

# Inteligencia artificial y datos masivos en archivos digitales sonoros y audiovisuales

*Perla Olivia Rodríguez Reséndiz*  
Coordinadora



**Q335** Inteligencia artificial y datos masivos en archivos digitales  
**I57** sonoros y audiovisuales / Coordinadora Perla Olivia Rodríguez  
Reséndiz. - México: UNAM. Instituto de Investigaciones  
Bibliotecológicas y de la Información, 2020.

xviii, 182 p. - (Tecnologías de la información)

ISBN:

Investigación realizada gracias al programa

DGAPA - PAPIIT IT400118.

1. Inteligencia artificial - Procesamiento de datos. 2. Internet  
de las cosas. 3. Archivos sonoros. 4. Big data. I. Rodríguez  
Reséndiz, Perla Olivia, coordinadora. II. ser.

Diseño de portada: Oscar Fernando Arcos Casañas

Imágenes:

Envato Elements

(<https://elements.envato.com/es-419/>)

Primera edición, 2020

D.R. © UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Ciudad Universitaria, 04510, México D.F.

Impreso y hecho en México

ISBN: En trámite

Publicación dictaminada

# Using Computational Tools and Experts to Improve Access to Digital Media Archives

KAREN CARIANI  
DAVID O IVES

*Executive Director WGBH Media Library  
and Archives Boston, MA*

## INTRODUCTION

This paper describes projects involving machine learning (ML) and artificial intelligence (AI) to create descriptive metadata for the American Archive of Public Broadcasting (AAPB), and the importance of archives and libraries collaborating with experts in the AI community. There is a challenge to improve access to audio visual collections, and especially the ever-growing digital collections. These collections represent our cultural heritage and need to be seen, used, and available. However, most cultural heritage archives and libraries are not acquiring more funding to manage the new digital collections, just more files and items, and audiovisual (A/V) collections in particular pose special challenges. Digital A/V collections offer great opportunities to use the new technologies for search and access. How can archivists continue to make these important A/V collections available through all the wonderful platforms the new digital technology has to offer? It is necessary to have good descriptive metadata to enhance computer search discoverability or better search tools that can use sound and images. An opportunity exists to collaborate with other professions and disciplines that are expert in computational tools, machine learning and artificial intelligence (AI). There is enough benefit to both areas of expertise, that working together to improve the tools, more can be accomplished. The AI and

machine learning community need good data sets. The archive community has datasets that need to be improved. By working together, the datasets can be improved by improving the quality of the meta-data and thus making the data set more viable for analysis. There is a saying that goes: If you want to go fast, go alone, if you want to go far, go together. This paper will outline AAPB past projects, and AAPB's current work with computational linguistics at Brandeis University through an Andrew W. Mellon Foundation funded project.

## BACKGROUND

WGBH is Boston's public television and radio station. WGBH produces fully one third of the content broadcast on the PBS network nationally. WGBH has been broadcasting since 1951 with radio and 1955 with television. The WGBH archive holds about 750,000 items, mostly audio, video or film materials. This does not include the born digital media files which are now created daily. The programming produced includes public affairs, science programming historical documentaries, children's programming and "how to's" like cooking, woodwork, home improvement. This paper focuses on the American Archive of Public Broadcasting (AAPB) collection which is a collaboration between the Library of Congress and WGBH.

The AAPB goal is to coordinate a national effort to preserve and make accessible, as much as possible, rights permitting, historically significant public tv and radio programming. The AAPB is a digital archive with a website at <http://americanarchive.org>. Users anywhere in the US can access a wide range of historical public television and radio programs from the late 1940s to the present. The AAPB supports current stewards of the materials and facilitates the use of historical public broadcasting by researchers, educators, students, and others. The Library is primarily responsible for the long-term preservation of the digital files. WGBH spearheads access and outreach, and together share overall governance, policy, collection development, ingest and access and rights decisions. As an aggregator of content, AAPB hopes to provide a centralized web portal of discovery for public media materials. The collection is growing. Access is for research, educational,

and informational purposes only. Due to rights restrictions, half is available through the On-line Reading Room anywhere in the US. The entire collection of over 100,000 items is available for viewing on location at the Library of Congress and WGBH.

