VAN Archival Team

Archival Best Practices Recommendations
&
Archival Survey Summary

May 2015

This document's online link:
http://bit.ly/VatBest2015

TABLE OF CONTENTS

OVERVIEW

The 25 members of the Vermont Access Network have a large and ever-growing collection of video assets that are unique and valuable. The VAN Archival Team (VAT) was formed in 2014 to develop recommendations for these 25 VT Community Media Centers (CMCs) to archive their assets.

Over the past year, VAT (drawn from staff from seven of the 25 VAN stations) has met on a monthly basis to hone in on the archival problems CMCs face and to research archival topics and available resources. VAT also conducted a survey of CMCs in Vermont (22 to date) from late November 2014 to early January 2015, asking them each about their archival practices and needs. The survey results are found throughout this report.

In our research, VAT has identified a number of common archival challenges and recommendations toward the goal of developing a statewide archival system with a built-in web interface.

1. Common PEG CMC Archival Needs & Challenges

While station size and resources vary, a trio of **needs** are apparent:
1) Organizing video assets in-house
2) Identifying/preserving/organizing video assets for the future
3) Sharing video assets with the public

Common archival **challenges** include:
1) Staff/volunteer time to organize assets & enter related metadata
2) Storage choices, cost and physical space
3) Linking metadata about video assets to assets themselves
4) Keeping up-to-date about archival tools/options
5) Processing/storing both older analogue assets and digital-born video

2. Archival Planning Considerations

The following table includes general considerations for the development of an archival plan and the ongoing management of assets and metadata:

VISION	CMCs need to develop and scale an overall policy for which new content will be preserved and made available to the public. This includes making predictions about the future value of assets and exploring possibilities of sharing the archival responsibilities with stakeholders (ex. municipalities).
LEGAL CONCERNS	Responsibilities of both producers & partner organizations with regards to future of video materials should be specified in contracts. CMCs need to take legal rights of content into account from the start to create archives that will be accessible to public (e.g. restrictions on some musical content).
DISASTERS	CMCs need to include data recovery information and procedures in a Disaster Plan.
MIGRATION	CMCs need to plan for inevitable, constant consistent/ongoing migration of content.
BUDGET STAFF TIME	Labor time, identifying the best archival tools - as well as resistance to changing existing workflows - are often big challenges. CMCs should consider how archiving could best fit into their existing workflows and how to make use of automated archival tools.

3. Metadata/Cataloging

Managing metadata requires an investment of some time, but it is essential to have some basic info about each piece of media to help identify it, catalog it, and, ultimately, archive it. Here are some recommendation basics:

- CMCs should decide on and stick to a consistent file-naming protocol that works for them for naming video files. Eg. CubanBridgeEp3033.
- CMCs should also use a unique cataloguing identifier for each asset/program. Eg. 3022
- CMCs should pick whatever recording system they have available to record and store this metadata - e.g. Excel, Numbers, Google Docs, in-house database, web application, website database, broadcast server database, etc.
- Each video asset should have <u>at least</u> the following metadata associated with it:
    (1) ***Descriptive metadata*** - Title, creator, date created, location, type of video, rights info

(2) ***Source metadata*** - Physical characteristics of source video- i.e. S-VHS from Burlington Town Meeting 1999
(3) ***Technical metadata*** - File name, format, codec, file size. (Ideally this is automated.)
(4) ***Legal/rights metadata*** - Asset should be assigned provisions for future/web use.

- We recommend a metadata standard like PBCore and a backend Application Programming Interface (API) with the ability to deliver machine readable metadata to search engines via HTML using https://schema.org or similar. To gain the most utility from our collective archival metadata and content across the state we should seriously consider moving to a metadata standard that allows us to easily discover and retrieve content across channels and Community Media Centers.
- Consider metadata options at the beginning of any project. Some stations require volunteer producers to enter the info about their program before submitting the actual programs themselves.
- Developments in voice-recognition software may be a powerful tool in future years, with searchable transcripts automatically created to accompany video content.
- For more info about establishing a cataloguing/metadata system, see **Appendix: Handy Links.**

4. File Types/Codes

When it comes to the best format/codec used to save files to, most CMCs are governed by 1) their playback format 2) their editing system 3) their web formats 4) their DVD formats.

When we talk about encoding media with a codec, we are really talking about using a codec to reduce the data rate of the recorded digital signals for transmission and storage. In the context of archiving, this is also done to preserve the perceptual quality of the video. See **VAT Archival Best Practices: Additional Technical Info online document** (http://bit.ly/VATBest2015Tech) for more information and links concerning the hows and whats of video digitization with codecs and video codec comparisons.

