IJDC | *Peer-Reviewed Paper*

Creating an Online Television Archive, 1987–2013

Robert X. Browning
Purdue University

Abstract

The growth of television, and in particular television news, has created a challenge in preserving and providing access to the resulting material. At the same time, technology has opened many opportunities to capture this information and make it more widely available. In some ways, it is a race of technology against the speed of content creation. In this paper, we describe a very successful archival project that records, indexes, archives and makes available the totality of the programming of the U.S. based C-SPAN television network, a nonprofit network that telecasts the entirety of the U.S. congressional proceedings, hearings, presidential speeches and other public policy events. As such, it is an archive of unedited primary source events. The use of evolving technology over 25 years has made this archive possible and it exists free on the Internet for world-wide access.

Received 12 September 2013 | *Accepted* 8 May 2014

Correspondence should be addressed to Robert X. Browning, Purdue University, C-SPAN Archives, West Lafayette, IN, USA, 47906. Email: rxb@purdue.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v9i1.288

Introduction

Television public affairs programming is simultaneously ubiquitous and ephemeral. It can disappear as quickly as it appears. Yet, it is so important as a record of our modern society that it creates a challenge for those who want to foster the preservation and access to television programming. This paper is a case study of a successful 25 year project to do just that – build an accessible online digital television archive. Because the project relied so heavily on innovation in technology, we use this case study to illuminate the methods and lessons of the project.

Background

In 1987, Purdue University created an off-satellite archive to record, index and archive all programming telecast by the C-SPAN television networks. This was a large undertaking. Vanderbilt University had pioneered the idea of a university creating an off-air archive when it started recording the three network evening news programs in 1968. That amounts to only 1,100 hours per year and Purdue would record two networks, 24 hours per day, seven days per week, or 17,520 hours per year.

C-SPAN is a Washington, DC nonprofit, cable television network that telecasts coverage of the U.S. House of Representatives, the U.S. Senate, congressional hearings, news conferences, presidential speeches and news conferences, political and campaign events, book and history programming. C-SPAN does not edit, but telecasts events in their entirety. As such, Purdue sought to create an archive of primary source audio-video documents that contained the debates and discussions of the American democracy.

The origin of the idea came from a meeting of Brian Lamb, the founder of the C-SPAN networks, and myself, a professor of political science at Purdue University, and other faculty in communication and history. Lamb grew up in Lafayette, Indiana and attended Purdue University. He was interested in expanding the use of C-SPAN programming in teaching and research. As faculty, we shared that goal, but pointed out the importance of an archive to preserve and organize the content. Thus, at that meeting in 1986, the idea was broached and I was tapped to research and undertake creating an archive.

As a small television network, C-SPAN maintained a limited production library of tapes, but recycled many tapes to reuse them. The Purdue goal was to preserve and to provide access to this material for educators and researchers who wanted to review the video record of important political events. The archive would be independent of C-SPAN and Purdue would decide how the archive should be organized and indexed. C-SPAN granted to Purdue a license to distribute the content to educators for a fee. Over time, as Purdue demonstrated the efficiency of its systems for retrieving and duplicating content quickly, C-SPAN also contracted with Purdue to provide all of its duplication services.

This arrangement was unique in that it was the first known archive of a single network what would contain all the material telecast by the network. Secondly, it was the first university to be licensed by a network to distribute copies of its contents for a fee. Purdue created the Public Affairs Video Archives, operating under the auspices of the School of Liberal Arts, and began recording C-SPAN content in September and

October of 1987. In this paper, we will discuss the technical issues that the archive confronted and how it solved them in its 25 year history. The archive today holds over 200,000 hours of digital content and won a Peabody Award in 2010 for its creation of the C-SPAN Video Library[1]. In 1998, management of the archive was transferred from Purdue University to C-SPAN and the archive operates as a unit of C-SPAN independent of the university, but is still located in the Purdue Research Park.

Developments

A project of this size could only succeed through the effective use of technology. Over 25 years that technology has evolved and it was very important to keep abreast and keep investing in new technology to harness the innovations. This paper outlines the challenges and how technology was used to manage the enormous amount of information that the archive processed.

Dealing with the Volume of Data

The problem was that there were 24 hours of content on two networks to be recorded each day. There was no network database and a limited schedule to guide us on what we were receiving. It was also necessary to devise an indexing scheme to capture the right information on each program so that the program could be retrieved later. In addition, we had to link the content indexes to the physical media on which the content was being stored.