Given the collection, there are, of course, problems and challenges. The material comes from over 100 sources around the US. It has great variety, which is a treasure and a challenge. The content of the collection varies from a single speaker, like a news announcer at a desk with a single microphone, to a man on the street with a heavy accent and background noise, to a musical performance, to foreign language, and potentially all mixed into 1 program. It comes from all regions of the country with people speaking in different accents, dialects, speech patterns, and speed of talking.

The metadata is also variable. Often coming from local public media stations with no archivists or librarians on staff, much of the content has limited metadata. And the AAPB is growing annually with each new collection adding thousands of files, increasing the problem, a common challenge with many audiovisual archives. For example, just looking at a digital audio file tells a researcher nothing about the content. The audio file needs to be played and listen to for possibly an hour or more to determine the content and add appropriate descriptive metadata.

Some items have a transcript which can provide rich data. The transcript can be indexed, and the terms searched. It also allows for an easy read or scan of content, and roughly a time location within the file to reach topics of interest. Obviously, in order to find things in the collection it needs to be described and catalogued, but human cataloguing is too slow, given the volume of material constantly being added. Can machines can do it faster and describe everything quicker with the same accuracy?

Most current search engines (machines) use words to locate an item. To add those appropriate texts and terms, knowledge of the program is necessary. There are thousands if not millions of footage, frames, and audio tracks. It would take a human many years to fully catalogue and describe the AAPB collection. There is a great opportunity to utilize machine learning and computational tools to help

create metadata to improve discoverability. Creating text from the audio, using Natural Language Processing (NLP) tools is an easy start, in addition to identifying people and sounds by matching audio sound waves, and to use crowd sourcing to help correct machine originated errors and verify the machine results.

## PROJECT RESULTS

The initial AAPB project collaborated with Pop Up Archive and University of Texas (UT), Austin on an IMLS project to 1) create transcripts using a speech to text tool (Kaldi) for 68,000 items in the AAPB collection, 2) to use other NLP tools to pull out named entities and locations from transcripts and 3) to test use of audio identification tools to identify key speakers in the collection working with UT Austin, High Performance Sound Technologies for Access and Scholarship (HiPSTAS). Crowdsourcing games were created to help correct or fix the computer-generated transcripts.

The results of the audio fingerprinting project with HiPSTAS were mixed. The goal was to use identified speaker soundwaves to find identical sound waves in the AAPB collection to identify speakers, music, applause, laughter, etc. HiPSTAS had a sample size of 4,000 hours (from the AAPB 68,000 item collection). WGBH identified 10 speakers: Hillary Clinton, Bill Clinton, Ronald Reagan, Julia Child, James Baldwin, Malcolm X, Martin Luther King Jr., Lyndon Johnson, Richard Nixon, and Gloria Steinem. These key people were probably in the AAPB collection, and nationally identifiable. The end product was to create a reference database of sound waves, or sound fingerprints, for speakers that could be available for others to use.

The process was the following. WGBH gave HiPSTAS a set of files with the 10 identified speakers. Humans identified the time code where these named people existed in the audio files for a set of 103 hours of content. Supervised learning was used to train the machine to identify those people soundwaves. A specific machine learning model needed to be built for each speaker. There was another set of files that might possibly have those speakers, to see if the machine

tool could get a sound wave match and identify the speakers. A random batch of about 1000 hours was provided to see if the machine could identify the speakers unsupervised.

After 3 years, only 1 speaker model was created for Bill Clinton. HiPSTAS tools were indeed capable of identifying Bill Clinton for a key set of files. However, the compute power and resources for the machine learning alone, in addition to a human tagging a speaker in a sampling of items, building an algorithm specifically for 1 speaker, and then to search for the speaker across a larger set, is not yet a feasible way of identifying speakers in 100,000 items of local tv and radio programs.

