ISSN 2161-3974
Full-Length Paper

Journal of eScience Librarianship 13 (1): e805
DOI: https://doi.org/10.7191/jeslib.805

Journal of eScience Librarianship
putting the pieces together: theory and practice

Responsible AI at the Vanderbilt Television News Archive: A Case Study

Clifford Blake Anderson, Vanderbilt University, Nashville, TN, USA, clifford.anderson@vanderbilt.edu
Jim Duran, Vanderbilt University, Nashville, TN, USA

Abstract

We provide an overview of the use of machine-learning and artificial intelligence at the Vanderbilt Television News Archive (VTNA). After surveying our major initiatives to date, which include the full transcription of the collection using a custom language model deployed on Amazon Web Services (AWS), we address some ethical considerations we encountered, including the possibility of staff downsizing and misidentification of individuals in news recordings.

Received: October 2, 2023 **Accepted**: February 5, 2024 **Published**: March 5, 2024

Keywords: artificial intelligence, AI, audiovisual archives, television news, metadata automation

Citation: Anderson, Clifford Blake and Jim Duran. 2024. “Responsible AI at the Vanderbilt Television News Archive: A Case Study.” *Journal of eScience Librarianship* 13 (1): e805. https://doi.org/10.7191/jeslib.805.

Summary

Our goal is to increase the research value of the Vanderbilt Television News Archive by making our collection computationally tractable. We began by using automated speech recognition (ASR) to create transcripts and then applied AI tools to refine and enhance the textual dataset by organizing the collection and increasing the quality and quantity of its metadata. We are currently testing computer vision tools to extract information from the video stream accompanying the audio track of a recorded broadcast TV signal and to automate metadata production and discovery.

Project Details

Founded in August 1968, the Television News Archive at Vanderbilt University (VTNA) numbers among the longest-running audiovisual archives for broadcast news in the world. Over the course of its existence, the TV News Archive has negotiated multiple technology shifts. In 2016, Clifford Anderson assumed responsibility for the direction of the TV News Archive as part of his Associate University Librarian, or AUL, portfolio at the University. He quickly recognized the need for upgrading several systems in light of newly available computational tools and newly emerging research requests. In 2018, Jim Duran was named the new director and pointed out another problem common to many archives: a growing backlog of undescribed materials. Beyond the backlog, the search interface lacked an open API and did not allow for computational analysis, by then a frequent request from VTNA's researchers.[1]

We agreed that the limiting factor in both challenges was metadata. Writing abstracts, the key metadata field in the VTNA database, is a labor-intensive process that requires training, consistency, and persistence. The VTNA adds 3.5 hours of evening news content every day, including weekends. At its peak, the VTNA employed thirteen staff members but, by 2018, was down to five, only two of whom dedicated themselves full-time to writing abstracts. The VTNA held about 3,000 episodes of news needing metadata before they could be made available to researchers.

Duran set out on an effort to find ways to automate either abstract creation or transcript generation. The primary goal was to eliminate the backlog and make the collection completely available to researchers. If the project could demonstrate how to speed up the acquisition, processing, and description of the collection, then he hoped to find a path toward increasing the daily recording capacity and expanding the VTNA's collection scope.

At the same time, Anderson laid down plans for the creation of a data lake at the university library. A data lake is effectively a repository of data sets that provides a computational environment, including on-demand

[1] An API (application programming interface) connects different software systems and enables them to communicate and share data seamlessly. In this context, an API would allow researchers to write code on their computers that queries the VTNA database in a secure and repeatable manner. From our perspective, an API would need to provide us with logs and reports on activity as well as a method for limiting use per user and session.

cluster computing, to analyze those data. Marcus Weaver, a recently hired cloud engineer at the VTNA, imported the TV News data set into the nascent data lake, allowing staff members to query and analyze the corpus as whole.

In what follows, we outline the key steps that allowed us to transcribe, annotate, and render the VTNA's collection computationally tractable in a data lake.

