NewsScape: preserving TV news

Tim Groeling | Oct. 13, 2016 | UCLA Charles E. Young Research Library
Dodging the Memory Hole 2016

TIM GROELING: ***[00:08]*** So, I'm going to be talking about the NewsScape Project, which is put together by Professor Francis Steen. I got this picture in part because it is one of the rare pictures you'll find of Paul Rosenthal that also has Francis over his shoulder right there, so it's a twofer, but you actually can see Francis live right there, if you want a better quality picture. Francis has been the director of our Communication Studies archive for several years. I was previously chair of our department and I'm leading the analog portion of the collection, the digitization of our old analog collection, which was started by my predecessor as chair and the founder of the archive, Professor Paul Rosenthal. It's kind of interesting.

[00:52] This is a rare picture of Paul Rosenthal, speaking of archival collections, I think I've only seen one other photo of him from the 70s and he looked exactly the same, so infer from that what you will. The UCLA Library has been a big part of our collection, especially in recent years. They have helped support the collection and they help host the collection. Huge amount of files are produced by this that they are in fact helping us host. If you want to go and see the collection, I think you can even see the existing digital collection now, since you are on the UCLA domain @TVnews.library.UCLA.edu, although you should pay attention to me instead, temporarily. You can do it after we're done.

[01:29] Other supporters include UCLA's Chancellor, the dean of the Social Sciences, Arcadia Fund, Social Sciences Computing, the NSF and some other funding organizations. The collections we have, our oldest collection dates to the 1950s. It was a collection of audio recordings of campus speakers. So, when Hollywood celebrities or politicians or media figures would come to campus, my predecessor at the Department of Speech, before we were the Department of Communications Studies, would go and record those events live and we'd digitize those. Originally we were just going to host those on our website to celebrate our department's 40th anniversary.

[02:10] We ended up deciding to host them on YouTube, which has been an interesting experience. I say "interesting," in part, because we've gotten a lot more traffic there than I would have thought was possible, and most of it is organic from YouTube. We haven't put any money in advertising. We don't promote this very much, other than a few links on our website. This is our analytics they give us, so we've had over 7 million views, which is... I'm sorry 728,000 views, 7 million minutes. If you look here, by the way, that big spike, we had a speech by Martin Luther King he gave on campus that will be posted on the 40th anniversary of that visit. That got picked up by local media and was a large part of the hits, but we also have a lot of other people just coming across a variety of speakers.

[02:54] As you might expect from YouTube, one of the issues we've had is the commenters on YouTube, the notorious commenters. When we have campus speakers that used to be more... we'll just say "diverse" than they are in the modern era. It was common to have the founder of the American Nazi Party come and speak to the college campus in the 1960s, which leads to commenters online that you would expect when you have the founder the American Nazi Party on the video. So, things like that have been a problem, and we have had some copyright issues, some of which we've been able to resolve by pointing out to the organization that, in fact, this is a UCLA speaker on the UCLA campus who is employed by UCLA while they're speaking.

[03:30] So, you might be violating our copyright. Step down. Situations like that are kind of fun. The NewsScape project though is mostly what I'll be talking about today. It's the largest collection we have, and it basically was intended to be TV news and public affairs programming, both local and national. It was started during the Watergate era. As you heard previously, the people at Vanderbilt have been doing an excellent job cataloging the three major evening newscasts since the 1960s. But during the Watergate hearings, it was clear to Professor Rosenthal that there was a lot of TV information, that was civically consequential, that was just being lost.

[04:07] Local news was one example, but these hearings if you're just collecting the three major evening newscasts, you were losing a lot of raw information. Using a very, very thin budget, $10,000 a year plus volunteer labors and donated equipment, it grew to become, in 1979, a recording of all the local news and national news one would see if one lived in Los Angeles and had a television. It was intended to be a complete collection, not of necessarily local news, but at least local news in L.A., and then all the national news, not just the evening newscasts. In 2006 Francis led the effort to go straight to digital, so you're talking about born-digital recording.

[04:42] That's when we really started taking things off of the videotapes and, instead, recording them on hard drives. Since 2006, we've expanded the collection to cover a lot of other cities. I'll tell you basically what we have. Prior to 2006, though, we don't really know what we have. It's kind of spread over multiple organizations, if you can imagine the volume of videotapes that we're talking about here. You might have a sense of exactly how disastrous that is. My department has changed offices four times since I've been here since 2001. Each time something else has gotten sort of moved someplace else. We're still not sure what we have. There's good records for some portions, but basically Paul's belief, in some sense, was: "I'm going to record everything I can and then someone in the future will take care of this."