However, useful tools were created as a result of this project. (Figure 1) There is a workflow chart, an audio tagging toolkit, a dockerized Jupyter notebook for machine learning tools, 2 notebooks to show how to use the tools using the Bill Clinton model, 2 notebooks to detect music, and other models created. Each effort forward creates more opportunities even if the original goals are not completely fulfilled.

*Figure 1.*

**Tools created for audio fingerprinting**

- Workflow charts are available here: [https://github.com/hipstas/aapb-data/tree/master/Workflow\\_flowcharts](https://github.com/hipstas/aapb-data/tree/master/Workflow_flowcharts).
- The team developed tools (available and described at <https://github.com/hipstas>) including an Audio Tagging Toolkit, Audio Labeler, and the Kaldi + Pop Up Archive Docker Image.
- HiPSTAS created a Dockerized Jupyter notebook environment with pre-installed audio machine learning tools, a package we were calling the Audio ML Lab : <https://hub.docker.com/r/hipstas/audio-ml-lab/> and <https://github.com/hipstas/audio-ml-lab>.
- Two notebooks to show how to use these tools using a Bill Clinton model here: [https://github.com/hipstas/sida/tree/master/Bill\\_Clinton\\_template](https://github.com/hipstas/sida/tree/master/Bill_Clinton_template).
- They also created two notebooks for detecting music: [https://github.com/hipstas/sida/tree/master/Music\\_template](https://github.com/hipstas/sida/tree/master/Music_template); and for detecting applause: [https://github.com/hipstas/sida/tree/master/Applause\\_template](https://github.com/hipstas/sida/tree/master/Applause_template).
- Other models are here, including those that HiPSTAS produced specifically from AAPB and those they used to create the Kaldi instance: <http://xtra.arloproject.com/datasets/>.



AAPB also worked with Pop Up Archive to create speech to text transcripts using the open source tool Kaldi, that was developed by the BBC. Having a transcript for a digital media file allows search within the transcript and the ability to highlight the found words. The audio/video can be synced to a timecode in the transcript, allowing users to navigate directly to content of interest. The key words can be indexed for search engines to find related items. The goal is to create transcripts at scale, pushing a large number of files through the tool without first categorizing the sounds in the program.

Many of the videos in the collection start with bars and tone, or have music, or other sounds that Kaldi tries to transcribe into text and because of this, the transcript is not accurate. Kaldi tries to turn the sounds incorrectly into words. In addition, accurate transcripts are dependent on audio quality, speaker accents, background noise, etc. Given that the AAPB collection is from 100 different local tv and radio stations across the country, the variety of audio and audio quality varies widely. WGBH used 3 tools to allow the public to help fix and correct the English transcripts and add additional metadata.

The AAPB used Zooniverse, a crowd sourcing platform, to utilize the public to transcribe the credit information on the screen that a speech to text tool could not transcribe because there is no audio. This is valuable metadata which includes copyright notice, date of broadcast, validating program title, producer names, etc. The tool was called Roll the Credits and within 3 months, because of the large user base of Zooniverse, we had completed a data set of 917 program credits which included 5 verification passes. However, to set the tool up, we needed a screen grab of rolling credits that became 29, 206 frames, which took significant human time.