Voice-to-Text Transcription Project

The goal was to make transcripts and captions available to patrons of the VTNA as well as a data mining resource in the data lake. Many longtime users of the VTNA know the recordings unfortunately do not include captions—the text of spoken words used by people with hearing impairments, or by anyone preferring to read rather than listen to the news. In 2022, the VTNA partnered with Vanderbilt University Library administration and the College of Arts and Science to utilize ASR to generate transcripts and captions for 62,000 hours of recorded television news from August 1968 to June 2022. This project was partially funded by a budgetary surplus in the last quarter of fiscal year 2022-23. We had to complete the project in roughly four months before the start of the next budget cycle.

VTNA staff recognized the importance of captions and transcripts for a long time, but developments and improvements in automated tools made this project feasible. Working with Clifford Anderson, Duran devised a plan that took advantage of our existing usage of Amazon Web Services (AWS), which offered the ability to immediately begin transcribing with current operational funding. He created a workflow that used four Python scripts running on three different computers to request 3,000 to 5,000 transcripts at once. When those transcripts finished, Duran would order another batch until he completed 89,000 transcripts. The team decided on this approach, rather than using an ASR vendor such as Trint because there was no need for segmentation or human interaction with the transcripts. Trint works best with a person interacting with each transcript, whereas the goal of this project was to automate the entire process to finish before the funding deadline.

The AWS Transcribe service uses ASR to generate transcripts of the audio tracks in the digital video files. Out of the box, the service provides a proprietary language model based on commercial data, like call center recordings. Duran wanted to increase the accuracy of the transcripts by applying a custom language model. Amazon allows users to feed the Transcribe service examples of text that resemble the desired output of a transcript job to improve the accuracy through machine learning. A sample needed to be 250,000 to 500,000 words of near-perfect text, also known as a training set.

To get these near-perfect sample transcripts, we serendipitously discovered Vanderbilt University had just licensed 3Play Media, a professional transcription service that employs transcriptionists for 99% accurate transcripts. Susan Grider, then the Administrative Manager of VTNA, worked quickly to create a new account for the VTNA, and Duran ordered approximately 75 transcripts for each American Presidential administration from Richard Nixon to Donald Trump. He decided to split the collection by president because of the likely transition of frequent names in the news.

Regarding frequent names, 3Play Media allowed users to provide transcriptionists with notes on the spelling of tricky domain-specific terms. Duran worked with Dana Currier, Metadata Specialist and Student Supervisor, to dig into the VTNA's existing database that included the proper spelling of names and places featured in the news. Once they determined a strategy, Duran wrote a new Python script that found the top 600 names in the news for each presidential administration. That was formatted into a PDF and shared with the transcriptionists at 3Play Media for even better spelling accuracy.

Once the near-perfect transcripts were returned to the Archive, Duran compiled them into a training set for AWS and created a custom language model. That model was then used on all news recordings for the specific date range.

Named Entity Recognition (NER)

Generating a title for a news segment was a more challenging endeavor in support of which we once again turned to AI tools for help. Each news story and commercial required a title that matched the existing title style, which roughly followed the pattern **Location / Main Subject / [optional: Subtopic]**, for example: "California / Earthquake / FEMA." Titles were of course written by trained abstractors before the introduction of ASR transcripts. Duran needed a new solution that was fast and consistent. Using Python scripts, the Amazon Web Services (AWS) command line interface (CLI), and AWS Comprehend, Duran extracted named entities from the body of the transcript for each segment. As documentation at AWS describes this service, "Amazon Comprehend is a natural-language processing (NLP) service that uses machine learning to uncover valuable insights and connections in text" (Amazon Web Services 2023). A named entity is a person, location, organization, or commercial product, and AWS Comprehend outputs all entities in a list outputted as formatted JSON. Our script takes the list of entities, ranks them based on frequency and concatenates them into a string to match the style of the titles written by abstractors.

