



NATIONAL LEADERSHIP GRANTS FOR LIBRARIES

Sample Application LG-260441-OLS-26
Project Category: Applied Research

College of William and Mary

Amount awarded by IMLS: \$750,000
Amount of cost share: \$0

The College of William and Mary in collaboration with Old Dominion University and the Internet Archive, will identify and preserve web resources that are referenced in archived US local and national television news. This project will provide context for both collections and enable the public, journalists, and researchers to trace how information moves between the web and broadcast media. The main research outcomes are developing novel artificial intelligence tools for effectively identifying web resources that are referenced in US TV news and building a public database for them, adding context banners on archived pages that are referenced in TV news, and investigating the web to TV news information flow. The project's tasks include developing novel multi-modal machine learning/deep learning web citation detection and context classification models; developing a large-scale web/TV news citation database; and linking archived web pages to citing TV news programs; characterizing information flowing from the web to US TV news.

Attached are the following components excerpted from the original application.

- Narrative
- Schedule of completion

When preparing an application for the next deadline, be sure to follow the instructions in the most recent Notice of Funding Opportunity for the grant program to which you are applying.

CiteCast: Linking US local & national television news to cited web resources

1 Introduction

William & Mary, in collaboration with Old Dominion University (ODU) and the Internet Archive (IA), respectfully request \$750,000 for a three-year National Leadership Applied Research Grant. We propose to identify and preserve web resources that are referenced in archived US local and national television news. The Internet Archive TV news archive preserves a critical record of how major events, policies, and social issues are presented to the public through television journalism. By linking television news segments to cited web resources preserved in the Wayback Machine, this project will provide richer context for both collections and enable the public, journalists, and researchers to trace how information moves between the web and broadcast media. These connections will improve discovery, transparency, and long-term preservation of the digital public record. This work is valuable for the visitors of the Internet Archive's TV news web archive to have full contextual information of the primary web sources that are referenced on TV news. Additionally, our work would enrich the largest public web archive by linking it to an important agenda-setting mass media – TV news. In short, our main research outcomes are 1) developing novel artificial intelligence (AI) tools for effectively identifying and preserving web resources that are referenced in US local and national TV news and building a public database for them, 2) adding context banners on archived pages that are referenced in TV news thereby establishing a bidirectional link between TV news and cited web resources, and 3) investigating the extent and nature of information flowing from the web to TV news by addressing the research questions outlined below. Our project outcomes align with the IMLS priority of creatively and effectively using AI to improve visitor experiences at our nation's libraries, and the NLG-L Program Objective 1.3, to provide broad access to and preservation of information and collections through libraries and archives. Moreover, they align with IMLS Goal 3 (advance collections stewardship and access), and Objectives 3.1 (support collections care and management) and 3.2 (promote access to museum and library collection).

If funded, this grant would enable us conduct research to understand the link between two consequential modes of mass communication, the web and TV news. We propose to address the following important research questions (RQs) that are relevant to the public and libraries or archives:

- RQ1: How often do TV news reports cite web content and how well are the cited resources archived?
- RQ2: What is the distribution of the types of web resources (e.g., other news stories on the web, government/company websites, social media posts) that are referenced in television news?
- RQ3: In what contexts (e.g., evidence, audience engagement) does TV news cite web resources?
- RQ4: To what extent does linking TV news content to web content increase public engagement with the TV news archives?

2 Project Justification

The hyperlink, or simply *link*, is the fundamental building block of the web, which is arguably one of the most important technological inventions of the 21st century. Links also exist in TV news reports even though they are not always as explicit as the blue hyperlinks that we click. Fig. 1a illustrates a March 2013 KQED (PBS) TV news clip [1] referring to a now defunct website ([noodle.org](#)) for finding affordable schools. Also, TV news presenters regularly cite web content, to provide evidence for major political or economic news, to amplify popular memes (e.g., *the dress*), or to share eye-witness accounts, etc. For example, on April 9, 2025, President Trump posted on Truth Social announcing to his approximately 10 million followers that his administration had placed a 125% tariff on Chinese goods [2]. Similarly, on June 21, 2025, he also announced in a social media post that the US military had bombed Iranian nuclear sites [3]. Both social media posts from the president were cited by CNN (Fig. 1c) and by NBC News (Fig. 1d). When an author cites a work, the citation can give credit to another author, or provide the evidence needed to support an argument or reproduce