WGBH also built a game to fix the transcripts, called FIXIT. There are actually 3 games in one – identify errors, suggest fixes, and validate fixes. The pipeline to output a finished corrected transcript was too slow. It took too long, and too many players for 1 transcript to be completed, corrected, and validated. The most successful tool was FIX-IT+ which uses a transcript editor tool that New York Public Library



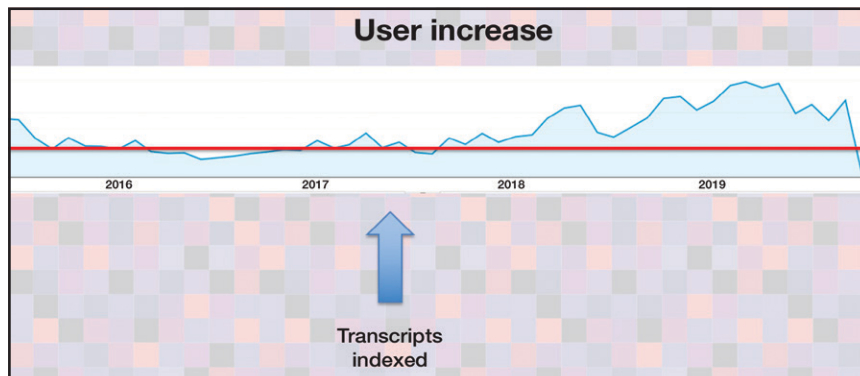
developed. The tool is a straight editor rather than a game and users are able to more effectively and efficiently correct transcripts. Two people have to agree on a fix for it to be validated.

A small percentage of transcripts have been corrected through this effort. The crowdsourcing campaign has engaged the general community and users intimately with the archive but has not corrected transcripts at scale. The most successful effort occurred when a vendor challenged AAPB contributors to push their communities to fix more transcripts. For every transcript corrected, the station would get a free tape digitized, up to 100, but they had to reach a 20 transcript threshold first. As a result, several stations engaged their volunteers with the challenge and there are now 530 completed transcripts. On the one hand, stations jumped on the challenge and people got involved, however, there are still thousands of transcripts to correct. Crowdsourcing corrections is not an efficient method to correct volumes of transcripts.

Once the transcripts have been verified, the JSON transcripts are stored in the AAPB's Amazon S3 account and indexed for keyword searching on the AAPB website. The transcripts will be made available alongside the media on the record page and can be played like captions within the video player. They will be able to be harvested via an API to be used as a data set for research. Researchers will be able to use the AAPB collection as a data set and start analyzing trends from programming over the last 60 years such as how language has been used in reporting, how it has changed over the years, and how it is different in different areas of the country. But to use this collection effectively as a data set, we need accurate good descriptive metadata.

The uncorrected transcripts were indexed in April 2017. There was an overall general increase in users coming to the site. (Figure 2) Of course, this is one of many factors that could have cause an increase in users, but it does show that search engines were driving traffic to the site even with inaccurate transcripts. The increased traffic is due to speech to text transcripts enhancing the metadata and thus discoverability and search. The question is whether users were actually finding what they were looking for.

Figure 2.



There is a lot of variety in the content and many of those indexed words are not accurate depiction of the content. For example, Kaldi does not recognize a foreign language. It tries to take the foreign language sounds and assign them English words that match the sounds. Understandable that there are tools for Spanish language transcription and translation, but many of the programs have mixed language and it is unknown which items or files have which languages. It would be helpful if a foreign language could be identified automatically, and then skipped over, instead of spending the effort to transcribe non-English into English. The same is true for music or other sounds. It is not very helpful and a waste of computational power and time to attempt to turn these sounds into English words. A helpful tool would characterize the language as not English and skip over it.

Our results from the Pop Up Archive version of Kaldi were for English transcripts, with an average of 81% accuracy, across the collection including all the mistakes around music and foreign languages. Other problems are punctuation errors and speakers with strong accents that were not well transcribed. Common errors are personal names. The project success is the creation of transcripts for the initial collection of 68,000 items which took 6-8 months to process. The quality of the transcripts varied. Specifically, the accuracy of programs from a single source single speaker, formal announcer was about 95% accurate (no accents, one speaker). However, transcripts

for a television program from Mississippi with a strong Southern U.S. accent was only 55% accurate. Data of named entities and locations from the Pop Up tools, was not useable. By sampling entities for 101 recordings, the conclusion was the number of tags, usefulness of those tags, and confidence in the accuracy of those tags would not significantly enhance the metadata records to warrant building the necessary new workflows and technologies. Many of the topic tags were not specific enough to be useful, such as “United States” “Entertainment and Culture” and “Human Interest.”