The results were satisfactory but not ideal. First, a reporter's name was almost always the first result, because they are typically mentioned several times in the story. Reporters should be identified, but not in the title. Duran turned to Steve Baskauf, Data Science and Data Curation Specialist, for advice. Given a list of known reporters for each network, Baskauf developed a Python function that utilized fuzzy matching to filter the reporters out of the title generation script. Fuzzy matching increased the accuracy of the filter by recognizing common misspellings or spelling variations like John and Jon. Once a reporter's name was identified, that data point was stored in a field for reporters.

The second issue was that some stories were less about an entity and more about a concept or topic, like gun violence, climate change, or foreign policy. These subject terms are not entities; they often will not even be mentioned in a transcript explicitly but are implied in the newsworthiness of the entire news report. Television news follows certain themes, specific aspects of society, that go beyond a news cycle and NER cannot recognize those concepts. We have not yet resolved this issue but are exploring options adding subject headings using auto-classification AI. Nevertheless, the people, places, and other details of each news story are highly valuable, so NER is useful as a tool to recognize the entities.

Creating a Data Lake for Computational Analysis

With the newly-created transcripts and supplied titles, the next goal was to make these data searchable and available for machine learning. In another context, Anderson had already implemented a preliminary data lake for the university library. He recommended adding the transcripts and other show-related metadata to that nascent data lake[2] with the goal of creating a repository of news-related data for use by Vanderbilt University faculty and students. This data solution will be a new resource for data scientists interested in researching events as they are documented in multiple news sources, from periodical literature to broadcast news. The data lake will also make it possible to create and fine-tune machine learning models on news sources.

VTNA receives one to three requests for computational access per month on average, but previously lacked the infrastructure and transcripts for most AI/ML projects. We are excited about the research potential of the newly-created transcripts, especially when combined with the existing database of time-coded, titled, and abstracted news stories and commercials. By merging these datasets, users will have a truly unique and powerful source for studying news media across many decades.

In addition to the textual datasets of transcripts and abstracts, we hope to build in capacity for the data lake to accommodate the study of visual and audio elements of the digital video. For example, a researcher could use machine learning to identify the usage of specific imagery like wildfires, airline accidents, or law enforcement activities, or even more abstruse studies like color schemes or sound effects.

Ultimately, the data lake will provide the environment for supervised and unsupervised machine learning projects at the VTNA. Several of the prospective projects described below assume the data lake as their computational environment.

[2] A 'data lake' describes a big data environment for aggregating data sets of heterogeneous provenance and format. For more information, see Anderson 2022.

Background

The Vanderbilt Television News Archive is like many other audiovisual archives. It consists of a collection of digitized tapes, a catalog that describes the video content, and researchers who request access to material. Unlike many archives, our collection is highly homogenous, consisting entirely of national broadcasts of television news. This is split into two sub-collections: special reports and evening news. The evening news collection demands the most of our time and resources, because unlike the specials, which can be briefly summarized, an evening news report consists of segments of news reports and commercials. In 1973, the VTNA office created a workflow in which abstractors watched an episode of news and completed three tasks for each news segment and commercial break: 1. identify the start and stop time of the segment; 2. summarize the news segment, focusing on people, places and the main topic of the report; and finally, 3. give each news story a title and for each commercial break, listing all the products and services advertised. For nearly fifty years, this metadata creation process continued unchanged; today, the VTNA consists of 1.4 million records, dating from 1968 to present. The TV News Archive operates today essentially as a database resource. Users of the resource interact with a website search page to query the collection's metadata to determine if the collection includes clips of television news relevant to their research topic.

Like most distinctive collections held by libraries and archives, the metadata is key to discovery and use. Without a properly described collection, users won't know what secrets are held within the pages, photos, or in our case, forgotten news stories and TV commercials. But the lack of helpful descriptive text and context is not a problem unique to library databases. The corporate world has similar problems describing its commercial products or summarizing and sorting the vast amounts of data streaming into privately held data warehouses. We all needed tools to help us make sense of the data.

Ethical Considerations

Our project raised several ethical issues, including the possibility of downsizing our workforce, the misidentification of entities that we extract from transcripts, and the potential violation of privacy involved with facial recognition software.

AI Replacing Skilled Labor?

