White Paper Report

Report ID: 113674

Application Number: HK-50175-14

Project Director: Kerri Hoffman (kerri@prx.org)

Institution: PRX, Inc.

Reporting Period: 9/1/2014-8/31/2015

Report Due: 11/30/2015

Date Submitted: 11/25/2015

White Paper
HK-50175-14
Pop Up Archive: Saving culturally significant audio through preservation, searchability, and distribution
Project director: Kerri Hoffman
Grantee institution: PRX, Inc.
Date: November 30, 2015

Background on Pop Up Archive

Since 2012, Pop Up Archive has worked closely with oral historians, radio producers, and online content distributors to develop a simple, day to day system for ensuring that significant audio is preserved, backed up, searchable, reusable, and shareable — without requiring technical expertise or substantial resources from users.

In its first phase, as part of the Masters work of co-founders Bailey Smith and Anne Wootton at the UC-Berkeley School of Information, Pop Up Archive assessed and built a solution for the "accidental archive" of The Kitchen Sisters, award-winning producers Davia Nelson and Nikki Silva. The breadth of the Kitchen Sisters' archive gives a sense of the variety and sheer volume of culturally significant audio recordings in need of preservation in the United States: over the course of creating public radio documentaries and stories since 1979, The Kitchen Sisters have amassed thousands of recordings that chronicle the B-side of history, documenting the lives and stories of people, well known and unsung. The Kitchen Sisters frequently receive requests from scholars, producers, and the public to access this content, but it is difficult — and often impossible — to fulfill these requests. Files are stored in various formats on hard drives across multiple locations, searchable by no more than a catalog number.

Pop Up Archive's proof-of-concept for The Kitchen Sisters was built on top of open-source web exposition software, Omeka, to provide a digital home for archival audio. The initial Pop Up Archive release provided a plainspoken mapping to PBCore, creating a basic data model capable of representing audiovisual content from anywhere in the world. An NEH Digital Humanities Start-Up grant and ongoing support from the Knight Foundation supported this phase of Pop Up Archive, which resulted in Omeka plug-ins for digital preservation of material at the Internet Archive and accessibility/shareability through SoundCloud. To address the needs of those attempting to preserve and access culturally significant audio without the ability to maintain their own server or software, Pop Up Archive also hosts a centralized, searchable standalone database and website for archival audio, popuparchive.org, that provides automatic transcription and keyword extraction. As of its initial formal launch on November 21, 2013, the Pop Up Archive service:

- Enabled anyone to search, filter, and access sound — whether their own collection or material from national media organizations and archives
- Enabled anyone to preserve their sound in a cloud-based, accessible digital location.

Project Activities

For the NEH Digital Humanities Implementation grant, in collaboration with PRX, Pop Up Archive identified and integrated speech-to-text software that performs with a much higher degree of accuracy than initial proof-of-concept work with the Google web speech API. Pop Up Archive conducted a pilot program under the advice of the BBC R&D team and researchers at the International Computer Science Institute in Berkeley, CA. In 2014, we put out a call for content partners to contribute training materials and be the beneficiaries of the three month speech-to-text pilot period. Pilot partners included the Hoover Institution at Stanford University, the Princeton Theological Seminary, KQED, KCRW, CUNY-TV, Illinois Public Media, Snap Judgment, and the Studs Terkel Radio Archive. We first worked with a team at ICSI to train our own speech-to-text acoustic and language models, using audio and transcripts from pilot participants and Pop Up Archive partners from 2012-2014, as well as text materials collected from major news sites. The domain-specific language models that resulted from this work included a general public media/oral history model, a religious model, and a 1980s model.

We conducted our speech-to-text experiments using Kaldi, a relatively new open source speech-to-text software that is quickly being adopted by research communities around the world. As the BBC speech-to-text work with CMU Sphinx (part of the COMMA initiative) wound to a close and they started working with Kaldi, we identified researchers who have trained Kaldi acoustic and language models for U.S. public media to degrees of accuracy far surpassing anything we accomplished in our experiments. We became their first partners and began providing public access to their software through Pop Up Archive in September 2014.

