broadcasts than ever in history. And nobody’s capturing any of that.

DR. HOYT: That’s one of the things you’re doing, Jeremy, right, with the podcasts?

MR. MORRIS: Trying to. I mean, it’s, again, still from RSS feeds and stuff. But it is, I mean, there’s lots of stuff there that’s usable. But trying to capture it across a whole bunch of digital audio that’s up there it’s, I mean, we’re literally restricting ourselves to a podcast of sorts.

DR. HOYT: Yes.

CHAIR MEYERS: Mark has a question.

DR. WILLIAMS: Yes. Three aspects of Media Ecology Project that we hope will help to facilitate answers to some of these points, one is we utilized Mediathread which accesses streaming online content and provides a capacity to create time-based annotations which will develop Linked Data kinds of capacities for search.

We also have a tool called onomy.org. This is my little joke. It’s like taxonomy without the tax. I’m from New Hampshire. You get bonus points --

(Laughter.)

DR. WILLIAMS: - when you take away the tax. Silly joke, but it helps you to remember what it is. And it’s to help build controlled vocabularies so that we can all be on the same page about what’s the best kind of taxonomy for radio broadcasts, right. We’d love to have your help in developing that.

MR. FREEDMAN: And we will never be on the same page. How long has PBCore been around? I mean, it’s just like it’s never going to happen. And I really --

DR. WILLIAMS: Well --

(Simultaneous speaking.)

MR. FREEDMAN: -- agree with what you

say about minimally viable metadata. I think it's the way to go. I really think less is more.

DR. WILLIAMS: I'm going to beg to disagree in a respectful way. I think if we don't use seven terms for the same thing it's actually to our advantage. And that's what we want to do, is try to harmonize whatever terms we're using, you know.

The third thing, we just were very delighted to have a new NEH grant to develop a semantic annotation tool. And we would love your input. We have a survey to help us understand how you would like to use such a tool. And we will put the URL for that onto the feed for the conference. We'd really love to have participation.

MR. BRIGHTON: Personally, I think, going back to this minimal viable metadata question and PBCore, which I apparently represent --

(Laughter.)

MR. BRIGHTON: I think that RSS is the wonderful first example of an aggregation format that allows you to build collections and to, you know, download the essence and to, you know, actually have a catalogue of a whole collection of materials.

And I think PBCore came along as a way to extend RSS to say, okay, who was the creator, contributor, location, you know, duration, format, a lot of other metadata that went along with that.

But you don’t have to have all that. You can start with the basics which is RSS. And then on the other end, nobody’s mentioned the Public Media Platform. And I don’t know if everybody knows about the Public Media Platform or the PMP.

It’s a thing that was built by a consortium of NPR, Public Media International, and a few other, PRX, a few other really good technical partners and content partners.

It allows anyone to contribute, if they choose to, to this pool of metadata. And each item in that, which would include a name or an audio file, or a video file, or a base or location, each of those has a unique identifier. So you can really very, either get, like, a high level view of a collection or drill down, like, incredibly detailed.

And, you know, it's something that I've been playing around with. And it seems like an incredibly powerful way to go about this. And it's not like, you know, the sort of typical way a library, you know, would go about organizing their collections. But it's for producers and distributors if you want to share content.

You could build library services on top of that, you know. I don't know if people know about it, but it's pretty cool.

MR. FREEDMAN: I just think there's a complete disconnect between this kind of approach and the kind of approach that a radio station could actually internally adopt. And you have a huge pool of potential archivists and metadata transcribers at radio stations. But they're not going to go anywhere near PBCore. It'll scare them away immediately.

MR. BRIGHTON: But I’m saying they don’t have to, right. I mean I started out as a radio producer. And I realized real quickly that, you know, producers don’t want to create metadata unless it serves some interest of theirs that they care about. And typically, they don’t care about preservation. It’s not their problem, right?

MR. FREEDMAN: But if you don’t try to enlist radio listeners, and radio station staffs and managers, then it’s all falling into your lap.

MR. BRIGHTON: Yes.

MR. FREEDMAN: And then we’d have the problem that we have which is this gigantic backlog of historic stuff waiting to be archived, and, you know, and catalogued and indexed. And meanwhile, the current stuff, the stuff that’s happening every day, 24/7, is just going out and not getting captured in any way.

MR. BRIGHTON: But they are putting those, like, current content. I deeply care about not letting the present slip away, like you said, exactly right. We can’t just pay attention to the past without paying attention to the future.

They are creating content in technical systems that handle metadata. You know, they have to put it into an automation system or their website where there is a way to get at that. And, you know, to your point, I mean, you can just broadcast stream. You know, the radio shows what’s playing, who the artist is.

MR. FREEDMAN: Right, RDS.