There is a growing concern that artificial intelligence will take work away from people, but this was not a concern for us. When the VTNA sought AI/ML tools to generate transcripts and enhance metadata, the goal was not to replace skilled labor. The VTNA saw the same reduction in staffing that most libraries experienced during the past three to four decades. But that reduction in staff came at the same time as the material being collected by libraries grew in volume and density. We were tasked to do more with less—a problematic challenge. But the growth in information was also accompanied by a transformation from analog to digital, which opened the door for AI/ML tools to assist with the challenge. For the time being, adopting AI/ML tools does not threaten staffing levels because cultural heritage institutions are already short-staffed, and the field is trying to ride the wave of digital information threatening to flood our repositories. So while we do see automation tools as a way to replace labor, we regard AI as a way to create more equitable and sustainable workloads for our staff.

In our case, we even explored the option of deploying a new cohort of workers to tackle the backlog. Not only did we find this option cost-prohibitive, but we also determined the task of writing abstracts cannot be done by temporary, entry level, or outsourced staffing. We found that a fully trained and skilled abstractor could complete one hour of video content in six to eight hours. Additionally, the worker needs to stay on task completing the episode, but they also need breaks to avoid burnout. In essence, abstract summaries require full-time skilled labor and we had too many episodes to complete and not enough funding. Finding computer-based alternatives to human labor was the only option to meet our needs.

That said, we do foresee that artificial intelligence and machine learning tools will affect the type of skills and experience we will seek in new staff members in the future. For example, we recently hired a cloud engineer to assist with automating our workflows and improving our discoverability systems. As AI/ML make it possible to automate repetitive tasks, we expect that moves to these automated systems will free staff members to work in other areas, particularly reference, outreach, and marketing. As AI tools for abstracting, indexing, and summarization improve, we will likely not rehire in these areas. Of course, these tools are not perfect and will need human review, so it will be important to keep at least one expert metadata specialist on staff to review any machine-generated metadata.

The Correct Spelling of Names

Spelling first and last names correctly has been a priority of the VTNA from the very beginning. A reason abstracts were so difficult to produce was because the abstractor would take the time to look up a person's name if it wasn't displayed on the screen. ASR transcripts only use spoken words, and the transcript may not include the person's name at all, because the news network left the person's name out of the script and the speaker was only identified with a screen graphic. So, as we moved to ASR transcripts, we had to accept an increase in inaccuracy and missing name percentages as a consequence.

Weighing the importance of quantity over quality may not be an ethical dilemma, but it certainly plays a key role in adopting AI/ML tools. We had to accept a small increase in spelling mistakes in order to move forward with this workflow. It is no different than the core principle of "More Product, Less Process" (MPLP) introduced by Mark A. Greene and Dennis Meissner, where archivists were encouraged to reconsider the expectations of arrangement and description of archival collections (Greene and Meissner 2005). In both situations, MPLP and AI/ML tool adoption, the goal is to make archival material discoverable and accessible to researchers in an efficient and timely manner. Nevertheless, we believe that AI/ML tools should not be adopted without taking a critical look at their shortcomings. We need to consider our reference instruction and search strategies, then communicate any changes regarding past data collection and description practices to our user community.

Additionally, we need to consider the impact on individuals with non-English names. The ASR models we used were trained with the English language, using sample data from American media. The model will be the most accurate with the common names in the U.S. Non-English names will have more spelling mistakes and should raise concerns of equity and bias. We will need to recognize this problem and prioritize its correction with future projects. Screen reading algorithms, for example, can assist with this problem by identifying any names spelled on the screen.

Facial Recognition

Finally, we have elected to pause some projects due to the inherent ethical privacy concerns. We have contemplated, for instance, developing a prosopography of major figures in the news. The idea was to trace the chronological appearances of public figures from different domains of culture (politics, business, society, sports, music, etc.) across networks. This project could be accomplished with off-the shelf tools such as AWS Rekognition. After consulting with experts at the Internet Archive and the GDELT Project, we elected to suspend this project because of the potential for misidentifying individuals. We also worried about the incidental exposure of nonpublic figures who might appear in the background of news programs, which could lead to a loss of privacy. We would like to resume this project when we have better protocols in place to address these ethical hazards.