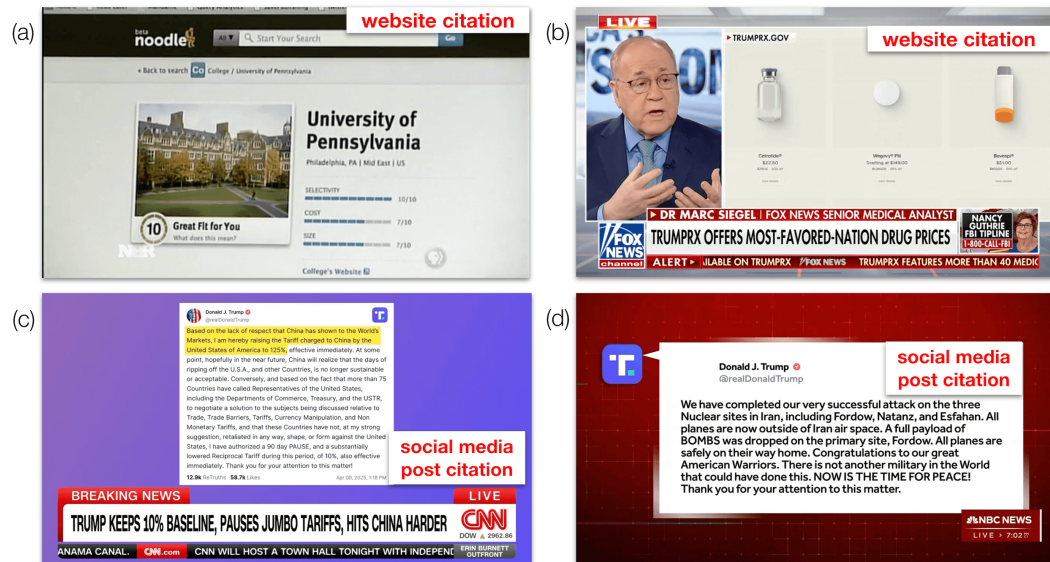


Figure 1: Example TV news citing web resources. (a) A March 2013 KQED (PBS) TV news clip [1] referring to a now defunct website (noodle.org). (b) A February 2026 FoxNews TV news clip [4] citing (trumprx.gov). (c) A CNN TV news clip [5] citing President Trump's April 2025 post about his administration's placement of a 125% Tariff on Chinese goods [2]. (d) An NBC TV news clip [6] citing President Trump's June 2025 post that the US military had bombed Iranian nuclear sites [3].

a method. The collective body of scholarly citations and their networks are valuable; showing the intellectual lineage of ideas, capturing the impact of research(ers), and are fundamental objects of study in bibliometrics. Citations also exist in television news as we have shown, however, unlike the scholarly domain, which has bibliometrics to study the citation networks of authors, television media does not have an equivalent field for investigating its citation networks, which severely limits research. We argue that this is partly because existing research on the fundamental task of mining citations in mass media is sparse or absent, mainly focusing on citations of research papers [7] in news or vice versa [8]. We propose CiteCast to address this problem and thus open future research on citation analysis in TV news. By mining the link structure of the web which connects human knowledge, culture, and social interactions, all expressed on webpages, researchers have revealed valuable insight, such as how collective intelligence can approximate importance [9, 10], that the distance between individuals online is small [11] on average, and that online communities form even though there is no central authority governing how people link to webpages [12, 13].

What can we learn by making links in US local, broadcast, and cable TV news explicit and accessible to the public and researchers? What novel insight might we gain by a large-scale mining of the link structure of web resources referenced in TV news? Currently no research, datasets, or tools exist to explore these important questions to shed light on the link between the web to TV news. We propose the CiteCast project to address this important research gap by generating citations for web resources that are referenced in television news.

2.1 Why television news?

TV news existed almost half a century before the web and has been the subject of important research, such as McCombs and Shaw's seminal work on agenda-setting theory [14] in 1972, which revealed how traditional mass media like television and newspapers determine the issues that the public considers important through their choices about what stories to cover. In today's fragmented media landscape, social media algorithms and influencers, as well as TV editors contribute in setting the public agenda [15, 16], which in turn influence the public on economic, health, or other crucial matters [17, 18]. Even though people increasingly prefer to get

their news from digital platforms like social media, millions of Americans still watch broadcast and cable TV news daily [19, 20] and a majority of Americans access news via television, and among older adults, television is the preferred medium [21]. Despite this, TV news has received significantly less research attention compared to online news, especially since processing video data from TV news is more challenging than textual data. Consequently, researchers are unable study how information flows from the web to US local, broadcast, and cable TV news, since this requires processing terabyte-scale audio, visual, and textual data with expensive computer hardware and sophisticated custom-built machine-learning algorithms.