Our initial solution was to label each two-hour VHS tape with the year, month, day, hour, minute, network that was being recorded. This created a unique number for each cassette that could be linked back to the time of each master program. Since C-SPAN telecasts about 8,000 of live content in the 17,520 hours total telecast time per year on two networks, there were many duplicate airings to keep track of and to determine whether they were first-run content.

By 1987, the personal computer had been around for about five years. Software was being developed to link these computers in networks and to organize information in databases. We took advantage of the technical base at Purdue University to tap talent and students to assist with the project. One of the first employees was a computer science freshman who helped to write the first database programs to organize the information. Together, we developed a schema for entering data that is still being used, with modifications, 25 years later.

Each morning, we would retrieve 24 tapes that were recorded over the previous day. We would play each tape at high speeds, stopping when we saw a program beginning. The software would allow us to enter the begin date, time and network, and then the end date, time and network, and would calculate the length of the program. Then it would display the last programs entered by category so we could scan to see if this program had already been entered. If it had, we would select the program and add an airing to the selected program. If the airing was longer or shorter that the previously selected program, the software would issue a warning. The logic was that there could be no gaps in a 24 hour schedule and an airing longer than the master would indicate that the program being entered was not a duplicate or it was a longer version of the original content.

1 C-SPAN Video Library: http://c-spanvideo.org

 doi:10.2218/ijdc.v9i1.288

Additional information entered were program name, category, format, speakers, organization, summary abstract and keywords. Not all of this was entered initially. Some of the fields were added over the first year of operation. Initially, only one person could enter information at a time, so the first data were collected on paper.

One innovation of the C-SPAN Archives was to relate content to media. The video was recorded on two-hour VHS tapes initially, then SVHS tapes, then two-hour files, then one-hour files, and finally five-minute files. The five-minute files allowed the video to be played out within ten minutes of when a program began, allowing almost immediate access to the digital media collection. Whatever the medium, the video was recorded continuously and then the database was used to link the content to the media. Since the media were labeled by time, it was possible to retrieve a program by going directly to the tape or file if one knew the time of the original broadcast. Thus, for any live program, one could retrieve the tape without using the database.

This linking was so important in tying the content to the media. It was used in duplication, later in playing files, and in archiving. The video player built by the Archives staff could handle offsets – the time on the tape before the selected content began – and could play programs across files without interruption.

The success of the archive depended not just on the recording and indexing, but also on the distribution and access. Within days of establishing the archive, educators were calling to seek material. A system was quickly developed to make duplicate copies of the material. Since we were dealing with videotape, this was a real-time process that required retrieving the master tape, cuing it up to the starting point, and making a duplicate copy. Initially, this was all a manual process, but we began automating it through the printing of retrieval lists and duplicate tape labels. Because of the labor intensity and the risk to the master tapes, we began to duplicate multiple copies when we anticipated that there would be repeat requests for the content. We basically used a three-times rule. If it was current content, we would multiply the requested number by three to arrive at the number to be duplicated. These excess copies were held to handle future orders and then eventually recycled if they were not requested. If it was an unusual archival request, we made only the number requested.

Eventually, we were able to add other time-saving features. We used automated loaders and custom software to create duplication jobs that could run unattended overnight. When we converted to digital recording, we would play the master content from the servers to machine-controlled duplicators. Later, DVD burners were acquired. These DVD burners held stacks of blank DVDs and would be commanded to duplicate the requested master content. A duplication image was created and electronically stored. It could be called again when the content was requested by another user. Software was custom-coded to write the information from the database in menus directly to the header of the DVD. One could go online to purchase a DVD and no human would touch anything until the DVD was removed from the DVD burner. The credit card was cleared, the order created, the content pulled from the server, tape library, or image file, the data written to the header, and the DVD burned. At the height of the tape and DVD duplication, 35,000 tapes or DVDs were duplicated in a single year.

When the C-SPAN Video Library, an online digital collection of the entire archive, came online and provided free video playing, DVD sales fell off rapidly. Downloads began to be offered to those who wanted their own copies of material. These downloads were totally automated; there were no physical media to handle. Requests for downloads continue to grow each year, but are a fraction of the demand for physical media in past years. High-quality downloads of small video segments for licensing are demonstrating rapid growth as media companies and documentary producers discover

the online archive. Online viewing continues to grow; 23 million videos were viewed in 2013 alone.