Who Is Affected by This Project?

Staff: Metadata Creation

The staff members responsible for metadata were the most impacted by the development of AI/ML tools and their workflows were changed the most. Eliminating the backlog was the driving force for our search for new models. Personnel at the archive were aware of the usefulness of transcripts as an alternative to a summary description. Many researchers have requested transcripts, but the VTNA did not have them. The closed caption stream in a television broadcast is a common source for textual data, but caption streams were not captured by the VTNA. Without access to the source textual data, we explored options for generating text transcripts and captions automatically using automated speech recognition (ASR) software. This software continues to improve in accuracy and efficiency. By 2018, we were satisfied with the accuracy and pricing available by Trint. This product not only used a language model that matched our subject matter, but the user interface was easy to navigate and easy to learn.

With Trint, our metadata department established a new workflow that focused on segmentation and spelling names and places correctly. The new process could be performed by temporary and student workers. A new employee would finish training in two hours, and, within a week, they were working at full speed. Where abstracts took new employees eight hours to finish one episode after weeks of practice, the new workflow required as little as two hours of labor for the same video runtime. With this new platform, VTNA was able to eliminate the backlog in eighteen months.

The new ASR transcript program proved to be a successful replacement to our existing metadata creation process, although it still required additional work to integrate into our content management and database discovery system. The Trint web application offered a variety of output formats for transcripts and captions, but none of the options integrated with our existing database of 1.4 million records. We needed a crosswalk from time-based transcript files to MySQL database fields, including title and description. For the description we used the first 500 characters of the news segment's transcript. This limited use of a transcript offered a level of summarization but kept the usage well below any copyright infringement.

Patrons: Providing Closed Captions

The VTNA used the time-coded transcripts to embed closed captions in all access copies of the collection. Video captioning is an essential element of accessible collections, allowing users to read the spoken words. The videos recorded by the television archive did not include captions originally. Using the ASR transcripts, we add the text to the video streams using a technology called FFMPEG, managed with Python. The process took some time to complete—starting with the oldest files and moving to the present, we finished the project in three months.

Patrons: Data Use Agreements

As we build out the data lake, a key concern is developing terms of service for researchers. Should the VTNA impose any terms for research beyond those stipulated in our licensing agreements? If so, what additional terms would be reasonable to impose? We have consulted with our office of sponsored research about the potential of using so-called "data use agreements" when providing access to the data lake, but it is not clear that such agreements apply when data is being used internally and not shared with external partners.

Lessons Learned and Future Work

Future Work: Abstracting and Indexing

A near-term project is the summarization of news segments. As noted above, the VTNA relies on full-time staff to write abstracts for segments of television news programming. Given the time and expertise required, we decided to reduce the number of abstracts we create each week, limiting them to a subset of our collection. As we explained, we substituted full text transcripts for abstracts to foster the searchability of our collection. Still, abstracts provide a salient overview of segments and, crucially, do so in language that is not under copyright since our abstracts represent interpretations of the news. So, while we cannot display full text transcripts to the public (or, at least, cannot display more than limited snippets of those transcripts), we can display the abstract publicly. Currently, the abstract provides a key resource for external users searching our collection in advance of transacting a loan.

Commercial and open-source machine learning models for summarizing text have existed for some years. But we must differentiate between tools that perform *extractive* and *abstractive* summarization. Extractive summarization works by identifying the most significant sentences in any given text, essentially "boiling down" that text to key passages. By contrast, abstractive summarization rewrites the text, drawing on new language to represent the meaning of the passage. For our purposes, extractive summarization does not satisfy our abstracting requirements since the technique cannot learn our implicit patterns for writing abstracts and does not free us from copyright constraints since it reuses the literal text. Abstractive summarization has the potential for satisfying both goals but was not effective until very recently.