While MPEG2 with mono MPEG layer 1 or 2 audio is the most commonly archived format, it has some shortcomings. It is one of the larger formats for a particular level of quality and often, while required for broadcast television, does not offer the best quality compared to rate of compression (downsizing of the file). Often, the MPEG2 PS or mpg/mpeg file produced by the mpeg2 codec used is very lossy, and may be stuffed with empty packets for when it replays as a transport stream over the network. Comparable bitrate encodings with more efficient codecs (H.264) have a much better Peak Signal to Noise Ratio (PSNR), and are closer in quality to the original copy.

While archiving our materials in Library of Congress-recommended formats, such as Motion JPEG-2000, would be desirable with its intra-frame image encoding, it is not widely supported outside of digital preservation organizations, and it has the very old and inefficient lossy compression algorithms from JPEG. H.264 may be the codec with the highest compression and most desirable file size with robust hardware and software decoders available. We could also entertain the idea of using H.264 with intra-frames, but we lose valuable compression, which shrinks necessary storage space. So our trade-offs are related to general interoperability, file size, encoding speed, and suitability of the file for editing at some later time. A study of intra-frame only H.264 encoding, file quality, size, and editing usage would be useful.

For HD we currently recommend MPEG4 part 10 (a.k.a) MPEG4 AVC made with the H.264 video codec (x264 encoder preferred), the AAC audio codec (fdk-aac encoder preferred), and the .mp4 file wrapper. Audio bitrates should be 128 kbps or 192 kbps, Video bitrates should be 3000 kbps at a minimum with the main or high AVC profile selected. The profile level is up to the CMC, but we currently encourage less sophisticated options to improve file compatibility and decode / encode speed.

Not all encoders are created equal, and so it is important that before you choose a particular bitrate for audio and video you take the time to look at your resultant encodes, and verify that the quality meets your expectations.

MPEG-2 audio and AC3 are some of the more common audio formats. Mp3 files are perfect for podcasts and audio backups of shows. Tip: A Zoom H4N recorder costs $200 to record sound to just about whatever format you want and archive it as a separate file in a folder with your show's video file.

5. HD Future

Vermont stations are increasingly entering a High Definition world, which requires by definition more storage space for both storing raw material and final programs. HD is now the foundation for nearly every phase of the video asset lifecycle, including production, post-production and distribution. As a general rule, CMCs should archive video assets in the highest quality possible, so to facilitate the reconstitution of a range of variable quality versions should the need arise.

As bandwidth resources continue to become more readily available at lower costs, HD video delivery is becoming more prevalent over non-linear platforms. CMCs that archive HD video assets in flexible file formats (see VAT recommendation above) will be better prepared to meet evolving consumer expectations.

In the future we may expect to move to HEVC H.265 video with AAC audio or royalty free Daala video with Opus audio.

6. Processing Analogue Materials

Ideally, materials from old analog formats (S-VHS, Beta, U-matic, mini-DV, Hi-8 etc) need to be digitized and catalogued, while originals are kept in climate-controlled storage.

- Materials can be digitized in house with a well-maintained deck employing capture software (e.g. Quicktime, VLC, or editing programs like Premiere). Use S-video cable when possible.
- The goal in capturing is to maintain source quality while achieving manageable file-size. High efficiency codecs are essential. Bitrates can be reduced because signals lack the resolution of current HD or SD workflows.
- Digitize in high enough quality to suit a variety of uses, such as online streaming, cablecasting, DVD mastering.
- While it can be costly, there are also professional digitizing services.
- See **VAT Archival Best Practices: Additional Technical Info** for file types/codec recommendations.

7. Storage Scenarios

In the context of archiving we are considering storage that is considered “cold” or “warm”, meaning that it is accessed far less and with lower performance requirements than storage used for the editing of content in an NLE or for delivery of content to end users over the Internet. Since we are storing files in an archival storage system, we are balancing reliability, ease of use, speed of operation, total aggregate storage needed, and cost. We reduce cost by using codecs to reduce file sizes, while preserving quality (PSNR).

- CMCs are currently running the gamut of digital storage options:
    - Stand Alone Hard Disk Drives (HDD)
    - RDX or LTO cartridges (Digital Tape Library)
    - DVD’s (Video or Data Optical Discs)
    - NAS Appliances (Drobo, Synology)
    - Servers with RAID Arrays of HDD (Dell, IBM, Supermicro)
    - Software defined storage clusters (ceph running on multiple servers with HDDs)

Ideally each CMC would have sufficient bandwidth to minimize the storage requirements at their facility, and we would collaborate on centralized storage in one or (preferably) more locations across the state, leaving physical archives (Digital Tape, DVD library, HDD library)

and a cloud option e.g. (archive.org, AWS Glacier) to be managed separately by CMC staff for the remaining copies of the archival data.