Near the end of the project, Pop Up Archive was bought by Apple, and shut down their Github account. The Kaldi tool being used was forked into the WGBH Github open account that anyone can access, but it is not the sophisticated trained tool being used for AAPB transcripts. WGBH has since dockerize the tools and is feeding programs through a Kaldi workflow on a regular basis. The accuracy these transcripts is about 56%, considerably less than the Pop Up trained tool of 81% average accuracy. WGBH continues to create transcripts using the lesser version of Kaldi for new content added to the AAPB. Even an inaccurate transcript gives some data about the program content. The crowd sourcing campaign to fix the output transcripts for accuracy is taking a long time. Another solution is needed to expedite the process, like perhaps trying to train the tool or using better machine learning tools.

Perhaps a commercial service such as Amazon speech to text is a bit more accurate, but 1) with this volume it gets expensive and 2) with this variety it is actually not that much more accurate overall. And this is not actually helping the open tool to learn and get better. Tools off the shelf, output the lowest common denominator and need programming to extract specific data beyond the common base. For example, Amazon web services facial recognition is probably not going to recognize many of the people in the AAPB collection. Better tools in the open source space, that are easy and affordable for archives to use is a much better long-term solution.

Now the AAPB is partnering with computational linguists at Brandeis University Computer Science department to develop workflows that utilize NLP and machine learning tools to extract key metadata for audio visual collections. Our goal is to begin communication

between the archives and computer science communities and develop open source tools that can be used by other cultural heritage organizations. Some key issues have been identified that would help with discoverability and access of the collection such as forced alignment of transcripts to time stamps, time stamping bars and tone at the beginning to improve the user experience, identify music and foreign languages to improve speech to text output, and identifying text on the screen and outputting it as data using Optical Character Recognition (OCR).

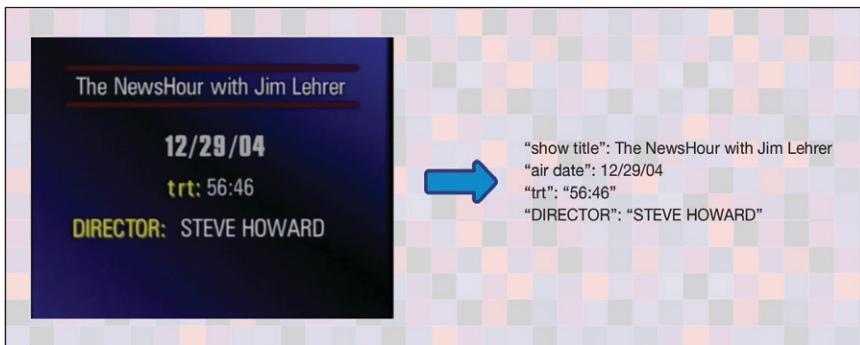
The data set focus is 30 years of NewsHour programs. The first step is to create key data that is verifiable such as program slates that can verify title, date aired, and producer; lower thirds identifying people on the screen; and credits at the end giving us production staff, participants, and copyright information. (Figure 3) That text can be used as metadata. If possible, the tools can also use the speaker lower third identification and verify against an announcer introducing the speaker, and then use facial recognition to find the same person again within the same video file. Eventually the tools can help with program type identification, better named entities, and better Kaldi output. For example, there is a single person at a desk therefore this might be a news broadcast.

Figure 3.