The dominance of TV news as a form of mass media for majority of Americans, its major role in agenda-setting even in a the fragmented media, the lack of existing research and tools for mining links in TV news, present a significant opportunity for us to fill in a research gap and enable future research on the interplay between TV news and the web, which is not currently possible.

2.2 The Internet Archive’s TV News Archive

We will utilize the TV News Archive (<https://archive.org/details/tv>), developed by our industry partner, the Internet Archive (IA). The Internet Archive was founded in 1996 and is the largest public web archive. It periodically saves web resources before they are lost forever and has preserved over 1 trillion webpages, 46 million books and texts, and provides free online access to these resources. Just like web resources such as news articles, television news is an ephemeral medium, so in 2009, the Internet Archive started the TV News Archive project and began preserving international, national, and local (US) TV news. The TV News Archive is a valuable resource for researchers and the public and it has been cited in over 2,000 articles and research publications [22].

Today the TV News Archive holds over 3 million TV news video clips from both a *static* and a growing *on-going* collection of 49 TV news stations (US: 19, International: 30). The static collection includes TV news captured from 2015 – 2016 from about 200 US local TV stations as part of the Political TV Ad project [23], which was funded mainly by the Knight Foundation, to provide a searchable online archive of US 2016 election cycle political TV ads. The ongoing collection includes TV news captured from 2009 – present, from broadcast (CSPAN1 and CSPAN2) and cable news stations (CNBC, CNN, CNNW, FOX Business, Fox News, and MSNBC). Moreover, it includes archived local news which predominantly serve the San Francisco Bay Area: KDTV (Univision), KGO (ABC), KNTV (NBC), KPIX (CBS), KQED (PBS), KRON (MyNetworkTV), KSTS (Telemundo), KTVU (FOX), and SFGovTV. For this project, we will use data from US broadcast, cable, and local news from the ongoing collection (2021 – 2026).

In addition to video data, there are three data modalities from the TV News Archive that are of interest for this project. First, *closed captions* text transcription of spoken news are provided and would be used to identify instances where web platforms or topics are mentioned verbally in the news. Second, *chyrons* [24] (aka lower thirds) are text in the lower third of a TV news display. Third, full-resolution *screen shots* captured at four-second intervals of TV news shows are available through the TV Visual Explorer [25, 26], developed by the GDELT project [27] in collaboration with the Internet Archive.

3 Project Work Plan

Here we explain our work plan for four deliverables for linking US local and national television news to cited web resources. Below each task, we mention the the PI lead (all Co-PIs will be involved in all tasks), and estimated time frame. PI Nwala will be coordinating all tasks. In Section 5.1.2, item 2, we detail how will engage various stakeholders’ input in the development and sustainability of CiteCast.

3.1 Task 1: Develop novel multi-modal ML/DL web citation detection and context classification models

- Lead: PI Nwala (Time frame: September 2026 – September 2027)

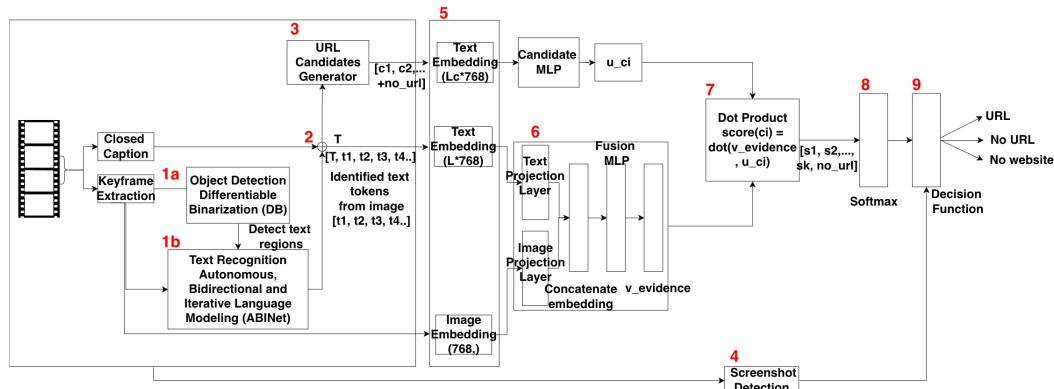


Figure 2: Novel multimodal web citation detection model based on intermediate fusion architecture.

Here we explain our plan to develop our training dataset and three separate novel models. The web citation detection model would detect websites cited in TV news, the citation segmenter will identify the start and end timestamps of citations, and the citation context classifiers will determine the various contexts of citations.

