Practical Cloud Archive

ETC Adaptive Production Archive Group

Denis Leconte, Erik Weaver (Chairs)
Nicholas Mitchell (Co-Author)

Published by

The Entertainment Technology Center at the University of Southern California

Table of Contents

Revisions

Version	Date	Description
Draft 0.0	05/27/2021	Early Draft
Draft 1.0	08/07/2021	DL Added the storage integrity discussion, tape vs cloud. Reorganized Best Practices section, added a bunch. Need to add more there.
Draft 1.1	08/23/2021	NM Added Alain's contribution
Draft 2.0	09/08/2021	More on tape and cloud integrity processes
Draft 3.0	09/27/2021	Tables
Draft 4.0	10/11/2021	Corrections on direction
Draft 5.0	10/17/2021	Added Illustrations
Draft 6.0	10/21/2021	Added Conclusion
Draft 7.0	10/29/2021	Update per feedback from A. Kalas, M. Frend
Draft 8.0	12/8/2021	Update per feedback of working group
Draft 9.0	1/31/2022	Update per ETC review

Introduction

This paper is a report on the activities of the Cloud Archive Working Group, which is sponsored by the Entertainment Technology Center at the University of Southern California. The target audience for this document is organizations and individuals with an interest in the long-term preservation of media assets.

In 2019, the ETC Archive Working Group released their white paper [1] titled "*Guidelines For The Preservation Of Digital Audio-visual Assets In The Cloud"*. The paper focused on fixity: the fundamental process of bit-loss detection in preservation assets, and Cloud archive.

Since the paper mentioned above was published, the group has met regularly to continue the discussions concerning practical Cloud archival. The issue is no less complex today - with the war against the digital dilemma[2,3] waging on.

This paper will provide the reader with an overview of the information collected through the proceedings of the Working Group, based on experiences and insights collected from a broad group of industry experts and practitioners.

The intent is to propose a number of avenues to introduce Cloud storage and Cloud technology as part of an overall archive solution, without compromising the basic tenets of preservation.

[1] "ETC APAWG Guidelines for the Preservation of Digital Audio-Visual Assets In The Cloud" 29 Mar. 2020, https://drive.google.com/file/d/1tQOUPCtQ6UWgJB0A-uUTGqRYWEyUsts1/view.
[2] "The Digital Dilemma | Oscars.org | Academy of Motion Picture Arts ...." https://www.oscars.org/science-technology/sci-tech-projects/digital-dilemma.
[3] "The Digital Dilemma 2 | Oscars.org | Academy of Motion Picture Arts ...." https://www.oscars.org/science-technology/sci-tech-projects/digital-dilemma-2.

For this working session, the Archive Working Group consisted of the following member organizations:

Amazon Studios/Web Services
Google Cloud
Iron Mountain Entertainment Services
Microsoft Azure
NBC Universal
Paramount Pictures
Seagate
Sony Pictures Entertainment
Technicolor
University of Southern California School of Cinematic Arts
Walt Disney Studios
Warner Bros. Entertainment Group

The Working Group has also consulted with several non-member organizations, including:

EPFL
MovieLabs
Walden Pond
Wasabi
WildAcre Company

The document consists of the following 3 sections:

- This introduction
- A report on the discussion of current preservation requirements pertaining to data durability and proof of retrievability, and exploration of potential avenues for interoperability between Cloud-native data integrity mechanism and these requirements.
- A number of potential avenues to implement digital preservation solutions using Cloud technologies, with special regard to meeting fixity verification requirements.

- A summary of a real-life experience in the evolution from a tape-based archive to a hybrid solution using Cloud technology, by the Montreux Jazz Festival.
- Appendices and references

Scope

The audience for this document is any entity familiar with - and potential using - a tape-based archive solution, and wanting to consider the use of Cloud storage and potentially Cloud computing as part of their archival solution. In particular, those considering very practical steps to safely implement that evolution, especially with regards to maintaining archive integrity and asset retrievability.

Out of Scope

While the document highlights the need for more precise experimental data on cloud storage characteristics in terms of durability as measured using the fixity process, such a study would be a fairly long-term process and would need to be undertaken as a follow-up of the current work.

Section 1: Fixity and Cloud Storage Data Integrity Mechanisms

One of the fundamental subjects the Working Group tackled was the very nature of the problem that current fixity methodologies approach, that of minimizing, to the greatest possible extent, the loss of even a single bit of a data archive, including a study into potential equivalence, and a direct path from the cloud data integrity systems to "traditional" fixity reports.

In order to better understand this, it's important to review the data integrity, verification and repair approaches of both traditional on-tape archive and cloud storage viewed from the vantage point of an archival application.

Storage characterization and archive maintenance processes

The fundamental motive behind the development of fixity checks and digital archive integrity maintenance technologies is, quite simply, that all current digital data storage technologies eventually fail. And while the failures may be rare, the quintessential goal of an archive is to guarantee the preservation of artifacts indefinitely. The general approach to establish this requires two components:

- Proper characterization of the storage and its durability. This is essential information as it determines the appropriate remediation approach, especially in terms of periodicity.
- Methods of verification and repair applied at intervals of duration that minimize the risk of any data loss (as determined during storage characterization).

For digital storage, this is implemented through various methods of storage redundancy using storage with properly characterized durability. And, it is maintained by fixity checks (going through every element and verifying, usually through checksum, that it has remained unchanged) restoring and maintaining redundancy when an error is found by discarding the failed element and replacing it with an exact copy from a redundant version of that same element that passes the fixity test.