There is not one single tool that can create all this metadata, but rather a series of tools where the output of one tool becomes information for the next tool to refine the specific characterization or data extraction desired. Just as each speaker sound wave needed a specific algorithm or 'model' to teach the machine to identify a speaker, a separate tool is needed to characterize the item or attribute, decide which tool is needed to isolate the data, and pull out the informative text. Time based media complicates this process. What makes media different and harder is that the information you are trying to capture is moving across frames or tracks, and there is a huge matrix of pixels across each frame or image. The data on this frame is needed as text metadata. Text on a video frame, however is an image, not text. It can identify air date, director, running time and show title. But first the machine has to find the slate frame within the video file, get rid of everything that is not a slate, isolate the frame. Bounding boxes then find the text and character recognition for the text in the bounding box is performed, and it needs to recognize the words and what they are. For example, to pull the slate information off a video frame and make it useable metadata, the workflow is 1) a tool finds and isolates the slate video frame 2) the text location is identified on the video frame 3) the text information on the video frame is put through an OCR tool to create text output 4) the output is in a form that is understandable and correctly labelled. (Figure 4)

Figure 4.



The Brandeis team has created CLAMS. CLAMS, Computational Linguistics Applications for Multimedia Services, is a workflow tool that can plug in the appropriate tool for the appropriate attribute, once the file has been characterized. (Figure 5 and 6) This pipeline, or noodling tools together, will be used to create workflows of tools. One problem across all these tools is that the developers create their own interchange format and the tools from different sources don't talk to each other or exchange data. CLAMS uses a standard format (MMIF) to allow the tools to be used together. A common language that all the tools can understand.

Figure 5.

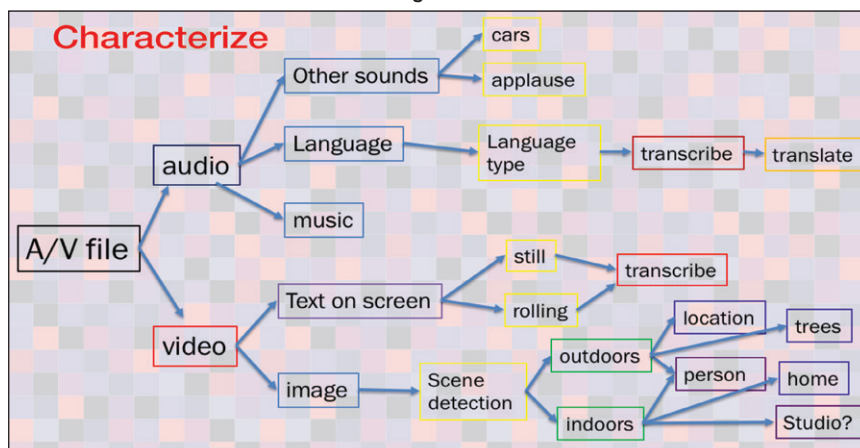
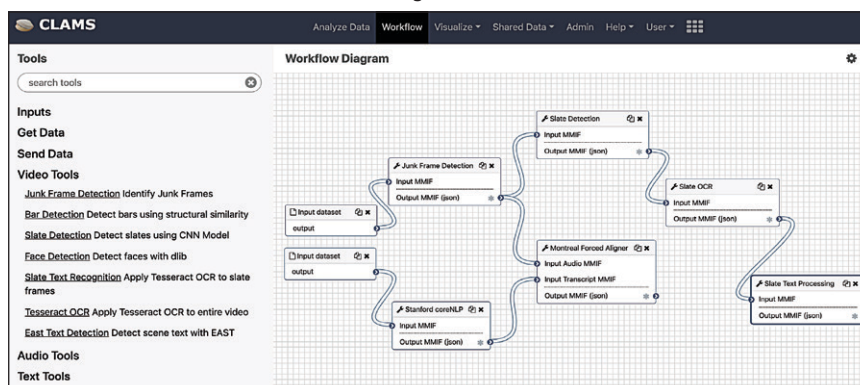


Figure 6.





The Brandeis team will write the algorithms, create the MMIF format, and have begun to build workflows and tools that have useful output for the AAPB. They are interested in developing an open source pipeline that will work with a variety of tools keyed to definitive specific tasks that A/V archivists might need - to enable an archivist to create a workflow by dragging and dropping tools from a tool shed to enable certain data outputs. This project is generously funded by the Andrew W. Mellon Foundation.