3.1.1 Gold-standard dataset of TV news citations of web resources

We will curate a dataset of web resource citations in TV news from the Internet Archives’s TV News Archive. First, we will query the TV News Archive with a list of web query terms: *web*, *website*, *webpage*, *homepage*, *landing page*, *blog*, *web resource*, *web document*, *online resource*, *online article*, *site*, *online*, *internet*, *world wide web*, *WWW*, *URL*, *social media site* *social media*, *social network*, *youtube*, *facebook*, *instagram*, *tiktok*, *twitter (X)*, *bluesky*, *threads*, *truth social*, *post*, *viral*, *meme*, *content creator*, *influencer*, etc. This would retrieve news videos containing candidate web references, with captions and/or chyrons matching the keywords along with potentially false positive cases. Additionally, we will query search engines, large-language models, and Wikipedia repositories for local and national news stories that were featured on TV news and subsequently search the TV news archive to find actual cases of web references. Second, we will conduct additional checks by manually inspecting all candidate video clips and filtering out false positives to ensure that our gold-standard dataset consists exclusively of TV news video clips with actual web references. The PI would supervise two PhD students in the labelling task.

We define a *textual web resource citation* as a verbal mention of a web resource by news hosts. These mentions, from closed captions, chyron, or OCR text of TV clips, use web query terms (e.g., *website*). We define a *visual web citation* as a video that includes a graphical representation of a textual citation, for example, a screenshot of a website (Fig. 1b) or social media post (Fig. 1c & d). Once we detect both textual and visual web references, we will subsequently segment them by marking when each citation starts and ends.

3.1.2 Web citation detector

The web citation detector will detect web resources cited in TV news, and will return their respective URLs which are either spoken (extracted from closed-caption text) or shown on the screen. If a website is detected when no URL is provided (e.g., social media post or news article), the system would return *Unknown*. For such web resources that are cited without URLs, we will use the *URL finder* described below to attempt to find their URLs. Our design emphasizes open-set generalization, meaning that the model can identify web resources/URLs not seen during training. Recall that the TV news video data includes two main channels of information - visual elements and textual ones from on-screen and closed-caption text (transcribed speech). Each of these channels can contain different but complementary clues to detect web citations. Multimodal learning enables us to combine these different inputs for more accurate and robust decisions making than if only one input channel was relied upon. This is especially important because web citations are not always all present visually or spoken (closed-caption text).

Preprocessing (Fig. 2.1-2.3): First, representative frames are sampled to capture on-screen cues without processing every single frame (Fig. 2 no. 1–3). For each sampled frame, DBNet [28] is applied to detect text regions (Fig. 2 no. 1a). Its differentiable binarization mechanism enables robust localization of text even when fonts are irregular or rotated. The detected regions are then passed through ABINet [29], which recognizes the text using a language model that benefits from bidirectional context and iterative refinement (Fig. 2 no. 1b). The recognized tokens are concatenated (Fig. 2 no. 2) with the closed-caption transcript, producing a single unified string that contains both spoken dialogue and on-screen text. This concatenation ensures no potential mention is lost, whether expressed in speech or only visually.

The combined string is tokenized, embedded, and standardized through padding or clipping so information from all input channels have a consistent shape for batch processing. Additionally before embedding the text, a separate stream of the text is sent to a *URL candidates generator* before embedding at Fig. 2 no. 5. This rule-based module is responsible for extracting possible complete URLs or URL components for the cited web resources from both close-caption (Fig. 2 no. 2) and OCR text (Fig. 2 no. 1a/1b). This component prioritizes high recall to ensure that no valid URL is missed. The extraction rules include regular expression patterns to extract complete URLs (e.g., <https://example.com>), or URL components (e.g., subdomains, domains, top-level domains) from closed-caption or OCR-corrected text from video frames. OCR corrections are applied to handle common errors such as character confusions (1 vs l, 0 vs o), extra spaces, and trailing punctuations. The final candidate list for each segment includes all extracted URLs and a special *Unknown* label for the URLs of websites that was not provided. The model will select from this candidate set during prediction.

On the visual side, the sampled frame is passed through an embedding network (Fig. 2 no. 5) that converts the raw pixels into a sequence of feature vectors. This sequence serves as the representation of the frame’s visual content, capturing distinctive patterns such as website artifacts such as icons or HTML elements like input boxes. After preprocessing, each frame is represented by a context-ready text sequence and a structured visual feature sequence (input to the multimodal model).