MR. BRIGHTON: Exactly. All that stuff is fodder, without the producer even knowing that they’re contributing to that.

MS. MACARTHUR: I wanted to just chime in with Ken. Because, I mean, it sounds like what you’ve done is crowdsource the creation of metadata.

MR. FREEDMAN: Partly.

MS. MACARTHUR: Partly.

MR. FREEDMAN: Partly.

MS. MACARTHUR: And I think that’s a huge resource that should be used. I mean, because there are all of these people who are obsessed with, like, whatever show or, you know.

MR. FREEDMAN: I mean, the staff --

MS. MACARTHUR: And they know what these things are.

MR. FREEDMAN: Yes, but the staff puts in the basic stuff, like song title, and who was interviewed on the show and what musicians performed on the show. And that stuff's really basic. But then the listeners can go in and continually add more and more to it.

MS. MACARTHUR: You know, and I think maybe librarians would be, like, but that's going to flawed metadata. Well, but it's better than nothing. And then you can go back and correct it.

I think the other thing is that so often, like, people are, you know, working in silos and, like, reinventing and reinventing. And they don't know what other people are doing.

And I don't know how to solve that problem. But I think it's really important. Like, if we could somehow, like, have some central location where any kind of projects you create metadata for radio had to be linked, like Library of Congress. I don’t know.

MS. DAVIS: I’d like to mention one project that the American Archive is working on. I forgot to mention it in my two and a half minutes. But it’s an IMLS funded project.

It’s working with the public archive and HiPSTAS, working on the ARLO analysis tool. So we’re going to be creating transcripts of the 40,000 hours that we digitized using speech to text tools, automated speech to text tools.

Then we’re building a game or a tool for the public to fix those transcripts, because they won’t be 100 percent accurate.

And all of this, all of the tools that we’re creating through this project will be available open source for other libraries, and archives and public media organizations to use. The tool will be open source, the crowdsourcing tool will be open source.

We’re also developing a national audio fingerprint. So we are identifying about 4,000 hours of the content that we digitized that feature speakers that are nationally known. And we’ll be doing audio wave form analysis on those speaker voices that can then be taken into our database.

So if someone, you know, if there’s Richard Nixon, you know, speaking, our database will know that it was Richard Nixon. And it will retrieve that in a search result.

And I think that the transcripts, our automated transcripts are becoming more and more, there are more and more services offering automated transcripts. But using automated transcripts enables you to have tons and tons of key word metadata that you could then index back into your system without, you know, spending hours cataloguing.

MS. MACARTHUR: So I just have a question. Are you paying Pop Up Archive?

MS. DAVIS: No. But the grant is. But there are other services that offer even, you know, less expensive options for those --

MS. MACARTHUR: Well, there’s one that’s free.

MS. DAVIS: Yes.

MR. PASSMORE: The Sphinx tool, are they using that?

MS. DAVIS: I’m sorry?

MR. PASSMORE: Are they using the Carnegie Mellon, the --

MR. BRIGHTON: Yes. Yes, we’ve been using it for the past couple of years and actually set up a PBCore feed into the Pop Up Archive that ingests the metadata, generates more metadata, ingests the audio file, sends the audio file and the metadata to the Internet Archive. So it’s kind of like this automated process of doing that harvest and then generate more.

MR. PASSMORE: The NYPL Labs tool though, is that what you’re basing --

MS. DAVIS: We’re working with them.

MR. PASSMORE: Okay. Because that’s available now. And you can use it. You can upload audio and then crowdsource the annotation to the key words, have people write abstracts. And then it kind of consolidates it by consensus.

So, like, if ten people write de Blasio one way and one person another, then it will pick-privilege the one that’s most often used. So you can actually, I don’t how people would sit down and listen to things on their couch and help you, but it’s worth a shot, you know.

PARTICIPANT: I think in terms of the last bits of the conversation that a lot of the things that you’re suggesting, none of it precludes anything else.

MR. BRIGHTON: Right. Both/and --

PARTICIPANT: Exactly. I mean, you can start with and certainly recommend a minimal viable product. And that can be enhanced in many different ways through the API, the crowdsourcing, you know, professional cataloguers, automated taxonomy.

So I think what, in some ways what you have to boil it down to is some sort of set of, a basic set of specs or standards that you could classify to kind of do no harm approach to production.

So at the production level, and I think the WFMU example is really kind of interesting and exciting, as long as you're not doing something that precludes some other activity down the line, you're not locking content into proprietary format, you're not hiding it behind a firewall.

To the extent that you can provide an API that allows someone else to enhance content, aggregate it, provide some tool on top of it. That's great. But I think that maybe the job is to define that kind of basic do no harm level that allows so many other things to happen. And then as funding is provided, and I'm a former recovering funder --

(Laughter.)