For on-site “warm” archival storage as we consider statewide collaboration:

- We recommend that every station have at least two - but better 3 - copies of each video asset and associated metadata, and two of those copies should be in different physical locations.
- We recommend that at least the two copies in different physical locations have at a minimum some form of Networked Attached Storage (Drobo / Synology), or availability for download / upload of their data. If that is not possible, it means that people will have to physically make a routine out of moving physical digital media periodically from one place to another. All stations should strive for at a minimum one copy of their archival data on a network. Bandwidth at remote locations or CMCs may also limit the feasibility of this networked option. Some collaboration makes sense with the idea of larger CMCs with more bandwidth and aggregate storage supporting smaller CMCs as one of their remote sites.
- We recommend any NAS, Server, or Desktop used as archival storage use a Redundant Array of Independent Disks (RAID) to achieve a higher total volume of storage than any single disk, potentially better performance on reading and writing data, and the ability to sustain one or more disk failures (depending on RAID level) without losing data. For sets of two disks, we recommend “RAID 1” , For three disks to five disks “RAID 5”, for more than four disks, but less than 8 “RAID 6” and for more than 8 disks, a “RAID 50” or a “RAID 60” might make sense. RAID uses up some of your storage space to store parity information, so you will store less than if you used the single disks separately. RAID is not a substitute for backing up your archival data. Calculate your storage with RAID levels.
- Choices of HDDs for use externally, internally, or in a RAID are important. We encourage the use of SATA HDD drives with a minimum 3 to 5 year Warranty, 5 preferred. Always check with the NAS manufacturer or hardware RAID card manufacturer for compatible models. Always buy one extra disk so you have it on hand as a spare if a drive in the array should fail, or designate a hot spare if your hardware supports it. We discourage the purchase of single HDD enclosures or drives from “Big Box” retailers as these generally have the lowest quality drives, with a warranty of 1 year.
- If you choose to use a rackmount server, keep in mind it will generally have much higher cooling requirements and will generate much more noise than a NAS, live studio switcher/server or playback server. Your CMC space may not be suitable for that type of equipment.

8. Sharing Content

Delivering our content, (both archival and otherwise) to clients across the Internet is paramount if we are to remain relevant in today's multi-platform media market. We are beginning a transition where our cable channels will likely be considered a secondary output of our CMCs, and the web presence of our content the primary output. As we move forward, we no longer just hand a signal to a cable system but we must manage, or pay for either directly or indirectly, the delivery of our content over the Internet.

We all do this in different ways, using different services, tools, software and hardware. As this flip in priorities takes place over the next few years, Vimeo, Archive.org, Youtube, playback ecosystem-based solutions like Cloudcast and PEG Central, as well as custom in-house systems make up our current methods.

To the extent that we want our content to be discovered in an overwhelming sea of web-based media content, we suggest a hybrid strategy of using the existing services above, while evaluating the ability to collaborate on a system of metadata standards and backend Application Programming Interfaces (API)s that allow for content discovery and search across the entire archives of the state, over the Internet. We also suggest exploring options to provide CMCs with increased bandwidth between our respective sites as well as between our sites and the Internet, so that we may more easily move files and live streams around the state.

Taking on a regional or web scale, on demand content delivery system will only be feasible with higher bandwidth access to regional fiber networks, access to national backbones (possibly via Internet2) and significant physical infrastructure and staff skillsets including: network engineering; storage engineering; system engineering; web application development; operations; and media streaming and encoding specialists. The purpose of projects like the Civic Cloud are geared to lower the barriers of entry for non-profit or public interest organizations delivering content over the Internet. Taking on a regional live streaming system and archival metadata system within the Civic Cloud may be more accessible and appropriately scaled endeavour at this stage.

In the future, with a common set of archival renditions that can be transferred over the network, we have the necessary basis for encoding lower bitrate renditions suitable for on-demand delivery over the Internet. Provided we can have the metadata API's specify the location of these renditions on the web, then we are in a position to deliver them both together.

9. Conclusion

While most Vermont CMCs are maintaining archives, they can take several steps to preserve materials in a more effective way for the future. These Archival Best Practices recommendations are geared to make this challenge easier.

With collective materials conservatively projected to grow to a total of 700TB by 2020 and the shift to HD underway, we recommend the following general steps be taken by Vermont CMCs: **budget and plan** for growing archives and networked “warm” on-site archival storage; **consider alternative codecs** (including H.264 and next generation codecs) for archival purposes; **digitize existing non-digital content** according to these; **maintain multiple copies** of assets; consciously **maintain file-naming protocols and metadata info**; **use cloud storage** for backups; **consider opportunities for larger collaboration** on metadata standardization and unified ways of storing and presenting that data for public use.

As a whole, VAN CMC’s are at a perfect time to address our collective archival needs with collaboration and some standardization to leverage the value of our important, diverse, and quickly accumulating content into the future. The VAT will continue to update its knowledge base and investigate options for a statewide archival system with a built-in web interface for the longer-term.