[05:30] It's important to save it, and that's exactly right if you have limited resources, but now we're the people who have to figure out what's actually on these tapes. There's also issues with — we get the tape and, kind of, know what should be on the tape and then it's preempted by a football game or someone actually lost power and didn't know it on the VCR or fun things like that. We won't really know what we have until we're done. It is kind of fun. So the tapes, earliest ones, actually we have some reel to reel and I don't know what to do with those. We have about 500 U-matic tapes. We have about 50,000 hours on Betamax, which I'm not looking forward to at all.

[06:05] Then we have a late period that we're focusing on now, with about 160,000 hours on VHS, with some redundancy, where we have duplicate recordings for that time period. That's going to take us a little while. The preservation — we actually are starting first with the VHS because their the most threatened. The VHS, ironically, even though they're newer, were recorded during a period where cable TV expansion and also some budget issues meant that we were cramming eight hours of recording on consumer grade tapes and consumer grade recording equipment.

[06:41] Actually Betamax was... better, I think, in some sense, as a recording medium, as you might have heard from other people. Those are holding up really well. The U-matic was professional grade equipment, very expensive, high-quality tapes. Those are holding up well, but our consumer grade VHS stuff is really quite threatened. We're currently working to pull things off of there, and we also are finding things like...that VCR was failing for quite a while before they recognized it and took it out of service. We have things that are kind of out of adjustment. Up until, literally, this month, even our faculty offices in my department didn't have air conditioning, so our tapes didn'thave air conditioning for long periods of time either. They got it before we did, so that's where our priorities are in life.

[07:59] It's basically run by me, when I'm not teaching or researching or serving on endless committees, which seems to be half my life. I have one part-time alumnus, who is in a paid position,

temporarily, and then work study students. Then Francis handles the files after they're out of the collection. Switching to Betamax, I am really dreading because, I don't know if you've gone on eBay recently. Betamax VCRs are expensive and seldom work. Our lab currently, we have 22 digitization stations, basically which are a VCR, an encoder, and a computer. We have local RAID storage that we use to temporarily hold the files and we use Filemaker as our data solution to manage the information about the files, the metadata.

[08:45] Basically, it's nice that we're using really cheap computers. You don't actually need, if you have hardware encoders, particularly good computers. We do not have good computers. We have very bad computers. I think some of the ones that we got off of eBay were used in "one laptop per child" school situations and dropped repeatedly and such, but they still work. We're using hardware encoders. We're going to MPEG 2, initially in part because we had trouble finding MPEG 4 encoders that will maintain the closed captioning. The closed captioning is a vital part of our indexing scheme. We couldn't give that up.

[09:21] If you care, JVC, VHS VCRs seemed to be the best ones for us after testing. Actually, consumer grade ones, because we found that some of the professional decks weren't very good at the super long play playback because they just weren't designed for it. They didn't really expect people to be using those Pro machines for extremely long play tape. We actually then use Filemaker and simple barcodes and have the students use their cell phones to do our inventory system, using barcode pictures with their phones. That actually works out pretty well and we didn't need much equipment for that.

[09:57] We use the Hoffman cluster, here on campus, to extract the closed captioning and compress the files to H264,which is our informat. Progress to date, this was one of my final priorities when I was chair was to deal with this huge backlog of tapes we hadn't dealt with. I started that process in Fall 2015, did some configuration testing, and we really just started working on this in winter 2016 and ramped up, moved to Filemaker in the summer. So basically, we've done about 36,000 hours of encoding so far this year. There's a step in the process that I'm dreading, which is going back and figuring out, based on the very limited information we have about each tape, what the individual shows are, splitting them up and dating and getting the time on them set correctly, which is a nightmare.

[10:44] Anyway, we'll deal with that later. Lots of other problems in this process: the original VCRs that were recording these were overworked and very cheap and gradually got further and further out of spec. The nice way we talk about that is to say they had "personalities," very strong "personalities," and the VCRs that we've bought also have personalities. Actually, if you find compatible personalities, they play back really well, and if you don't, they play back really poorly. We had to, over time, do some testing and figure out which of our VCRs were most like Paul's original VCRs, and find that love connection.