PARTICIPANT: -- so I can speak to that. If they’re few and far between, the options, but when those opportunities arise, then other things can happen. And they don’t have to do a lot of retrospective correction, normalization, standardization that will eat up the majority of the funds.

The other thing that I would say to add to that, again as a former funder, is a lot of what drives funding in terms of deciding between this good idea, and that good idea and the impossibility of funding all the good ideas, is impact, which is not a topic that’s really come up so much.

Because we, as archivists and librarians, tend to get hung up primarily on the standards and the specs in the technology. But you really have to formulate and articulate these ideas in terms of the audiences and use factors.

Or they just look like, you know, a bottomless pit of more money, you know. Yes, we can create the metadata until the end of time. So you have to put it in that framework of who’s going to use it and how and what impact that use is going to have.

PARTICIPANT: Agreed.

CHAIR MEYERS: Yes. And I would like to bring that, that’s a really good point. And I’d like to bring that back to the interface user issues.

I have to confess, as a user of catalogues and archives, I’m just looking for what I’m looking for. I don’t really want most of the metadata that’s there. I’m sure that metadata has a good function and use, but as a user I’m trying to get access. And so that’s my interest.

(Simultaneous speaking.)

PARTICIPANT: You can’t use what you find unless you have metadata.

PARTICIPANT: Exactly.

CHAIR MEYERS: No, I’m being hyperbolic here --

(Laughter.)

CHAIR MEYERS: -- for the purpose of

thinking about the end user rather than the archive.

MR. FREEDMAN: I’m obviously a huge fan of the Internet Archive. And I think the way that the Internet Archive approaches historical preservation is, considering the vast quantity of what they’re saving, they’re doing it incredibly inexpensively.

And, you know, they’re just sucking every television news broadcast onto their servers. And then they’re using voice to text. And they’re just putting it up on the Web. And then they’re letting their fans annotate it.

And nobody’s doing anything, including them, nobody is doing anything like that in radio. And that’s incredibly inexpensive. I think if they were getting hung up on copyright and metadata, they wouldn’t be doing any of that.

I feel like these incredibly complex metadata schemas that we have, as well as copyright psychosis, are the two biggest impediments that we have in terms of preserving radio and preserving all sorts of things.

DR. WILLIAMS: And maybe the psychosis thing, yes. But we're working with the Internet Archive. They're really delighted to work with us on our project. Because we can enhance the metadata they already have.

MR. FREEDMAN: Right.

DR. WILLIAMS: And we can pursue tools like computer vision and machine learning to, at some future date, be able to look for things that basic descriptive metadata --

MR. FREEDMAN: Right. But they're not waiting for you. They're just putting it out there. And then they're adding what you have to add later.

DR. WILLIAMS: But it's a both/and kind of thing. It's a both/and kind of thing.

DR. HOYT: If I could also just put out, a lot of what they're adding is being added by folks like the University of Maryland which scans it.

So, I mean, the Media History Digital Library worked to about 2 million pages of magazines and trade papers from film and broadcasting that were scanned. But really, a ton of credit goes to the University of Maryland that has the scanning center. The Library of Congress did close to, Mike Mashon, a million pages.

PARTICIPANT: And I paid for that.

DR. HOYT: Yes, personally. But then, I mean, so the Internet Archive, you’re right, they deserve a lot of credit, including as being what we’ve been talking about, an aggregator that lets lots of different sites scan things that are important to them and their users and share it.

And, you know, they’re incredibly flexible on metadata in the sense that the fields you can put in are completely arbitrary which, on the one hand, can make discovery quite difficult but on the other hand doesn’t create a barrier to entry for more people contributing.

MR. FREEDMAN: Yes, but discovery is possible.

DR. HOYT: Yes, yes.

MR. FREEDMAN: And it’s up there.

DR. HOYT: Yes.

MR. FREEDMAN: And it’s being preserved and backed up.

PARTICIPANT: Exactly. I think it’s probably going to horrify a lot of people who are working with metadata. I’m a musicologist, I’m coming from a music researcher’s perspective.

And to take a category like genre, to me that category being dynamic and descriptive in a timestamped kind of way is enormously helpful to me to try to discover things, like, the construction of genre which, as you pointed out, are dynamic after all.

And so, I don’t know. I agreed with the gentleman who said this hopefully can be a both/and instead of either/or which is speaking as a music researcher. That’s where I would find value.

PARTICIPANT: Yes. The other thing I would add to the conversation is that all of this is going to depend on a huge infrastructure of continuing education and outreach.

Because, I mean, one of the things that we’re discovering today is that, even though this is a great conversation, and I think the best I’ve heard in the conference so far, is that there’s a lot going on, but nobody else knows about it, all right.