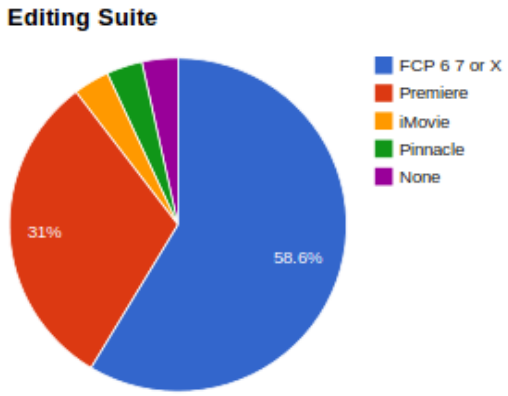
VAN ARCHIVAL TEAM

Angelike Contis (MMCTV), Andrew Crawford (CCTV), Drew Frazier (RETN), Joanne Montanye (NWATV), Rebecca Padula (LCATV), Jeremy Perkins (GNAT-TV), Alex Reichert (CAT-TV).

VAT Archival Survey Summary:
A Snapshot of Current VAN CMCs Archival Practices

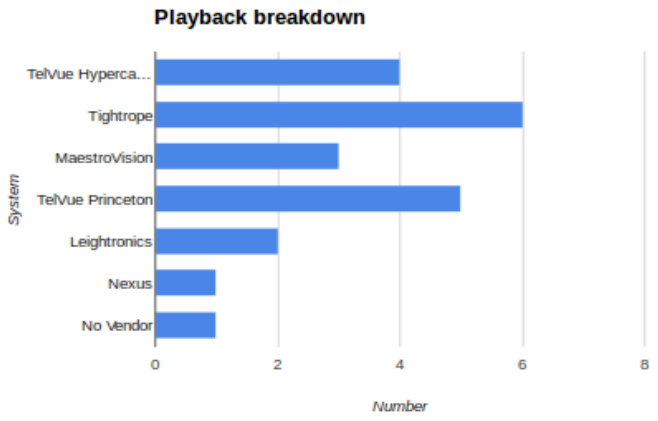
These are some results by a survey taken by the VAN Archival Team (VAT) in late 2014/early 2015 of 22 VAN CMCs.

Editing Suite



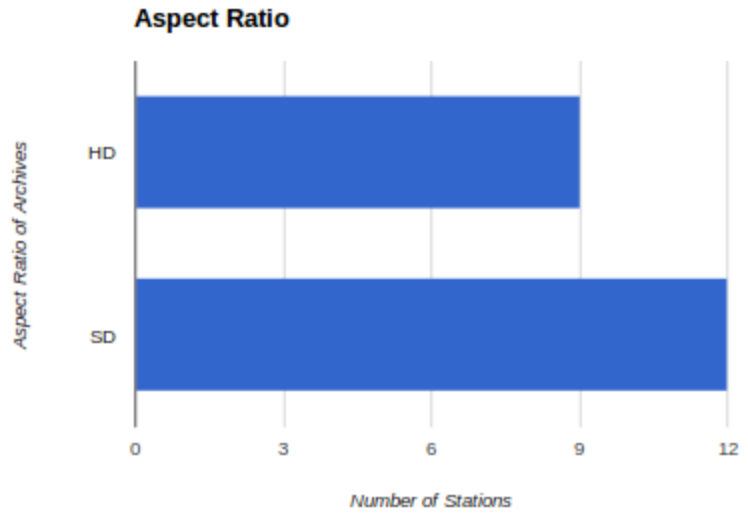
Existing workflows for archiving currently seem to be paths of least resistance steered by a combination of editing suite functionality, playout vendor requirements, video recording dimensions (aspect ratio SD/HD), current web databases, and any pre-existing system for cataloging content on legacy physical playback mediums (Video DVD,DVCAM tape, miniDV tape, SVHS tape...).The editing suite space is dominated by FCP and Premiere. One station is not using an NLE.

Playback Systems



Playback systems are spread widely across vendors, with many “older” systems still in use. One station uses a desktop video player for station playout, and only 18% have a workflow where IP transport streams are being switched (Telvue Hypercaster).