Shifts in Hardware

The initial recordings from 1987 to 1990 used "prosumer" quality video recorders. These were operated by timers and changed twice a day. Beginning in 1990, we installed professional quality SVHS recorders outfitted with tape loaders. These machines could be controlled by custom-written software that would start a machine, rewind a tape, eject a tape, and load another tape. Recording alternated between a deck that recorded for two hours and then sat idle for two hours while a second machine took over. Another set of these machines recorded the second network. The VHS machines were used for backup redundancy. Later, a set of four additional machines controlled by a different computer was installed to provide an identical backup system. The backup tapes were kept for 60 days and then recycled.

In 2002, the first video servers were purchased. These were Grass Valley MPEG2 encoders and servers. Two of these were purchased. One, primarily used for recording, had six inputs and two outputs. The other, used for duplication, had six outputs and two inputs. The two inputs were used to capture a backup recording stream. The recording format was 10 Mbps MPEG2. Every time a file was created it was named by the date and time to the second and network. This continued the same convention that we created for the two-hour tapes and had the added benefit that all our software that related content to media continued to work with little modification. Because the files were too large to play at the desktop, we built a proxy video system for cataloging. This was a separate recording system in MPEG1 format. These files much smaller and could be played from within the database at the desktop for cataloging. Thus, we were freed from physical tape handling and had immediate access to the video. Finally, we installed a Real® video recording system to provide video quality to be streamed on the web. This format, no longer used, was the dominant form for video playing on the web at that time.

Format Changes

Since we were now recording C-SPAN's third network, the amount of recording in the archive dramatically increased. Instead of recording 8,760 hours per year per network plus an equal number of hours in the backup system (35,040 total hours), we were recording 8,760 hours per year in three formats plus backup hours for a total of 105,120 hours per year. This system was implemented in late 2002 and tape recording was discontinued. Duplication continued in tape and DVD format, with DVDs quickly outpacing tape requests.

In 2009, a new format of recording was implemented. The Archives chose H.264 as its new recording format. This format gave much smaller files with higher quality. An added advantage was that we could use one format for archival recording and quickly transcode these files into web and DVD formats with off-the-shelf equipment using free ffmpeg software. The MPEG2, MPEG1, and Real® network encoding was discontinued. Now specialized encoders were installed for each network. Two different servers were installed to encode streams separately, providing the necessary redundancy. Both one-hour and five-minute files were recorded. A Flash® player was built to provide playout of the five-minute files in a continuous piece until the one-hour files were available. An

 doi:10.2218/ijdc.v9i1.288

HTML5 player enables playing on Apple iOS devices that do not support Flash®. When C-SPAN converted to HD, the Archives installed three HD H.264 encoders.

Figure 1 shows the video cycle of recorded video in the C-SPAN Archives.