## CONCLUSIONS

With the volume, variety and complexity of digital audio-visual collections, using machine learning and AI tools effectively, will take better knowledge of the tools than most archivists currently have to pipeline the tools into workflows. Libraries and archives have great data sets and computational scientists need large datasets to test tools, build tools and analyze trends. Libraries and archives need tools to improve the metadata thus improving the datasets. A collaboration between computational scientists and archives to improve computational tools, AI, and machine learning for better data in archives and libraries, would benefit both communities.

Computational linguists and AI experts have been working with and building tools for years that are just now being used by libraries and archives. AI experts can help build workflows and algorithms that will improve the tools archives and libraries need and make sure the output is useful to improve collections. Libraries and archives need to be able to talk to computer scientists using their definitions and terminology to understand the tools being built. The elasticity of the human brain to recognize variety and perform many tasks at once, is not yet there for machines. There is actually quite a lot of human effort that goes into machine learning or building the training for the machines. In order to use these tools for complicated collections they need to be adapted and trained. Archivists should work collaboratively with computer scientists toward better affordable, and open tools for archives and libraries.



## BIBLIOGRAPHY

Cariani, Karen, & Kaufman, Casey Davis. "Crowd Sourcing Metadata for Time Based Media in the American Archive of Public Media." In *Participatory Archives: Theory and Practice*, edited by Edward Benoit III and Alexandra Eveleigh, 95-101. London: Facet Publishing, 2019.

\_\_\_\_\_. "Improving Access to Time-Based Media through Crowdsourcing and CL Tools: WGBH Educational Foundation and the American Archive of Public Broadcasting." In *Proceedings of the CLARIN Annual Conference 2018* (2018): 66-71.

Ide, Nancy, Pustejovsky, James, Suderman, Keith, & Verhagen, Marc. "Enhancing Access to Media Collections and Archives Using Computational Linguistic Tools." In *Proceedings of the Workshop on Corpora in the Digital Humanities* (2017): 19-28.

Rim, K., Lynch, K., & Pustejovsky, J. (2019, June). *Computational Linguistics Applications for Multimedia Services*. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 91-97).

## URL REFERENCES:

WGBH Kaldi Github: <https://github.com/WGBH/kaldi-pop-up-archive>.

AAPB website: <http://americanarchive.org>.

FixIT: [http://fixit.americanarchive.org/?utm\\_source=aapb\\_help-us\\_promo1&utm\\_medium=website&utm\\_campaign=help-us\\_from\\_website](http://fixit.americanarchive.org/?utm_source=aapb_help-us_promo1&utm_medium=website&utm_campaign=help-us_from_website).

FixIT+: <http://fixitplus.americanarchive.org>.

Zooniverse: [https://www.zooniverse.org/projects/sroosa/roll-the-credits/?utm\\_source=aapb\\_help-us\\_promo3&utm\\_medium=website&utm\\_campaign=help-us\\_from\\_website](https://www.zooniverse.org/projects/sroosa/roll-the-credits/?utm_source=aapb_help-us_promo3&utm_medium=website&utm_campaign=help-us_from_website).

***Inteligencia artificial y datos masivos en archivos digitales sonoros y audiovisuales.***

Instituto de Investigaciones Bibliotecológicas y de la Información/UNAM. La edición consta de 100 ejemplares. Coordinación editorial, Israel Chávez Reséndiz; revisión especializada, Angélica Valenzuela y Valeria Guzmán González; revisión de pruebas, Valeria Guzmán González; formación editorial, Oscar Fernando Arcos Casañas. Fue impreso en papel cultural de 90 gr. en los talleres de Grupo Fogra. Año de Juárez 223. Col. Granjas San Antonio. Alcaldía Iztapalapa. Ciudad de México. Se terminó de imprimir en 2020.