Aspect Ratio

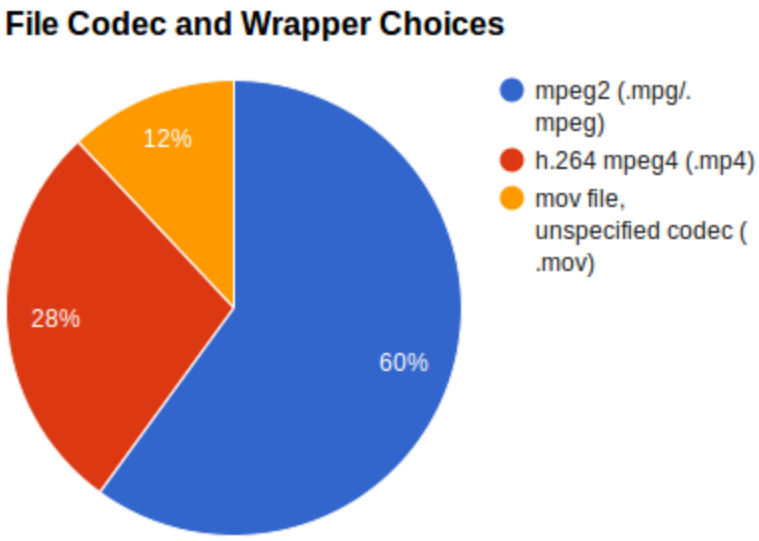


The left represents the currently used aspect ratios for the CMC's digital archiving workflow. Many CMC's have significant additional numbers of SD archives on legacy physical playback media.

Archiving Prevalence

Of the 22 stations surveyed, all have content archives. Twenty one of them (95%) also have digital file archives of some form. There are a number of ways to characterize these legacy and digital archives, but the systems in use vary widely, and do not fall into neat categories. In some cases, stations have multiple systems that have been subsumed within one another or exist in parallel. We will look first at various numbers concerning current digital archives and then mix in legacy physical playback media archives, and their overlap.

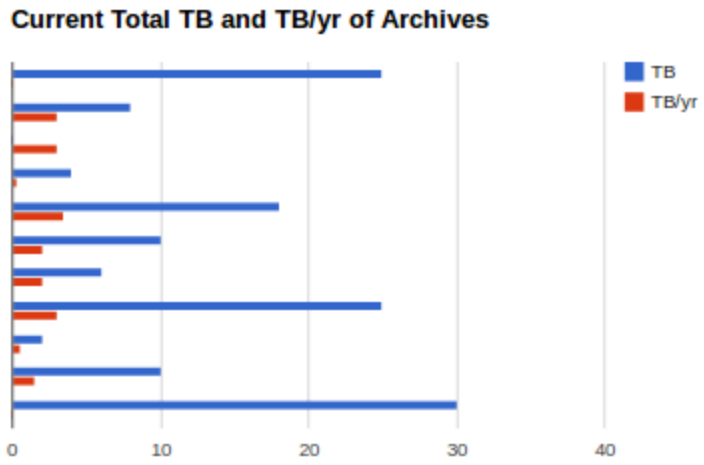
File Codecs/Wrappers



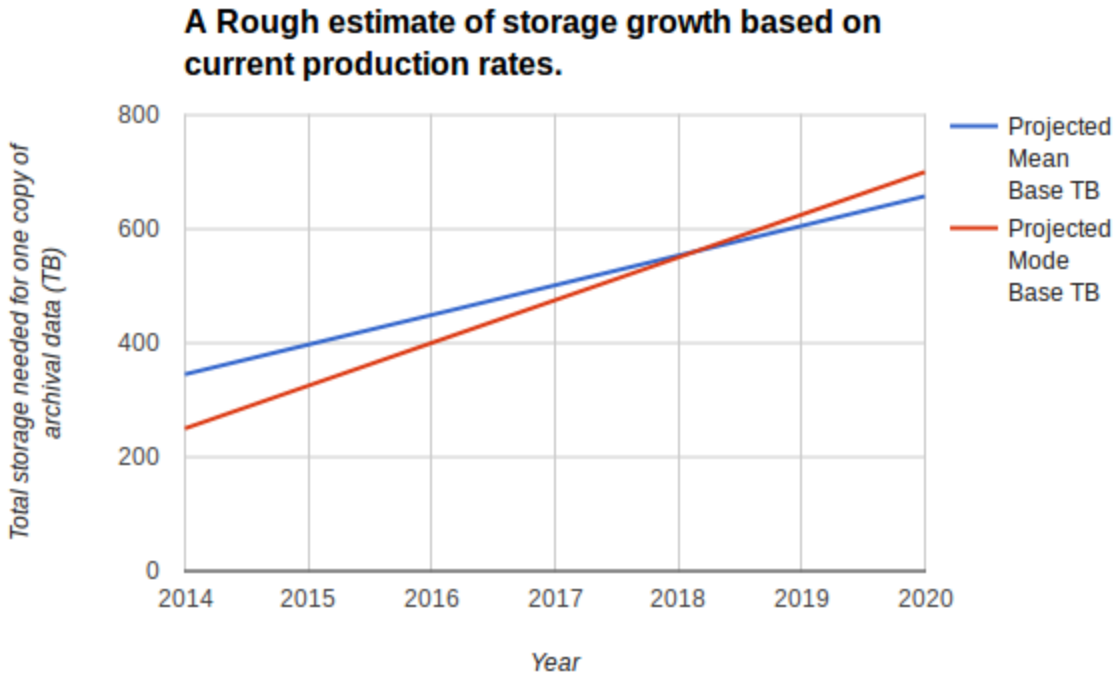
In general, people did not know off-hand the specific codec and the settings used in making the digital archival rendition, but did know the file wrapper. When we look at these numbers, keep in mind that some stations keep digital archival copies using more than one codec/wrapper.