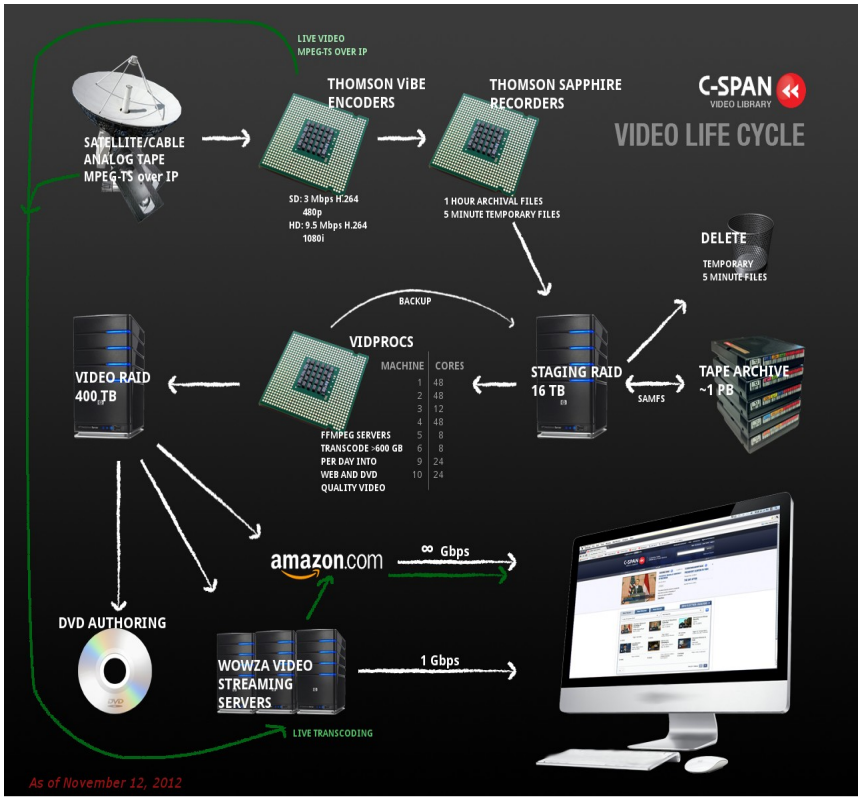

Figure 1. C-SPAN Archives video life cycle.

Indexing

Indexing the collection was a critical task that was tackled immediately when the recording began. The database fields developed in that first month are still being used 25 years later. They include the category of the program, the format of the event, the program name, abstract, sponsoring organization, and location. A nested set of keywords was developed based on Library of Congress subject headings in the first year when a librarian was brought into the project. Since C-SPAN programming is all public affairs event coverage, the scope of the content was manageable, albeit large. Figure 2 shows the arrangement of a physical record created for a single event.

Program Title Presidential News Conference
Program Date January 13, 2013
Program Sponsor White House, East Room
Program Location Washington, DC

Program Abstract President Obama held a news conference in the East Room of the White House. Among the topics he addressed were efforts to implement gun control measures following the school shooting tragedy in Newton, Connecticut, and the need to raise the debt ceiling in order for America to "pay its bills." In his remarks he called congressional Republican refusals to raise the debt ceiling "irresponsible" and "absurd," saying "they will not collect a ransom in exchange for not crashing the American economy."

Program Person Obama, Barack, President, United States, 2009-

Program Tags Business & Commerce -> Budget, Federal -> Federal Budget
Courts & Judicial Process -> Crime -> Gun Control

Figure 2. Detail of a C-SPAN archival record.

All the programming of the archive originates from one network, as recordings of public affairs events, so there was no need to note physical formats, genre, producers, and rights. Most of these characteristics were consistent for the entire collection. Instead we could concentrate just on the content. A sample of the keyword system is found in Figure 3.

Foreign Affairs and Defense	Foreign Policy	Communism
		Democracy
		Developing Countries
		Environmental Policy
		Foreign Aid
		Humanitarian Issues
		Immigration
		International Summits
		Iran-Contra Investigation
		Peace Corps
		War Powers Act
	International Security	[More keywords]
Health and Welfare	Drug Policy	[More keywords]
	Health Care Reform	
	Health Policy	
	Social Policy	

Figure 3. Hierarchy of the keyword system[2].

Another refinement to the indexing was the development of within program indexing. We were able to take electronic records C-SPAN created during the recording process and match these speakers to our recorded video record. The result was a speaker-by-speaker index within the program for a large number of new programs in the Video Library. Now, for a four-hour hearing, a researcher could move right to the speaker to see what they said. This was a powerful addition to our item-level indexing.

2 A full listing of these tags can be found at: http://www.c-spanvideo.org/browse?browse=tag

 doi:10.2218/ijdc.v9i1.288

Closed Captioning

Since 1994 the Archive has been capturing the closed captioning text transmitted with the video. This text is time-stamped by word and stored in separate files. Captions are captured for the three C-SPAN networks, plus five other broadcast and cable news networks. For the C-SPAN networks, the text is matched with the video and displayed online as a searching tool into the the video record. Matched with the speaker index cited above, it adds a powerful search tool for finding where a statement was made within hours of video.

For presidential events we went a step forward and linked the White House transcript to the video. This created the same level of word linkage, but it used the actual transcript rather than only the captioned text. The index, audio, video, transcript entry we created for presidential events provides an important historical record. We have thus pioneered in the development of a modern historical record of presidential activities. This feature also exists for significant congressional hearings where the committee has released a transcript.