[11:21] We have lots of issues, actually, that librarians, hopefully, can help us with, about how to handle the changing names of shows or it's the same time, same host, but different name. Is it the same show or not? That sort of information is going be kind of difficult. We have issues like, why are we getting audio buzzing? It's when you put a computer directly on top of VCR, you get RF interference. We're actually using a layer of dead VCRs to protect our good VCRs from the computers. Lots of other problems. If you'd like, and can get me some alcohol, I'll give you a lot of problems. I'll tell you about everything, but for now that's just fun.

[11:58] Most of our problems, if it's a problem, we tell by the quality and try again, as best we can. Some of the ones we're just frankly not going to be able to recover because nothing was ever recorded or such. The digital collection is actually, I think, much more interesting, potentially, for this group because it is going straight to digital and includes a lot of material that starts off as digital material. Right now, we're covering 46 networks in the U.S. and beyond, over 2000 series. I would say, actually, that this is data from a month ago. Francis could probably update this. We're nearly probably up to 400,000 video files, over 300,000 hours in duration, by this point.

[12:32] Closed captioning files are a big part of our index. We use them to basically find material based on text. We're also doing optical character recognition of all of our files, which is very helpful for knowing the topic that's being talked about, across long periods of time, and also other news that's happening at that time, useful also for separating out commercials. We also do thumbnail images, periodically, so you can visually search our collection and find things, even if there's not closed captioning, for a particular show in it. Again, we have access. They are using a search engine the library has helped us set up. In terms of unlocking the content, we view preservation as the first step, and we don't really want this thing to necessarily be the world's best DVR.

[13:16] Part of what we see as our value added here, especially as an academic department involved in this, is helping provide the tools that help people understand and interpret this data, as opposed to just watching the shows again. Also, we'd like for them to be able to find and share. For instance, it's good to know if you search for Barack Obama or Obama: "Oh, there's 155,000 videos or so. I will watch all 155." No, you're not going to watch all 155,000 videos of Barack Obama. It might be more useful to know tools, and see the coverage over time and visualize patterns of news and see the forest, not just the trees.

[13:52] We're working on some of these things. This is a quick example of visualizing where candidates were getting news coverage. What the terms were that were associated with them. Which candidates got more or less coverage over the course of the 2012 election. I should mention one of my colleagues, Jungseock Joo, who is not here today — I don't see him — has been helping develop some of these tools, as well. This is another one he had that, CNN is at the bottom of the access, Fox is on the upper right hand corner, MSNBC is in the top left corner. You might not be able to read the terms very carefully, but this is over the process for the time period where the Affordable Care Act was being debated in Congress.

[14:29] You can see, over time, what issues some channels talked about more than others, or terms some networks talked about more than others, change gradually. Interestingly, MSNBC tends to talk about Republican members of Congress more. You see Fox talks about Democratic members of Congress more in their coverage, and talks about things like trillion, mandate, wreck, premium, subsidy exchange, implement. CNN is talking about Angie for some reason, I don't know why, ceiling, negotiate, default, shutdown, debt. MSNBC is talking about continuing resolution, Ted Cruz, repeal, and some other people.

[15:06] Again, you can see how the patterns shift over time. You can start seeing that as it's actually introduced, you start seeing the navigator being talked about on Fox, canceled, deductables, things like that. Seeing those patterns over time is, I think, probably more useful than having, necessarily, people being able to see the individual videos that went into generating that. Developing those tools is part of our mission. We also have been trying to help people analyze visuals. This is, again, an area my department is trying to build expertise in. We already have the ability to recognize named entities in the collection, parts of speech, and topic collection, which I'll talk about a

little bit later.

[15:42] Sentiment is a little bit harder in the text, and visuals are challenging, but we're doing things like facial detection and facial analysis and being able to scale those more and more, over time. For instance, we can, potentially, understand patterns of news and do it to this automated coding and be less subject to subjectivity than if humans do it, building on machine learning and big data tools. We're very excited by that. We presented a paper at this year's American Political Science Association convention. We're doing things like looking at presidential candidates, locating their faces, and then looking at what proportion or how often they're shown smiling versus not smiling at the news, over three election cycles.