While we must shoulder per-hour costs for transcription through this service, its pricing is currently competitive with (and we anticipate will drop below) the cost of training, implementing, scaling, and maintaining our own servers and speech-to-text software. Our focus remains on refining the speech-to-text output through controlled vocabularies and speaker data as well as bringing this software to the public media and oral history communities in the U.S. through easy-to-use interfaces and widgets and ongoing educational outreach and strategic partnerships. As we monitor unfolding Kaldi developments, we continually revisit the question of whether it makes sense for us to employ our own speech-to-text implementation. In the meantime, we also plan to release our Kaldi models publicly through Github (www.github.com/popuparchive) so they can be accessed and potentially adapted by interested parties.

We have integrated controlled vocabularies, using DPBedia, a project that extracts structured content from the Wikipedia project, as a controlled vocabulary by which to reject or approve the automatically-generated entities.

In addition to presenting at conferences such as Code4Lib and the Association of Moving Image Archivists and publishing articles, blog posts, and newsletters to spread awareness of computational methods for searching and indexing audio (see appendix), two of our team members sit on PBCore subcommittee teams charged with education and communication around the PBCore audiovisual metadata standard, crosswalks with the PBCore schema, and documentation of real-world use cases.

Accomplishments

- Adapt BBC speech-to-text software and release software for use with U.S. public media/oral history collections ✓
- Begin integration of controlled vocabularies ✓
- Process audio from Studs Terkel Digital Archive, American Archive of Public Broadcasting, and other archival/public media partner organizations ✓
- Begin developing crosswalks with the Digital Public Library of America (DPLA) and Public Media Platform (PMP) ✓
- Begin integration of speaker differentiation ✓

One unexpected outcome of the grant period was the inclusion of automatic speaker differentiation in our speech-to-text output. Our next step, through an IMLS National Leadership Research Grant in collaboration with WGBH and the Library of Congress' American Archive of Public Broadcasting, will be to bridge the gap between speaker differentiation and speaker identification. That work will be conducted in partnership with the HiPSTAS project out of UT-Austin, and will entail the transcription of 40,000 hours of American Archive content, the open-source release of our Kaldi speech-to-text models, and the creation of an audio "fingerprint" through programmatic identification of audio qualities beyond transcripts (i.e. speaker identities, emotion, applause) combined with crowd-sourced improvements to machine transcripts and audio fingerprint training data. (See "Long Term Impact" for more.)

While we have not yet built technical crosswalks with the Digital Public Library of America or Public Media Platform, we are in contact with both organizations. As of November 2015 we have partnered formally with DPLA to offer discounted services to its 1,600 partner institutions. We also built an initial prototype for a PMP-Pop Up Archive integration that, if implemented, would enable full-text, timestamped search for PMP

partner content (PRX, NPR, APM, PRI, PBS), and might bear revisiting now that PMP has ceased running as a standalone organization and is functioning as a part of NPR.

Audiences

Pop Up Archive has over 3,000 individual or small/medium business users (up from 800 in February 2013, at the time the proposal for this grant was submitted), including over 50 larger institutional partnerships. Through our partnership with the American Archive of Public Broadcasting, in 2016 we will be indexing 68,000 audio files from over 100 public media stations across the United States.

Between September 1, 2014 and November 1, 2015, the Pop Up Archive website (www.popuparchive.com) saw 615,948 pageviews, 206,474 total sessions, and 156,456 unique users. (Monthly average traffic is 43,835 pageviews, 14,734 sessions, and 11,449 unique users.) By contrast, the site saw 74,174 pageviews and 14,014 unique visitors between its launch on November 21, 2013 and July 16th, 2014.

Evaluation

Pop Up Archive has been evaluated through a combination of quantitative metrics (quantity of audio processed, accounts created, paying users, institutional partners, website traffic) as well as qualitative metrics such as the range and caliber of our partners, press coverage, and feedback from our NEH Implementation Grant advisory board, which we convened in August 2015 for an in-person meeting.

As of November 24, 2015, Pop Up Archive has processed 1,165,406 minutes (almost 20,000 hours) of audio. This includes public audio as well as privately stored audio. Pop Up Archive is responsible for the creation of hundreds of Internet Archive audio collections housing over 10,700 items that have been viewed over 500,000 times. In addition to over 3000 individual and small/medium business Pop Up Archive accounts, we are supported through enterprise partnerships with over 50 organizations and institutions. We continue to process audio for the Studs Terkel Digital Archive and will soon begin processing material from the 40,000 hour American Archive of Public Broadcasting in partnership with WGBH through an IMLS National Leadership grant.