So not just the idea that somebody’s doing something at Dartmouth, or somebody is doing something in this, somebody is doing something at this conference, but then, you know, the iceberg beneath that is how you actually do it somewhere else, right. And what’s the mechanism for that?

So this is, I think, you know, the real practical problem is not just kind of figuring out what all the options are but then delivering the tools, and outreach and all the other things that allow somebody to take advantage of it.

And there’s very little capacity for that out there. And then, you know, we used to fund a lot of that or try to, not nearly enough money is available for those continuing education and training opportunities.

DR. WILLIAMS: And I concur. I think ultimately we want to create a long tail that enables fans, and listeners, and everybody to create and contribute some kind of metadata that will actually benefit the archives.

But I think that has to be part of our pedagogy going into the future, just the variety of scholarly responses here at this panel to the word metadata, you know, it can be very off-putting to a lot of people and probably to a lot of fans and listeners too.

So creating, like, best practices and ways to get people to think about what it means to give back to the archives, I think, would be incredibly useful.

CHAIR MEYERS: In the back.

PARTICIPANT: Let’s use the term information about information instead of metadata per se. Metadata is not, it turns off a lot people.

And when you create taxonomies in an aggregator setting, you’re always going to have people at a local level, whatever that may be -- it doesn’t matter whether it’s radio material, whether it’s art documentation, whatever it is -- who are going to need to adapt that to their own use.

And it requires whoever is creating the taxonomies, or the metadata schema, or the cross walks between schema or among schema, to review, in some kind of regular basis, what they’re doing.

It cannot ever be said it’s done. It’s constantly dynamic. And I have been doing this for 30 years. So I know that my group here at the University of Maryland when I worked here that, less often than I would have preferred, but it took us a year to review the taxonomy that we used for a particular problem, something for everybody that we were talking about.

So metadata isn’t anything that needs to be scary, but the term is so overdue in application. This is a cross-cultural situation.

PARTICIPANT: I just thought it was, I think a possible solution to the problems you're bringing up goes back to the idea of the hub, someone mentioned the DPLA although they didn't answer, D --

PARTICIPANT: PLA --

PARTICIPANT: You know it.

PARTICIPANT: -- the Digital Public Library. And they have hubs. And even Internet Archives, I believe that they have hubs. And also I was referring to the National, the NDSR, what was that?

PARTICIPANT: Stewardship.

PARTICIPANT: -- stewardship, that's right. They have the trainer program where they train somebody. And they agree to do training close to where they live.

So I do think geography is something to consider here and having hubs that will sort of proselytize to their area and for, at least, you know, kind of get a sense of what’s going on there and be the place where the information passes through to get to, from, like, the RPTF --

CHAIR MEYERS: One last, yes?

PARTICIPANT: So one plea, I would like to add my voice to those who ask for some spontaneity in metadata entry, especially for historians in order to get kind of a live opinion rather than something that’s educated away by us or from other or arguments or academics.

And, like, to me the model of Urban Dictionary is great to root out the new slang words. That’s where we go. And it’s completely free form entry and then we’ll make it a lot more convincing or less convincing. Something like that would be great for the problems we’re seeing.

DR. WILLIAMS: More folksonomic.

PARTICIPANT: Thank you. I just want to advocate for doing things well the first time. Because I can’t tell you how much of my time is spent redoing the work of others that didn’t do it well the first time. And I just really want to advocate for that, as opposed to just trying to do it real quick.

PARTICIPANT: Actually, the research from the BBC that's been doing crowdsourcing for quite some time, they've stopped doing crowdsourcing. Because they're redoing everything, because the information, if we're not going to use the ugly metadata word, is wrong. So you can't actually get the information.

PARTICIPANT: And especially with metadata cleanup, that is a huge amount of work. And so, you know, if people took the time and cared to do the metadata correctly and well the first time, redoing it all again is just so much work. So I just really want to advocate for, you know, letting archivists and metadata librarians do it and do it well.

MR. FREEDMAN: Well, radio stations want the metadata to be correct as well.

PARTICIPANT: Exactly.

MR. FREEDMAN: You know, radio stations --

PARTICIPANT: So do the people who did the show.

MR. FREEDMAN: -- their economic interests to have the name of the host, the name of the interviewee, you know, the date, the time, to have that all go out exactly correct. They're not going to put out all that information going out there incorrectly.

CHAIR MEYERS: So this is the beginning of the discussion. It's not the end of the discussion. Please feel free to contact the RPTF. I don't know if my contact information is on the program. But you can find me pretty easily.

But I'd like to hear more feedback about what you think the RPTF needs to do to move forward. And I think we've got some great people to continue to contribute to that discussion. So I want to thank everybody for participating. Thanks.

(Applause.)

(Whereupon, the above-entitled matter went off the record at 12:36 p.m.)