The advent of Large Language Models (LLMs) from OpenAI, Google, and others makes conceivable the application of abstractive[3] summarization to our corpus. These tools are trained on enormous quantities of text, making it possible to apply them to corpora without fine tuning. By reverse engineering the rules for writing abstracts, we should be able to write complex prompts for abstracting the transcripts of news segments according to our traditional standards. However, some of these models, including GPT-3.5, allow for fine-tuning. In this scenario, we would need to supply manually produced abstractions and transcripts for several hundred segments, essentially "teaching" the model how to "predict" or, in plainer terms, to write summaries of future segments. Now that the OpenAI API has significantly expanded its character limits (to 8k and 32k, respectively), such fine-tuning is within reach, though it would come at significant expense ($0.03 to $0.06 per 1k tokens).

Progress is also being made in techniques for **diarization** and **chapterization**. The goal of diarization is to detect changes in speakers. Working from audio files, diarization technologies detect and label speakers. By developing the **prosopography** of journalists and newsmakers, we could add valuable context to our transcripts, allowing researchers to study the interaction between journalists and their subjects. (But see our note about the related ethical concerns above.)

Chapterization divides heterogeneous A/V into semantic segments. In the case of television news, for example, chapterization would divide the shows into news segments. Tools such as automated scene detection (ASD), a machine-learning technique used in video production, can already divide video into units. As the name implies, the technique could be used to detect transitions in news shows, a possibility that media researchers have been investigating for more than two decades (Zhu et al. 2001). However, news segments frequently have many scene changes, making ASD too fine-grained an approach. Research into chapterization of the news continues, drawing on multimodal clues from video and text (screen OCR), for example, to distinguish segments in new shows (Rozsa and Mocofan 2022). We expect that pragmatic methods for both diarization and chapterization will be available at scale within the next decade, allowing us

[3] Abstractive summarization differs from extractive summarization by focusing on preserving the semantic context rather than identifying key phrases when condensing texts.

to automate key parts of our video capture and preservation process while still maintaining our distinctive approach to identifying speakers and describing news segments.

Future Work: Analytical Tools

The VTNA not only preserves the news, it also provides tools for researchers to analyze the news. At present, our main offering is a search engine, which permits researchers to look up shows by date and network, and to search for keywords in titles, abstracts, and transcripts. Our discovery system works reasonably well, allowing users to find related news stories by using key terms. However, we believe that we can provide superior tools for analysis. There are two near-term projects that we expect to enable qualitatively superior analysis.

The first is the deployment of a **graph database** to bring the latent social graph in television news to the surface. A graph would show linkages between news segments and news shows, allowing users to traverse segments based on their nearness or distance to events rather than on keywords. When combined with **natural language processing** techniques such as topic modeling and sentiment analysis, graphs would show patterns of coverage over time. So, for instance, a user could trace how different news networks cover evolving stories, exemplifying the amount of time that networks devote to topics, the language they use to cover them, and the tone they employ during their coverage. Pending budgetary approval, the VTNA is planning to deploy a graph database called **Neo4j** to enable such network analysis.

The second is the provisioning of a **vector database** for vector-based or semantic searching of the VTNA's collection. A vector database requires that you create **word embeddings** from your search documents. In simple terms, this requires converting each word into a number and then storing those numbers in a tensor or multidimensional array. Unlike traditional 'bag of words' approaches, words in the embedding maintain their relationships to other words in the tensor, allowing the use of techniques like cosine similarity to measure the similarity or distance between vectors of words, i.e., sentences. The techniques for creating word embeddings have grown in sophistication since the invention of the Word2Vec algorithm in 2013, but storing and retrieving information from these embeddings for production use has proved challenging. Recently, a new class of vector databases has emerged to meet this need, offering databases with search engine-like capabilities that provide results based on semantic similarity rather than keyword. So, for example, a patron could search for the phrase "space shuttle" and receive hits from documents that mention only "Challenger," as well as provide related segments that satisfy some level of similarity. Tentatively, we plan to use **OpenAI's** embeddings to create word embeddings from our abstracts and transcripts and the **Pinecone** vector database to productionize our semantic search engine.