MPEG2 and MPEG4 account for 80% of current digital archival formats, with all MPEG4 encoded by the H.264 codec. If stations were likely to keep archival renditions with more than one codec, the top choices were unsurprisingly MPEG2 and H.264.

Storage Space



As we look at the space that these digital file archives consume, we notice that while all stations have archives, many stations do not know how much data they have or how much of it they have been generating from year to year. Of the 22 stations surveyed, only 10 (45%) knew the total amount of currently archived digital file data, and only 9 (40%) knew how much they were generating per year. Above is a graph of data gleaned from the survey. A lone red bar means it is only known how much data is produced per year, while a lone blue bar means it is only known how much data is stored in total. For stations that know both, both color bars are grouped together. Note that station names have been omitted on purpose.

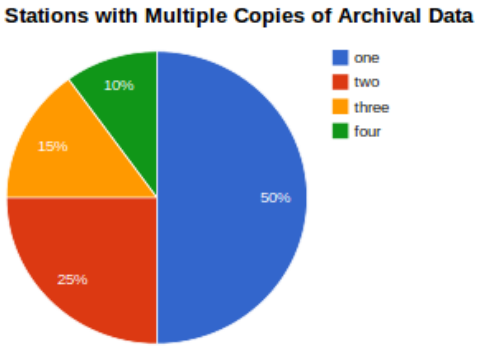


We also made some rough projections for statewide VAN system storage needs.

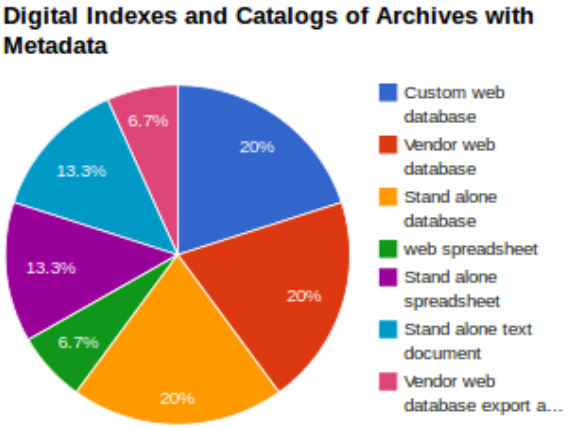
The chart represents a notional model only; changes to encoding parameters, the amount of content currently being produced, and other variable factors may alter projections on this dramatically. One projection uses the statistical mode and the other projection uses the statistical mean from the "Current Total TB" Chart. If anything, these estimates should overshoot an actual number, due to the fact that more of the larger stations participated in the survey, hence a larger representation of the statewide content. These numbers will go up if survey respondents did not include plans to digitize old archival content in their archival storage estimations. Also, these numbers do not include backups of archival data stored in the cloud or off-site.

Backup Copies

Many stations have backups of each program's archival rendition(s), some stored off-site. Half of the stations have only one copy of the archival data and no backups. The remainder do have a backup of archival data, sometimes in multiple renditions for up to 4 distinct copies of a particular program. (See "Stations with Multiple Copies…" Graph).