Data Entry

For most of the existence of the archive, all data have been entered by hand. There are 170,000 programs and 115,000 people linked to programs 700,000 times. In order to keep up and not have a backlog, efficient tools were built to permit fast, error-free data entry. Look-up tables for people, organizations and committees allow indexers to quickly pick repeat persons from the database and add to a program. Members of committees are displayed for committees to allow fast attachment to programs. Each person can have multiple biography records that document their titles by time and display the correct title for the program date. Today, C-SPAN schedules and rudimentary data are transmitted to us electronically, but all these data need to be edited into the Archives style for consistency and for assisting with searching.

In order to assist the catalogers with scanning large amounts of video, still images were captured every six seconds and displayed in what we refer to as contact sheets (Figure 4). Each of the images could be clicked on to open it full screen and each image was time-stamped to match the moving video in the different formats. A software tool was developed to allow the cataloger to set the begin and end time of programs by clicking a button below the picture. This prevented transposition errors when entering times down to the frame of video.

Using this tool, the catalogers are able to select and archive five pictures with the program record. These are the opening and closing shots, a wide, an audience shot, and a marketing shot that is used whenever the program is displayed. As they catalog the program, they also save a shot of each person. These pictures allow the catalogers to quickly see the essence of every program without accessing the moving video. In addition, the pictures of individuals allow the cataloger to avoid linking the wrong person to a program when there are two persons with the same name and they need to determine which one is appearing. People often appear in one capacity and then could be elected to another office or represent a different group. The picture allows another level of information in the cataloging process. It also documents how people change over their public lives that we document. Over 700,000 pictures have been indexed in the database.

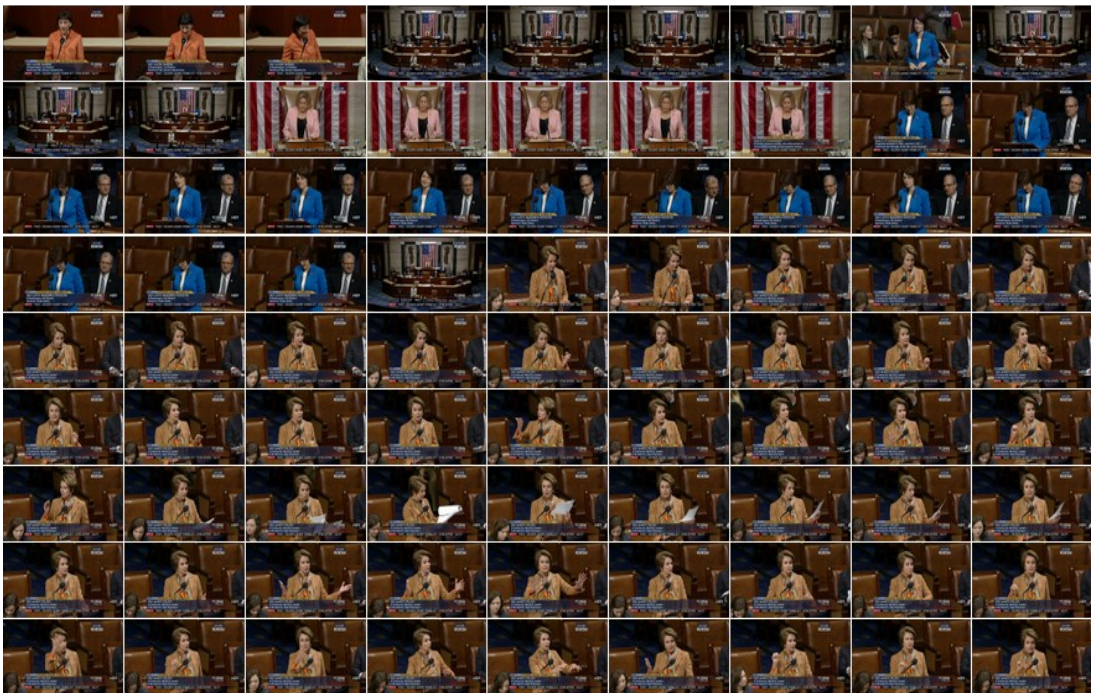

Figure 4. Archival contact sheet.