In addition to user stories that can be accessed at www.popuparchive.org, we appended media coverage and reviews in the appendix of this report.

Continuation of the Project

Pop Up Archive will live on beyond the scope of the NEH Implementation Grant period. Since receiving the grant, we have gained name recognition and created new revenue

streams through partnerships with major public media stations (KQED, WGBH, KCRW, KUOW), libraries (NYPL, Library of Congress), archives (Stanford Hoover Institution, Princeton Theological Seminary, Wisconsin Historical Society), and radio shows/podcasts (This American Life, The Moth, BackStory with the American History Guys).

Our current focus is on improving/standardizing our existing automatically generated metadata and facilitating the implementation and improvement of time-stamped automatic transcripts through software plugins, widgets, embeddable players, and crowdsourcing efforts.

The Pop Up Archive service generates revenue that contributes to operating expenses and our ongoing development of a sustainable business model. We anticipate being financially sustainable by September 2016, supported by fees for the Pop Up Archive tools and services that that our partner institutions either budget for directly or include as part of grant funds for digitization and more research-focused audiovisual cataloging and access efforts.

One ongoing challenge for Pop Up Archive is the development and maintenance of a hybrid business model, through which expenses (primarily Pop Up staff salaries, research and development, and audio processing) are covered by a combination of direct revenue for monthly recurring subscriptions and one-time processing as well as income from grant-funded partnerships with non-profit institutions such as WGBH and the New York Public Library (NYPL).

The 2014 National Digital Stewardship Agenda notes the need to “engage and encourage relationships between private/commercial and heritage organizations to collaborate on the development of standards and workflows that will ensure long-term access to our recorded and moving image heritage.” These partnerships are critical in order to move the needle on audiovisual access issues of national significance. We are eager to continue building such relationships so that the innovations in technology, workflows, and data analysis advanced by the private sector can be fully and sustainably leveraged for U.S. public media and cultural heritage organizations.

Our initial goal was for Pop Up Archive to sustain itself independent of grant funding. Without ruling out that possibility, we have also experienced firsthand the realities of current digital preservation and access standards and methodologies for audiovisual collections (whether analogue, digitized, or born-digital), as well as the integral role that grant funding can play in helping leading institutions (i.e. WGBH and NYPL) partner with

initiatives like Pop Up Archive to create scalable, widely applicable best practices for opening up these collections. We recognize that our collaborative role in these grant-funded initiatives — and our ongoing work to educate and connect the stakeholders in these communities — is critical not only for the immediate survival of Pop Up Archive, but also to foster an ecosystem of capable technology, informed practitioners, and clear approaches so that tools like Pop Up Archive have a clear value proposition and are increasingly expected components of audiovisual preservation and organization efforts.

Long Term Impact

We see Pop Up Archive's potential for long-term impact through its increasing use by newsrooms and audiovisual archives, producers, podcasters, and librarians alike. Pop Up Archive's software is used to make audio source material searchable, its web-based tools are used to correct machine transcripts and make them accessible to new audiences, including the hearing impaired, and its embeddable widgets are used to enable search engines to index audio, to help audiences pinpoint exact words or phrases within audio files, and to facilitate the sharing of audio across multiple platforms and networks.

Pop Up Archive's growing partnerships result in non-federally generated income in the form of subscription and one-time fees paid by customers. The subscription fees paid by our customers are sometimes subsidized through federal and non-federal grants, e.g. a Knight Foundation Prototype Grant (which is funding collaboration between The Moth and the NYPL) and an IMLS National Leadership Research grant (which was awarded to WGBH to unite human and automated methods for cataloging audiovisual content, and for which we are the major partner).