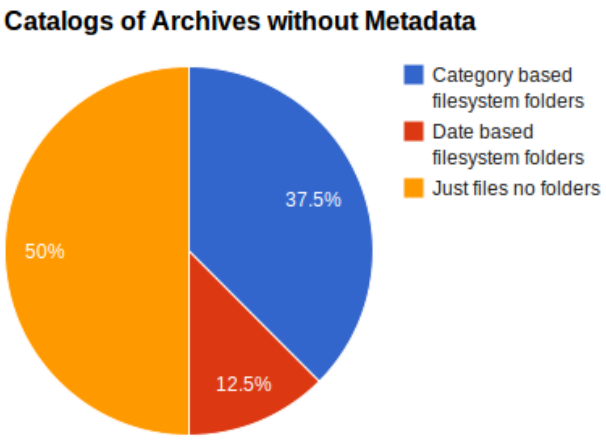


Digital Indexes/Catalogs/Metadata



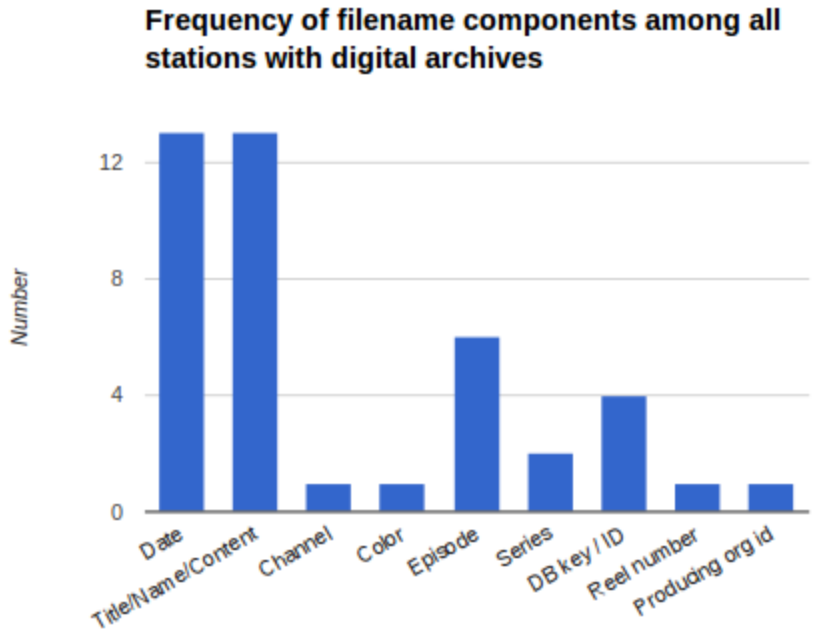
When it comes to linking content and descriptive metadata, many methods are used. Of the 22 stations surveyed, 15 (68%) keep descriptive content and information about the archival file or physical media linked together in a digital system. Here is how those 15 digital metadata systems break down: 60% of stations keep that connection via a database; the other 40% keep it in a digital document or a database exported to a digital document. These stations also normally search a database or digital document to find archival content.

Of the 22 stations surveyed, 8 (36%) have an archive catalog system that involves no digital link between metadata and digital archival content. These 8 stations rely on filesystem or filename conventions to find the specified content. Some stations that use the linked systems above also use filesystem and/or filename conventions like this, but it is not the *only* way they find content. The 8 stations represented above use just filenames and directory structure to find content.



Filenames

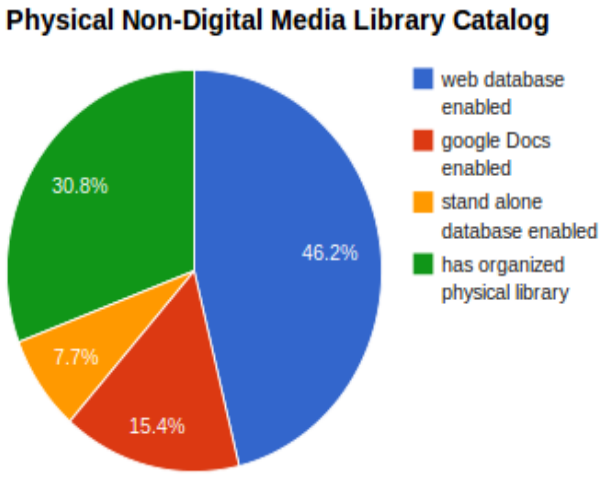
In both cases, media filenames or categories are of key importance in identifying the content. 86.6% of the stations use a system for naming files in a unique manner, the remaining 13.64% have no standardized file naming system. Many different approaches are used to create this unique catalog system. In addition, this system generally has links to older physical media archives and may also have components held over from that system. Below you can see the frequency of components used in the filename.



The favored conventions use date and some form of content descriptor. No two stations use the same format, and all use varying numbers of components in the actual filename.

Storage On & Off Site

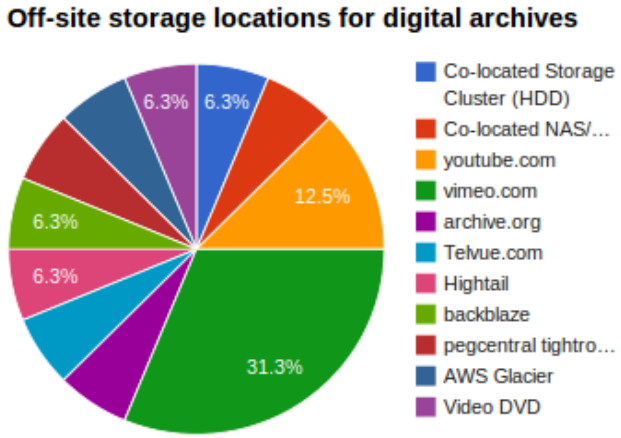
Thirteen stations (59%) also have an organized physical media archive catalog. Some tape, DVD, and hard disk libraries may hold digital content as well. The difference with this type of library is that it requires you to physically retrieve a piece of media (e.g. S-VHS, DVCAM, stand alone HDD, or DVD), from a shelf somewhere.