Architecture: The multimodal model consists of three encoders (Fig. 2 no. 5) and two lightweight neural layers - Candidate and Fusion Multi-Layer Perceptron (MLP). A pretrained image encoder transforms the frame into a visual embedding. A shared text encoder maps caption text and OCR text into a shared semantic embedding space. The same text encoder is also used to encode candidate URL strings.

Intermediate fusion (Fig. 2 no. 6) is performed by concatenating the image, caption, and OCR embeddings and passing them through a small fusion MLP. This produces a unified evidence representation capturing the combined multimodal context. Each candidate URL is encoded using the text encoder and optionally passed through a candidate-specific MLP to produce a candidate representation. A dot-product similarity (Fig. 2 no. 7) is computed between the fused evidence representation and each candidate representation to produce a score for every candidate, including *Unknown*.

Training: Each training example consists of an image frame, caption text, OCR text, candidate URLs, and a single ground-truth label indicating the website URL, *Unknown*, or no website. The model is trained using cross-entropy loss over the candidate sets. The model learns how to align multimodal evidence with the correct candidate without memorizing specific web resources. Because the system ranks dynamically generated URL candidates of websites rather than predicting from a fixed list of URLs, it naturally generalizes to unseen websites during inference.

Inference: the system first extracts candidate URLs from closed-caption and OCR text, and then detects screenshots. The image, caption, and OCR inputs are encoded and fused, and each candidate URL is scored. If the selected candidate web resource URL is not *Unknown*, its corresponding URL is returned. If the selected candidate web resource URL is *Unknown* and a screenshot was detected, the system returns *Unknown*. If no screenshot evidence exists and no URL is extracted, the system returns that no website was detected.

URL finder: Our multi-modal citation web citation detection model is able to identify the URL of detected web resources when they are either spoken verbally (Fig. 1a) or displayed on the screen (Fig. 1b). In some cases however, web resources are cited with no URLs provided. This is more common for social media posts. For example, in Fig. 1c/d the Trump tweets were cited with no URLs provided. For such cases, we will develop a heuristic model to issue a search query to search engines using text extracted from the screenshot of the website, to attempt to find the URL.

3.1.3 Web citation segmenter

Once we detect web citations in a TV news video clip, we will segment it by marking when it starts and ends. This can be done by re-running our detector starting from the detection time point (t_c) to mark a time point $t_c - t_k$ before, and a time point $t_c + t_i$ after the web reference is not detectable.

3.1.4 Web citation context classifiers

We will develop multiple web citation context classifiers to determine various contexts of web citations.

What-context: This classifier determines the topic (e.g., economics, sports, entertainment) of the news segment when the web resource was cited. We will employ state-of-the-art language models such as BERTopic [30] for topic modeling. Additionally, we will mark the specific website from our citation detection model and the TV news station that did the citation.

How-context: We will use our web citation detector to derive the *how-contexts* of citations, since it can determine whether web content was cited verbally and/or visually (screenshot or closed-caption textual reference).

Why-context: This classifier will determine why web content is cited. Through our preliminary investigation which we will expand over the course of this project, we have identified the following initial four *why-contexts* that explain why web resources are cited in TV news.

Web resource-is-topic involves cases where some web resources (e.g., website, social media post) is the subject of the TV news coverage. Fig. 1 illustrates four different examples: the March 2013 KQED (PBS) TV news clip about finding affordable schools using the now defunct website noodle.org, the February 2026 FoxNews clip discussing how trumprx.gov was addressing affordability of drug prices, and the CNN and NBC clips, respectively discussing that President Trump had placed a 125% Tariff on Chinese goods, ordered the US military to bomb Iranian nuclear sites.

Reference involves news stories a web resource is referenced to support a claim, illustrate a point, or add depth to the story. In *request-for-engagement*, a TV host prompts viewers to engage with them online (post, vote, follow, etc.), either explicitly through verbal cues or implicitly through on-screen graphics. Finally, *passing reference* happens when a web resource is mentioned briefly as part of casual conversation or humorous commentary, without being central to the story.

We will treat detecting the context of the citation of web resources as a supervised learning problem where we train a classifier (using web resource citations gold-standard dataset) to learn the relationship between features representation of web citations and citation contexts. First, we will generate informative features that accurately signal the specific context of web resource citations. Specifically, we will research using a combination of latent feature (embeddings from the web citation detector in Fig. 2) and hand-crafted features (*duration*, textual and/or visual elements present, location of citation, etc.) from textual and visual web resource citations. Third, we will investigate the effectiveness of deep-learning and traditional ML models for the task of detecting the specific context of a web resource citation.