[16:42] We can find — there's my little face validity joke here — we're generating a score of how much they're smiling based on this and then setting a cut point of higher than that level of confidence, they're smiling. If you have any curiosity, actually Hillary Clinton was smiling a lot more this election cycle than Donald Trump. I'll just leave that where it lies. We also have a site that you might be interested in, Viz 2016.com, viz with a "z," where we have set up some interactive tools based on both Twitter and our collection to look at how much coverage different candidates got, the sentiment, at least on Twitter, of some of that coverage, and have done some topic tracking tools.

[17:02] Again, we're pretty excited about this tool, where you're able to see in both individual channels or more broadly, what the topics are that different news organizations focus on and how those topics evolve over time, who the main figures are that were associated with them, and what those topics are and where they're located. That's, I think, in the broader collection, going to be a very fast and accessible tool to see these changes over time. We also have done another version that's interactive, where you can do that same sort of triangle I showed you earlier, but just focusing on presidential candidates that you select or news organizations you select to show their coverage.

[17:44] This is Donald Trump for the month of June and looking at what FOX versus CNN versus MSNBC tend to be covering. For some reason, CNN did a lot more stories talking about Mexicans and Donald Trump than other channels, not MSNBC. If you go to the equivalent of Hillary Clinton, this one dot, way over by Fox, is talking about her e-mails, so you can see patterns that maybe you have intuitively and see them over time as well. We'd also like to — although we haven't developed these tools yet — help people, potentially, share what they find in the collection.

[18:58] It would be transformative in sharing that online. Again, we're fairly early in the development of those tools, but that is something we think is a nice use of this collection. We have preservation as the goal and we're also starting with preservation. We would like to see this collection be one of the premier tools for understanding big picture changes in media, over time. I think we all realize there have been huge changes in the news media, over the time period we cover, and I think it would be fun to study that. We would appreciate any advice or funding leads you have to help us do this collection.

[19:35] Again, we're doing this very efficiently, but it would be nice to have more money. Again, the Beta just scares me. $400 for a really terrible VCR these days, if it's working. That's all I have for today. Questions? You taking notes Francis? Who is that again?

[20:15] That's already a good question. Yes.

[20:45] So this is the other reason why I'm terrified of the Betamaxes is those are three hours or fewer tapes. Right now, the nice thing about the VHS is that we're relatively low labor because the students will come in and start the recording process and enter the metadata we have for that particular tape and make sure everything's working relatively well and leave. Then we have the lab unsupervised while all 22 machines are encoding and then another student comes in at the end of eight or nine hours to make sure that the files have exported correctly, that the information is set and it's on the server, and does a quality control check that either makes us know we have to do this tape again because it didn't work right, or sends it on the process so that Francis can process it.

[21:24] That process is actually relatively straightforward. I mentioned we haven't actually done the step of slicing up the shows into individual portions. That's, in part, because our collection is not currently well-ordered, in chronological order, and we only know, with some gaps, what's supposed to be on some of the tapes. We don't know what's supposed to be on most of these tapes. The only information that's on the tapes themselves, for this period of the time, is the date that it was pulled from the VCR. That doesn't even tell you what the date of the programming is, and the VCR number it came from. Normally they followed a regular program and some of them are very regular like, there's a CNN for multiple hours per day, that just is that.

[22:03] As long as CNN doesn't change the name of their shows, which they seem to do constantly, just constantly, then we can know what that is, as long as we know the day it came out of VCR, but there are other ones where there's breaking news or other events where we actually don't know what's on the tape until we view it. The plan is to get everything from that sort of time period encoded and then start working back in time, taking advantage of the similarity in the same VCR over time, and also resources like archive TV Guides and such to try to figure out what's on each tape.

[22:34] Actually, the plan is to involve our alumni and crowdsource that online, and have people who are retired and knowledgeable about news and very generous with their time help us by cutting the split points and identifying the shows online, using a custom program we're setting up but that's still down the road, as well. So, actually, one of our works students is interested in learning VCR repair, which is like what you see on late night TV ads or such for ITT tech. We're fortunate, and one of the reasons why I selected JVC is that was what the campus classroom VCRs were set to and we were able to find somebody who is still employed at UCLA who has expertise in repairing JVC/ VCRs who's giving us a really good rate for doing that.

[23:47] That's been a very useful arrangement recently, as a matter of fact... I guess I was the last question.