For the facilities with these physical libraries, there are different ways of finding the data. For these 13 stations, 69.3% of the time it is a database query, while the rest of the time it is using the physical organization present in the library to find the content.

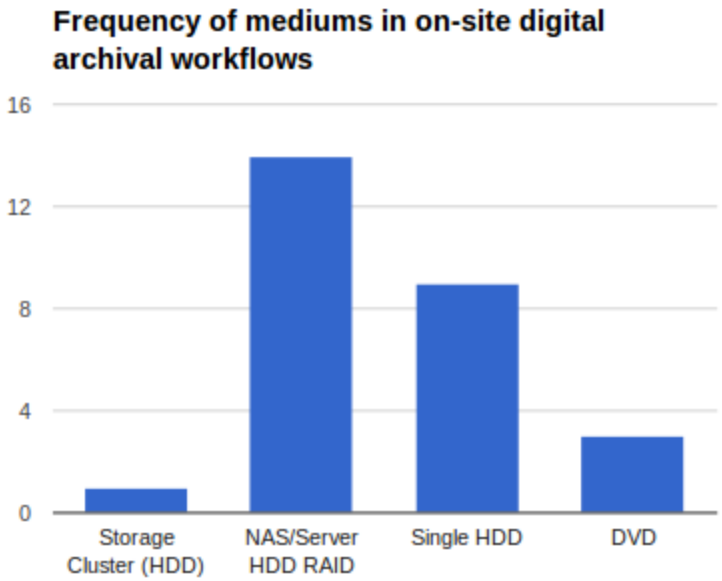
The survey did not go into great length about the precise number and type of physical media in the above libraries, nor did it take into account unorganized libraries of physical media, although these may also exist.

Some stations also store copies of digital data off-site. Here is the breakdown of the 16 (72%) of stations that store digital archives off-site or use a web service. Vimeo and YouTube are

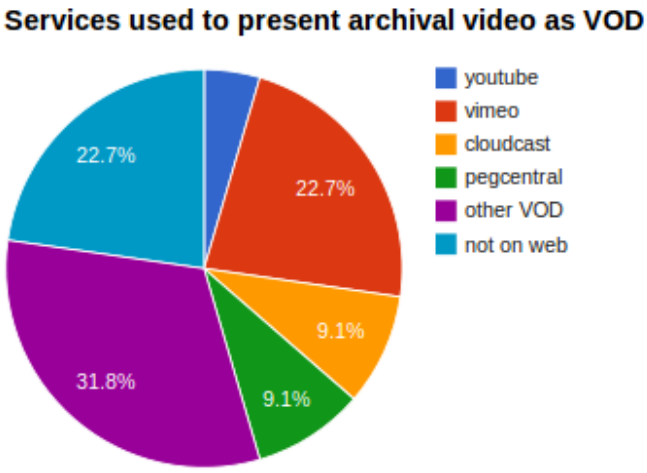


the most popular, used by 43.8% of those who store data off-site.

For on-site digital archival workflows, we see prevalent use of NAS or server-based RAID setups, and single HDD's used in off-site "sneakernet" backups or as libraries storing digital media.



Many digital media streaming services provide an archive resource as well, though it is not VAT's recommendation to rely solely on these services for primary archival use.



III. Appendix: Handy Links

Here are some online resources and links to archival organizations' pages that might be helpful as you embark on improving your station's video archives.

VAT Best Archival Practices: Additional Technical Info http://bit.ly/VATBest2015Tech

Organizations

- The Digital Public Library of America http://dp.la
- Community Media on Internet Archive https://archive.org/details/community_media
- Bay Area Video Coalition preservation http://www.bavc.org/preservation
- The Association of Moving Image Archivists http://www.amianet.org
- Activist Archivists (includes tips for cataloging) http://activist-archivists.org/wp/
- Library of Congress Digital Preservation landing page (with links to National Digital Stewardship Alliance, info and publication The Signal): http://www.digitalpreservation.gov
- WGBH Open Vault Media Library and Archives blog http://blog.openvault.wgbh.org/
- Witness' "Activists' Guide to Archiving Video" http://archiveguide.witness.org/

Cataloguing/Metadata Help

- What is Metadata? A video by Witness non-profit http://library.witness.org/product/video-metadata
- Independent Media Arts Preservation (IMAP) Quick Reference Guide (handy for beginning to catalogue/process metadata, with handy sample records) http://www.imappreserve.org
- PBCore. The site includes details on how to use this metadata system (currently being revamped) for audiovisual content. See "Baby Steps" section to start from zero. http://pbcore.org
- University of Texas metadata scheme (a good example of a complete system): http://www.lib.utexas.edu/schema/Video_Metadata_Guidelines_v1.pdf

Digitization Services

- Northeast Historic Film (Maine) http://www.oldfilm.org/