Converting to Digital

When the conversion to digital recording occurred in late 2002, there were 87,000 tapes containing 120,000 hours of first-run content in the tape archive. With the digital recording underway, we moved to tackle the problem of digitizing this material. We purchased a VHS robot that could hold 100 tapes and eight machine-controlled SVHS VTRs. An additional video server with eight inputs was purchased to handle this ingested material. Each day the robot would be loaded with tapes and throughout the day and night would play the material back into the server. The tapes were all barcoded and scanned by the robot on ingestion and file records were created for every tape. When we installed the H.264 encoders, we purchased 16 encoders for the archival material and installed a second playback robot. Now, in addition to the three C-SPAN Networks, we were recording 16 channels of archival material for about 20 hours per day. This technology allowed us to completely digitize 120,000 hours in about three years. This may be the fastest known digitization of an entire archive. As these files were ingested they were also transcoded into web quality H.264 so that they were available for online viewing.

The entire archive is backed up on a StorageTek digital tape robot. Today, each tape holds about 600 hours of H.264 format. All the web-quality video is stored on a 400TB RAID and available for immediate viewing. This video is also available in the Amazon cloud service (AWS) for backup and viewing. When demand for online video is high, we serve the video from AWS to cut down on bandwidth consumption at the Archives and to provide better user services. The digital tape robot stores two copies of all the content on separate tapes. In addition, a single copy of each tape is stored in Iron Mountain® remote storage.

 doi:10.2218/ijdc.v9i1.288

In 2010, the archive went public with a newly designed C-SPAN Video Library. The Video Library contains over 200,000 hours of indexed content all playable in H.264 files. The player allows clipping and sharing. The MyC-SPAN feature allows logged in users to receive program alerts based on people, organizations, committees, or tags that the user wants to follow. This feature also keeps track of all clips made by the user and allows for future sharing. The collection is indexed by program name, speakers, affiliations, organizations, committees, and tags as well as all words in the abstract or in the closed captioning.

A separate section of the site is called the Congressional Chronicle, which indexes every session of the House and Senate proceedings. For the first time, the Archives has an index of what actually happens on the House and Senate floor. The government-published Congressional Record allows members to insert remarks. There is, for example, an oft sought after speech that the Senator Barack Obama made on the Senate floor about the debt limit. However, this speech was never given and there is no video in the C-SPAN Video Library. Now, our index provides an easily accessible source to find when, how many times, and on what topics a House member or senator speaks on the floor of the Congress. The video is also linked to the printed Congressional Record so that one can search on the words and be immediately linked to the video.

Conclusions

The C-SPAN Archives has been very successful over 25 years through its effective use of technology and technical design. As we move forward into the next 25 years, we anticipate more within-program indexing and linking to text. Our servers and our read-only databases now operate in the cloud. We also use remote site computing for backup and for serving high-demand video. Cloud computing provides greater capacity and reliability over in-house servers as the size and use of the collection continues to grow. We are also seeing more opportunities for data exchanges through application programming interfaces (API) that we develop. We currently provide network schedule information to the Interactive Programming Guides (IPG) used to provide schedule data to set-top boxes and newer devices that are coming on line. We expect to see more development of APIs as new interest comes to our database.

This project has evolved with technology over the last 25 years. By utilizing technology at the beginning, we kept pace and never fell behind in the collection, organization, and management of the information we recorded. Today, we have almost a petabyte of video and related data. As new technology emerged, such as digital recording and storage, we implemented it and then adapted the newest technology as it came along. The result is one of world's largest online archive accessible free on the Internet.

At the same time, there were bumps in the road. This article addresses the technical developments. In the technical areas, archives do not change formats quickly, so the decisions to move from VHS, to SVHS, to MPEG2, to H.264 were significant. The key is to get the video digitized and linked to indexed data descriptors. Once it is digital in a nonproprietary format, it can easily be transcoded into alternative consumer formats.

Most of the problems encountered had to do with miscoded data. In the early days, when duplicate video was deleted, these errors had greater significance. In the technical area, few decisions other than format and database design had long-term implications. Understanding how people wanted to find and use the data shaped the database design.

The C-SPAN Archives collection was widely used from the beginning. The database was initially a tool to allow the staff to find and retrieve information. As the technology developed, it became a public database that allowed users to first find the video and subsequently to watch the video. The initial database design was for internal use. As the database became public, the feedback from users shaped the design of the database and search algorithm. C-SPAN conducted multiple focus groups to get input on design and search.

It is hoped that the exposition of the technical issues here will help others who are organizing similar large collections of data and video. Video collections are increasing daily and present challenges of size, organization, access, and retrieval. The experience of the C-SPAN Archives will not be relevant to all of these collections, but there may be some similarities that will apply in part to help others wrestle with these issues.