3.2 Task 2: Develop large-scale web/TV news citation database

- Lead: co-PI Nelson (Time frame: January 2027 – September 2028)

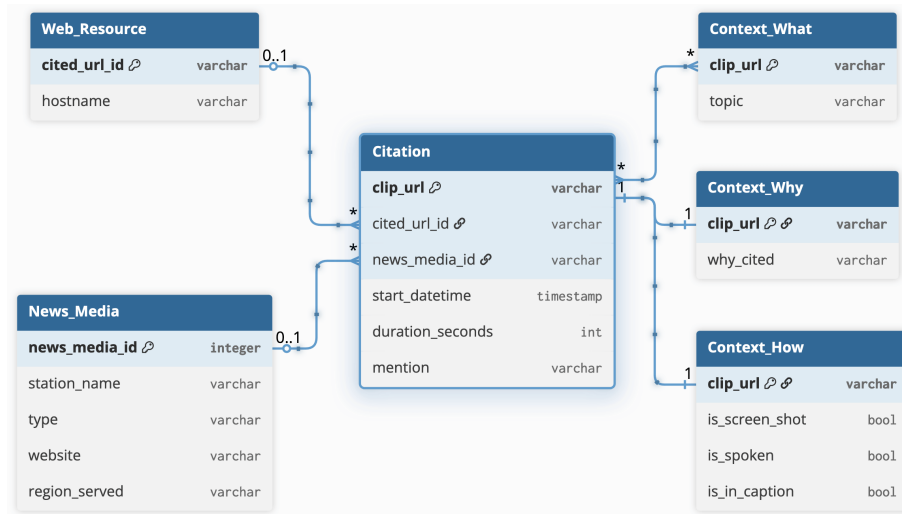


Figure 3: Schema for the CiteCast web citations database

Using the citation detection and context classification models (Task 1), we will conduct a large-scale data analysis by processing TV news data (from 2021 – 2026) to detect instances where local and national TV news have cited web content and identify their respective citation contexts. Next, for each of the cited web resources on TV, we will annotate their archived copies on the WayBack Machine at the Internet Archive with context banners, thereby establishing a bi-directional link between TV news and cited web resources.

3.2.1 CiteCast database schema

As outlined in Fig. 3, the CiteCast database would consist of a minimum of six tables. The states of the fields of all six tables collectively represent web citations. Henceforth, for brevity we will sometimes use the syntax *Table name.Field name*, e.g., *Citation.clip_url* to represent a given field (e.g., *clip_url*) from a given table (e.g., *Citation*), and use *Table name.** to reference all fields in a table.

The *Citation* table would contain basic information about a web resources that is cited on TV news. This includes a link to the video clip (*clip_url*) citing the web resource, the URL of the cited web resource (*cited_url_id*), and TV news station (*news_media_id*) which cited it. *Citation.start_datetime* would store the datetime, *Citation.duration_seconds* the length (in seconds), and *Citation.mention* (if applicable) would store textual information of the citation. The *Context_What* table would store the topic associated with the citation, while *Context_Why* would the store one of the various why-contexts (*Web resource-is-topic*, *Reference*, *Request-for-engagement*, *passing reference*) of a citation. The *Context_How* would store how a citation manifests: indicating whether it is a screenshot (*is_screen_shot*), is the web resource was spoken by the news host (*is_spoken*), and whether the citation was included in the news caption (*is_in_caption*).

Moreover, the *Web_Resource* table would contain the *hostname* and URL (*cited_url_id*) of the cited website, while the *News_Media* table would store information of the TV news station that did the citing. This includes the station name (*station_name*), the *type* (local, broadcast, or cable TV news), the *website*, and the geographical area where the program is visible (*region_served*). We will gather the information about each TV news media by consulting open-source and proprietary resources.

3.3 Task 3: Linking archived web pages to citing TV news programs

- Lead: co-PI Alam (Time frame: January 2027 – September 2028)

We will annotate the Wayback Machine at the Internet Archive using the URLs of cited web resources. Specifically, given the URL of a cited web resource (e.g., trumprx.gov), we will first check it is archived and preserve it if it is not. Subsequently, we create a hyperlink for the cited web resource, pointing to the TV news reports where it was cited. Fig. 4 illustrates a mockup showing a hyperlink with anchor text, *102 captures on TV news*,

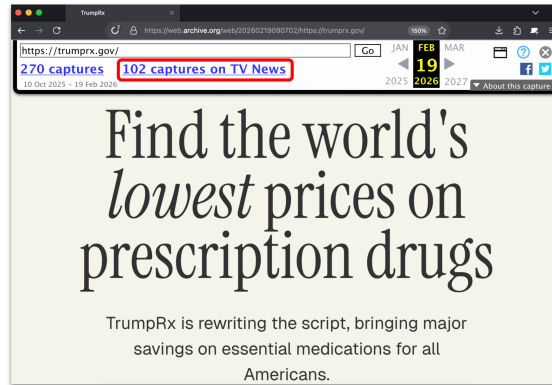


Figure 4: Mockup with annotation of cited web resource (trumpxr.gov) with link to the 102 TV references

pointing to the TV news reports that cited trumpxr.gov. Consequently, the daily Internet Archive users will have additional context that the current URL they are replaying has been cited by TV news outlet.