The use of word embeddings may address the problem noted above with variant spellings of personal names by allowing us to identify clusters of closely related names in a fashion akin to topic modeling and, if we desire, to create a thesaurus of names with known variants. But word embeddings also introduce a different problem, namely, the surfacing of implicit bias in the news. Given the historical nature of our collection, we may find problematic associations surfacing in the relationship between words, potentially reinforcing stereotypes of women, minorities, and people of other nationalities.

A different kind of analysis is made possible through the generation of **ngrams** from our corpus. An ngram is an ordered sequence of word tokens. As the name implies, ngrams can be of any length, but are usually two to three tokens in sequence. By using ngrams, it becomes easier to differentiate between topics in a corpus, for example "Washington, DC" and "George Washington." In the field of audiovisual media, Dr. Kalev Leetaru of the GDELT Project has released a ngrams dataset for television news, which can be used to analyze the topics discussed on different networks (Leetaru 2021). The VTNA is considering generating ngrams of its data to complement the dataset, but we need to work with legal counsel to assure that these ngrams are not considered derivative works of the copyrighted broadcasts. The GDELT Project has also pioneered the concept of so-called **visual ngrams**, which sample audiovisual broadcasts at defined intervals, taking snapshots of the news show (Leetaru 2022). These visual ngrams fall under the provisions of "fair use," and again allow comparison between the networks' coverage of topics.

Conclusion

The Vanderbilt Television News Archive has sustained its operation for more than 50 years, though its existence was at times challenged by the cost of capturing, describing, and preserving the news.[4] We also continually have sought ways to fulfill our public mission of providing access to the cultural record of television news, pushing the technological frontiers in order to expand access to our collection. By drawing on new AI/ML techniques, the VTNA has found ways to make itself both more sustainable by lowering the cost of indexing and abstracting while also expanding its audience, by providing new ways to search the collection. As we move forward in the AI/ML era, we hope to build on these early successes while keeping a careful eye on the potential ethically related pitfalls of misdescription, de-professionalization, and compromised privacy.

[4] On the sustainability challenges at the VTNA, see Marcum 2013.