We are particularly excited by the potential of human-computer collaboration for audiovisual collections at scale. PRX has been a leader in digital audio technology since its inception in 2003, and we started Pop Up Archive in 2012 because no solutions existed for cataloging and organizing spoken word audio through technological means. Now that Pop Up Archive is recognized as a leader in processing and creating metadata for audiovisual collections at scale, we are inspired to see growing interest in the potential for combining automated audiovisual cataloging methods with crowdsourced refinements and contributions. We will be contributing to the development of crowdsourcing platforms designed specifically for for time-based media through our work on the IMLS National Leadership Research Grant and Knight Foundation Prototype Grants referenced above as well as through our ongoing educational efforts within the archiving, oral history, and public media communities.

Grant Products

See appendix for screenshots of the primary grant product, which can be accessed directly at www.popuparchive.org, as well as materials that were distributed online and resulted from grant activities.

Our open-source code base is available at github.com/popuparchive.

Appendix

Conference presentations:

- Code4Lib 2015: "Helping Google (and scholars, researchers, educators, & the public) find archival audio," https://youtu.be/gCfpQgXcpTE?t=44m14s

Press:

- Press release announcing Pop Up Archive + DPLA partnership
- Knight Prototype Grant with NYPL + The Moth
- IMLS National Leadership Research Grant with WGBH and the American Archive of Public Broadcasting
- (CBC Spark) "Here's why your phone can't understand your accent"
- (Oxford University Press's Oral History Review Blog) "Using Pop Up Archive for oral history transcription"
- (Free Music Archive's Radio Free Culture podcast): Trapped in the Black Box of Sound with Emily Saltz of Pop Up Archive
- (AIR Public Media Scan) Found Sounds
- (PRI's The World in Words) A California startup tries to capture the elusive spoken word — and make it searchable
- (WILL Public Media) Pop Up Archive: breakthrough in speech-to-text
- (PBCore blog) PBCore data exchange with Pop Up Archive
- (Poynter) Pop Up Archive releases new tools for transcribing audio
- (Knight Lab) Pop Up Archive's Anne Wootton & Bailey Smith on born digital audio, search, and transcriptions
- (BBC Wales Research) Can't type, won't type

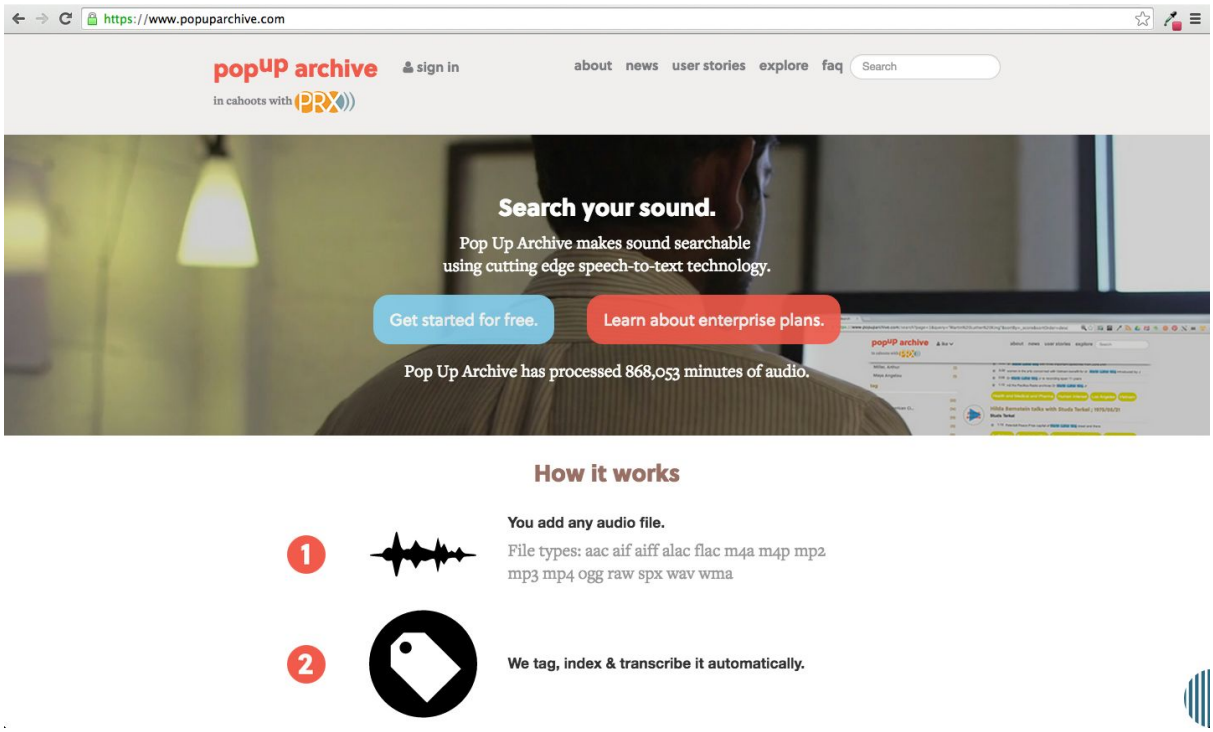
Blog posts/newsletter content:

- How to share moments from audio stories (Pop Up Archive Tumblr)
- Why producers use Pop Up Archive's tools (Pop Up Archive newsletter)
- Is public media ready for machine transcription? A Socratic dialogue. (Medium)
- What are the top challenges facing digital audio in 2015 (part two)? (Medium)
- What are the top challenges facing digital audio in 2015 (part one)? (Medium)
- Introducing the embeddable transcript player (Pop Up Archive Tumblr)
- Archival audio starter kit (Pop Up Archie Tumblr)
- Audio SEO pilot announcement (Pop Up Archive Tumblr)
- Speech recognition for media: it's not all about accuracy (PBS Idea Lab)
- Five unexpected insights from speech-to-text (Pop Up Archive Tumblr)
- How to add auto-tags to WordPress audio (Pop Up Archive Tumblr)
- UC-Berkeley Free Speech Movement Hackathon (Pop Up Archive Tumblr)

- [Introducing our speech-to-text partners](#) (Pop Up Archive Tumblr)
- [NEH Digital Humanities Startup Grant White Paper](#)

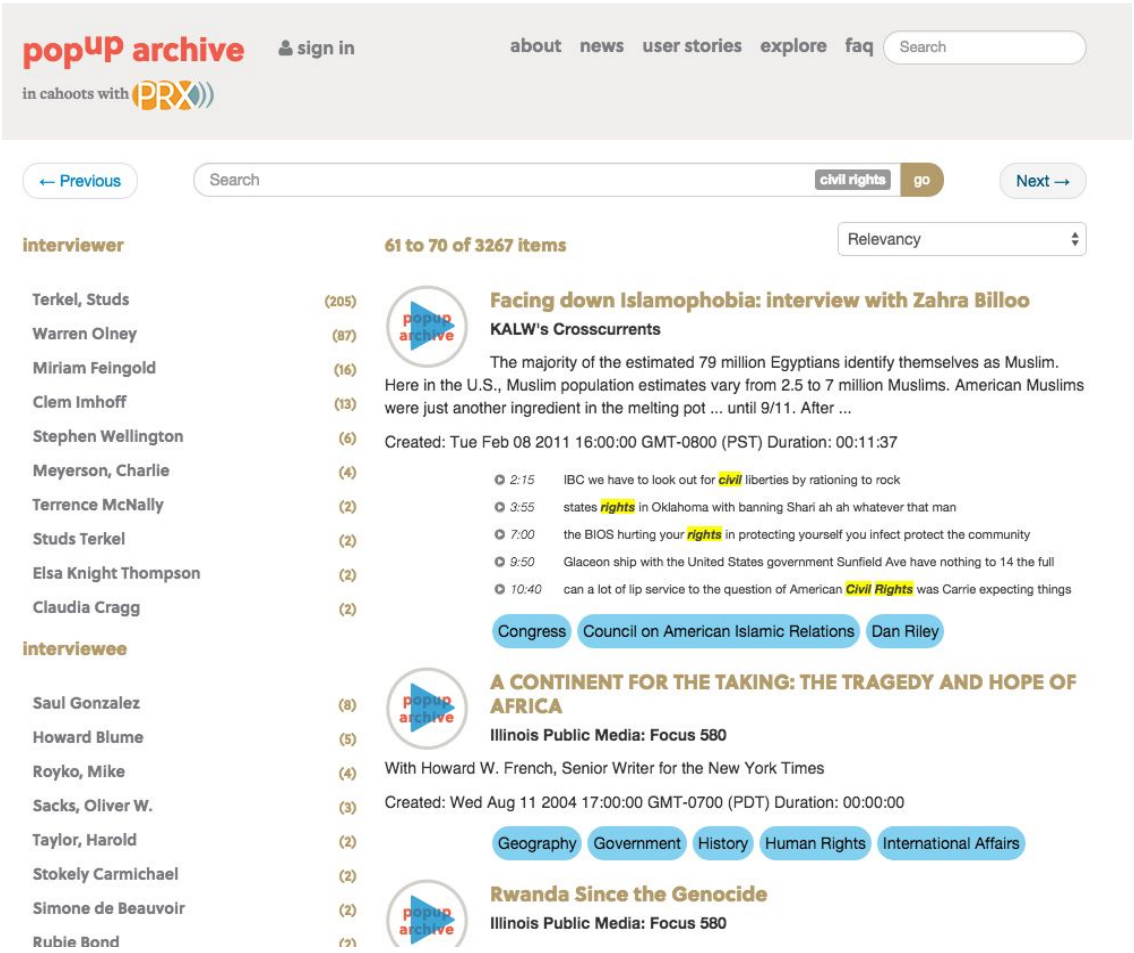
Screenshots:

1. Pop Up Archive homepage

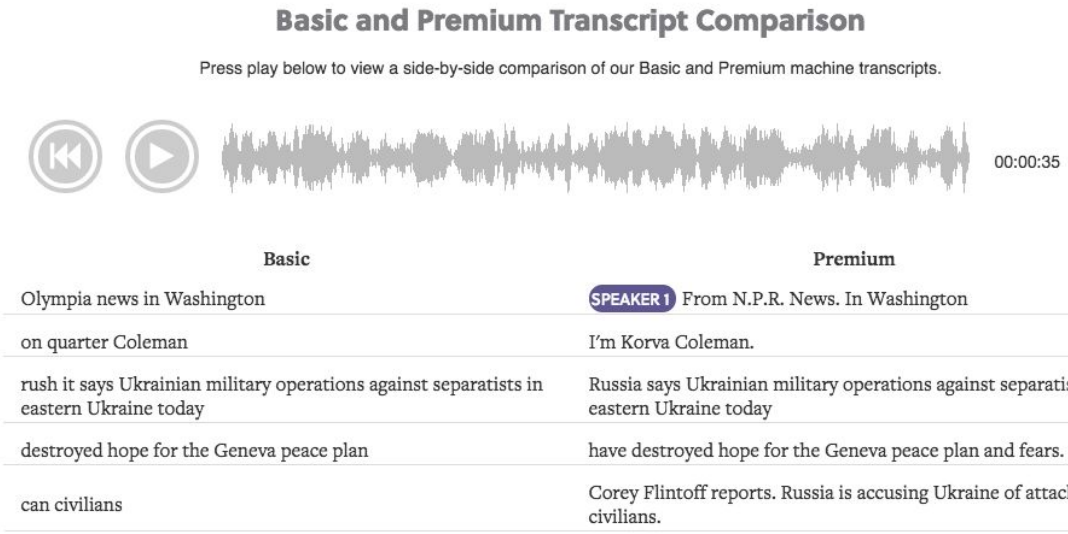


2. “Explore” page (popuparchive.org/explore) featuring over 10,000 public, machine transcribed audio items

3. Time-stamped search results (i.e. for “civil rights”) return relevant audio items by searching metadata and full-text transcripts



4. Comparison of basic and premium speech-to-text software



Basic and Premium Transcript Comparison

Press play below to view a side-by-side comparison of our Basic and Premium machine transcripts.

00:00:35

Basic	Premium
Olympia news in Washington	SPEAKER 1 From N.P.R. News. In Washington
on quarter Coleman	I'm Korva Coleman.
rush it says Ukrainian military operations against separatists in eastern Ukraine today	Russia says Ukrainian military operations against separatists in eastern Ukraine today
destroyed hope for the Geneva peace plan	have destroyed hope for the Geneva peace plan and fears.
can civilians	Corey Flintoff reports. Russia is accusing Ukraine of attacking civilians.