3.4 Task 4: Characterizing information flowing from the web to US TV news

- Lead: PI Nwala (Time frame: January 2027 – August 2029). RQ1/4 Lead: co-PI Alam, RQ2 Lead: co-PI Weigle, and RQ3 Lead: PI Nwala.

Using the CiteCast database of web citations and their respective contexts discussed in Task 2, we will conduct a comprehensive longitudinal investigation to map and characterize the information flow from web to US local and national TV news, by addressing our four main research questions.

RQ1: How often does TV news reports cite web content and how well are the cited resources archived?

We will utilize our CiteCast database to count all unique instances of social media citations and their respective timestamps and additional metadata to generate a distribution of web citations over time and summary statistics such a number of citations per day, for local, broadcast, or cable TV station.

RQ2: What is the distribution of the types of web resources (e.g., other news stories on the web, government/company websites, social media posts) that are referenced in television news? We will deploy either custom-built web resource classifiers using transformers models or LLMs which have been shown to be effective for various NLP tasks. The final models we use would depend best-performing model trained/evaluated on the task of classifying web resources into pre-defined categories using a curated labeled dataset.

RQ3: In what contexts (e.g., evidence, audience engagement) does TV news cite web resources? We will deploy our web citation context classifiers (Section 3.1.4) for this task to determine the *What*, *How*, and *Why*-contexts of web citations.

RQ4: To what extent does linking TV news content to web content increase public engagement with the TV news archives?

In Section 3.3 we discussed how we will create a hyperlink pointing cited web resources to the respective TV news reports that cite them, as illustrated by Fig. 4. We will implement a system to detect when Internet Archive visitors click on these hyperlinks to measure the degree to which linking web resources to citing TV news content increase public engagement.

4 Project Evaluation

We will evaluate the performance of our web content detector/context classification models on the task of detecting web citations using our labeled dataset. For each of our classifiers, we will use stratified k-fold cross validation to obtain performance metrics such as precision, recall, F_1 , and Matthews Correlation Coefficient (MCC). Additionally, since we will be processing millions of video clips from the TV news archive, algorithmic

efficiency is important, and will be included in our evaluation. In addition to the formal evaluation outlined, we will assess our success based on the usage of the web services, number of students successfully mentored, the number of publications, citations, and user feedback.

5 Project Results

This project will have four main deliverables (D1 – D4). To ensure that each deliverable is readily adaptable and usable by other institutions and communities, we will continue in the tradition of the Internet Archive of producing free valuable web resources (e.g., TV News Archive, Wayback Machine) by implementing CiteCast as a public data infrastructure. We will ensure all code and data deliverables open-source or released under the most permissive license possible (see Digital Products Plan and Dissemination Plan below).

D1 - Multi-modal machine/deep-learning (ML/DL) web citation detection/segmenter and context classification models: The citation networks of authors have been extensively studied by bibliometrics researchers to track the provenance of knowledge and measure impact and influence. Citations exist in television news as we have shown, but not studied, partly because existing research on the fundamental task of mining citations in mass media is sparse or absent. We will address this need by researching the representations of web resources cited in US TV news and develop multi-modal ML/DL models to detect them. Next, we will research the contexts of web resource citations in TV news, and develop classifiers to detect them (Section 3.1.4).

D2 - Large-scale web/TV news citation database: Using models from D1, we will conduct a large-scale analysis of a half-decade (2021 – 2026) of TV news data to detect instances where local and national TV news have cited web content and identify their respective citation contexts. This would enable us study the nature of information flowing (See D4) from the web to TV news, which is currently unknown. We will release this database for public access to enable the public and researchers study the link between the web and TV news.

D3 - Linking the archived web resources on Wayback Machine to TV news reports that cite them: The Internet Archive is the largest public web archive in the world. By linking the archived web resources on Wayback Machine to TV news reports that cite them, this work would enable these visitors to have additional valuable contextual information of the primary web sources that are referenced on TV news and establish a bidirectional link between TV news and cited web resources.