Documentation

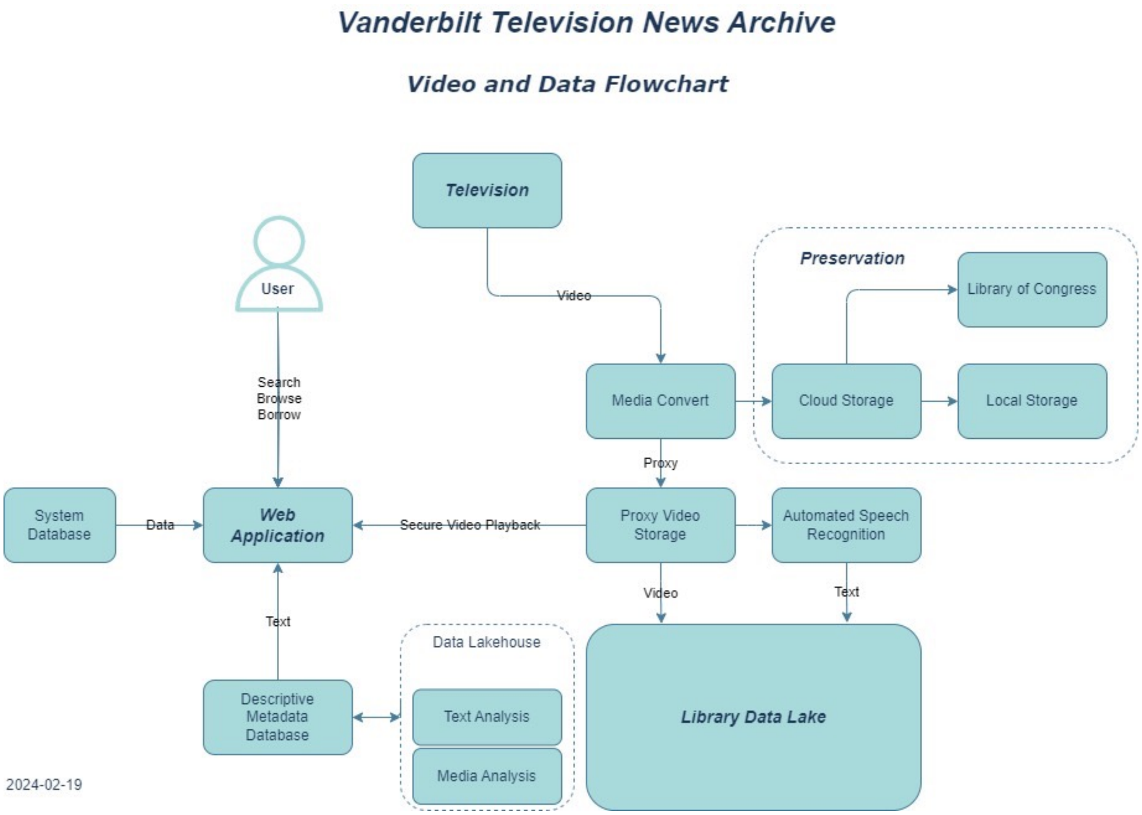

Figure 1: Complete Workflow, 2024. This workflow depicts the entire process in which video streams are acquired, processed, described, and accessed by the various user groups of the TV News Archive.

Acknowledgements

Vanderbilt University is a leader in the development of artificial intelligence and machine learning. The computer science department has several faculty who specialize in these fields. The College of Arts and Science at Vanderbilt has also commissioned an interdisciplinary group of faculty from the humanities and social sciences to address the Grand Challenge of Artificial Intelligence. One of the writers of this case study is participating in that Grand Challenge on behalf of the university library.

The research case study was developed as part of an IMLS-funded Responsible AI project, through grant number LG-252307-OLS-22.

Competing Interests

The authors declare that they have no competing interests.

References

Amazon Web Services. "Amazon Comprehend." Accessed March 21, 2023. https://aws.amazon.com/comprehend.

Anderson, Clifford B. 2022. "An Introduction to Data Lakes for Academic Librarians." *Information Services & Use* 42 (3–4): 397–407. https://doi.org/10.3233/ISU-220176.

Greene, Mark, and Dennis Meissner. 2005. "More Product, Less Process: Revamping Traditional Archival Processing." *The American Archivist* 68 (2): 208–63. https://doi.org/10.17723/aarc.68.2.c741823776k65863.

Leetaru, Kalev. 2021. "Announcing the New Web News Ngrams 3.0 Dataset." *The GDELT Project Blog*. December 15, 2021. https://blog.gdeltproject.org/announcing-the-new-web-news-ngrams-3-0-dataset.

Leetaru, Kalev. 2022. "Visual Explorer: 5.25 Million Broadcasts Totaling 12.3 Billion Seconds Of Airtime = 3 Billion Analyzable Images Spanning 50 Countries And 1 Quadrillion Pixels." *The GDELT Project Blog*. November 8, 2022. https://blog.gdeltproject.org/visual-explorer-5-25-million-broadcasts-totaling-12-3-billion-seconds-of-airtime-3-billion-analyzable-images-spanning-50-countries-and-1-quadrillion-pixels.

Marcum, Deanna. 2013. "Vanderbilt Television News Archive." *Ithaka S+R Case Study*. https://doi.org/10.18665/sr.22672.

Rozsa, Benjamin, and Muguras Mocofan. 2022. "TV News Database Indexing System with Video Structure Analysis, Representative Images Extractions and OCR for News Titles." In *2022 International Symposium on Electronics and Telecommunications (ISETC)* 1–4. https://doi.org/10.1109/ISETC56213.2022.10010319.

Zhu, Xingquan, Lide Wu, Xiangyang Xue, Xiaoye Lu, and Jianping Fan. 2001. "Automatic Scene Detection in News Program by Integrating Visual Feature and Rules." In *Advances in Multimedia Information Processing — PCM 2001*, edited by Heung-Yeung Shum, Mark Liao, and Shih-Fu Chang, 843–48. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-45453-5_109.