D4 - Mapping and characterizing information flowing from the web to US local and national TV news: Using the database from D2, we will conduct a longitudinal study to address our four primary research questions to understand the extent and nature of information flowing from the web to TV news: RQ1) What is the distribution of web resources (e.g., government/company websites, social media posts) referenced in television news? RQ2) How often does TV news reports cite web content and how well are the cited resources archived? RQ3) In what contexts (e.g., evidence, audience engagement) does TV news cite web resources? RQ4) To what extent does linking TV news content to web content increase public engagement with the TV news archives?

5.1 Dissemination and Availability of Findings

Here we summarize our plan to disseminate our research output to the public and research communities, and our plan to ensure that our deliverables are sustained beyond the period of performance.

5.1.1 Dissemination Plan

The tangible products of this work include software (D1), data (D2) and public web interfaces to access them (D2/D3), and the research findings of this project (D1–D4). All code and non-database data will be shared publicly with a permissive license (e.g., Creative Commons license, GPL) on GitHub using standard formats like CSV, JSON, etc., to encourage reuse and additional research. The Internet Archive will host the CiteCast database indefinitely. We will disseminate relevant data and research findings at venues such as the ACM/IEEE Joint Conference on Digital Libraries (JCDL), ACM Web Science, the ACM Conference on Information and Knowledge Management (CIKM), and the International Journal on Digital Libraries (IJDL).

In addition to these formal venues we will also describe our research findings in blog posts and preprints. We will use our standing in the digital library community to demonstrate and present the work at workshops such as International Internet Preservation Consortium (IIPC) and the annual Archive-It Partners Meeting. As researchers engaged in digital preservation and ties to the Internet Archive, we are well-positioned to ensure the long-term preservation of our research output.

5.1.2 Sustainability Plan

We care deeply about digital preservation and sustainability especially as web archiving researchers. Our industry partner, the Internet Archive, is the largest public web archive with a core mission to provide universal access to knowledge in perpetuity. Combined, we have decades of experience sustaining our research output through consistent grant and institutional support. For example, our StoryGraph¹ service — which generates a longitudinal national news dataset and quantifies news attention — has been running for almost a decade. The Internet Archive has committed to host all data generated to build the SocioCast database (2021 – 2026), indefinitely. Next, we describe our plan to sustain CiteCast beyond the expiration of this project.

1. Infrastructure and data management: as outlined in our Digital Products Plan, the Internet Archive would provide the storage for the CiteCast database and all derivative data produced during the course of this project. William & Mary IT would provide the infrastructure for all CiteCast online services, as they have provided for other services of the PI, such as *Local Memory Project* and *Agwagram*. These services are currently maintained by undergraduate research assistants at the supervision of PI Nwala. Similarly, CiteCast would be sustained by undergraduate research assistants and the community of CiteCast users as detailed next.

2. Personnel/research community engagement: the sustainability of CiteCast partly hinges on support from an active community of users including, students, researchers, and the public. Therefore, we will incorporate CiteCast in class research projects surrounding various news spreading phenomena from web to TV news. Highly-ranked students will become candidate research assistants to support CiteCast. Additionally, we will publicize CiteCast via the Internet Archive's platforms, formal venues (conferences/workshops) and informal venues (social media/blog posts).

3. Institutional support, funding, and collaborations: Our respective institutions (William & Mary, Old Dominion University, and the Internet Archive) are committed and have a history of supporting research output that have outlived their active funding timelines. In addition to the institutional support provided by our respective institutions, we will actively solicit future funding to sustain CiteCast specifically for additional data generation, software/hardware maintenance after the performance period. Additionally, we will collaborate with other researchers on collaborative proposals. Specifically, members of the CiteCast team could be included into grants as sub-awardees, in support of additional features or collaborating on research.

¹All tools available at: <https://newsresearch.lab.wm.edu/>

Schedule of Completion

Task (Lead Org.)	Year 1	Year 2	Year 3
Task 1: Develop labeled dataset of web citations (ODU)	X		
Task 1: Develop multi-modal ML/DL web citation detection & context classification models (W&M)	X	X	
Task 2: Develop large-scale web/TV news citation database (ODU)		X	X
Task 3: Linking the archived web resources on Wayback Machine to TV news (IA)		X	X
Task 4: Map and characterize information flow from web to US TV news (W&M)		X	X
Evaluation	X	X	X
Dissemination	X	X	X

Table 1: Research task Gantt chart illustrating project timeline and organizations leading each task