Fixity checks and reports have become an essential fiduciary requirement of archive management, in particular for owners of extensive collections of intellectual property. Therefore, it is likely that even if fixity evolves, it will remain a requirement in future archives.

One notable aspect of the use of fixity in the current archiving landscape is that while there are common themes in terms of deliverables, there is also a fair amount of variety in the requirements depending on each content owner's internal processes.

Fixity Requirements for 4 major studios	**Studio 1**	**Studio 2**	**Studio 3**	**Studio 4**
Fixity Work Unit				
What is the "object" on that a fixity algorithm/report should be processed against?				
As Granular as possible: Frame level	X	X		
Potential limited grouping (shot level, asset type)	X	X	X	
Per-project (completed collection of files for a show)			X	X
Fixity Report Frequency				
In the scenario where a periodic report is prepared and made available, how often should this take place?				
Periodic Fixity report, quarterly	X			
Periodic Fixity report, bi-annual			X	
Periodic Fixity report, annual		X		
Off-Cycle Verification Ability				
Should there be a feature/possibility to perform fixity verification on specific portions of the archive, in addition to the periodic report?				
Ability to audit specific assets anytime	X		X	
Report Contents and availability				
What information is needed from the fixity reports, and are there any preferred methods to access the report information?				

Report on data repair		X	X	
Fixity Dashboard: Green, bits flipped, bits failed	X		X	
Availability of reporting via API	X	X		X
Reports on Data Location and Fixity repair within infrastructure	X			
Infrastructure Health Reports	X			

The application of fixity checks is therefore dependent on the durability of the storage medium. The current de facto standard verification cycle, usually between 6 and 12 months depending on individual cases and classes of assets, is well suited to LTO verification, as these are time ranges that fall well below the empirically observed mean time to first failure (MTTFF) for LTO tapes under regular use and storage conditions. And while that time depends on the exact circumstances of tape use and storage, it has generally been at least "a few years" under standard conditions, as seen in current audio-visual preservation applications.

But the use of new storage technologies for archival warrants a careful reexamination of the exact modalities of fixity checks and the overall integrity process. Cloud storage has very different implementation characteristics than LTO tape, and therefore, should be expected to have very different durability characteristics. However, these characteristics distinctions are not defined clearly, and there is growing cause for a review of the exact modalities of fixity checks as applied to cloud storage (frequency, general methodology).

There is, unfortunately, scant experimental data on the durability of cloud storage. The CSP provides targets "numbers of 9s" for each of their storage classes, but this is not something archivists consider sufficient to trust without considerably more experimental data, including information on the methodologies used to determine the numbers and whether they prove out in everyday use.

Therefore, this calls for a thorough examination of the data durability of cloud storage - with fixity methodology as a measurement instrument - to make a proper determination of how we can achieve fixity in the context of cloud storage, particularly in terms of frequency.

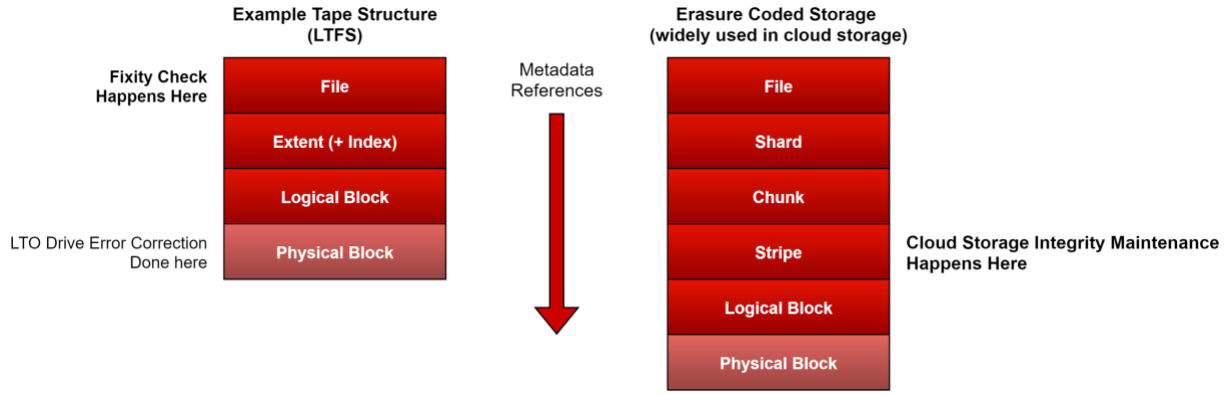
A possible offshoot of this working group, as suggested by one of the participants, would be a longitudinal study of fixity in the context of archival storage in the cloud - by storing a curated "standard archival set" in a number of cloud environments, and performing fixity checks periodically over an extended period of time. This, of course, is likely to be a fairly lengthy project and is unfortunately outside of the scope of the Working Group, but it is an effort that would provide interesting validation data to many of the concepts discussed herein.

Fixity implementation on Cloud storage

Based on the above findings, it is apparent that the fixity process (in particular, the need to have per-file integrity information over time) is necessary for the cloud archiving landscape. However, given the sizes of some cloud archives and the fact that fixity verification does consume resources, the group discussed possible ways to optimize the process at length.

One of the significant findings from the previous Working Group was the potential for some level of interoperability between digest-based fixity and cloud-based data durability solutions. As a result, this Working Group spent a significant amount of time researching the issue and discussing it with the cloud vendors in the group.

The group has determined that, while fixity checks get performed at the filesystem level in the storage stack (be it file or group of files), cloud data integrity checking gets done at a lower abstraction level, and the data integrity algorithm has no visibility into the filesystem structure. **Therefore, there is no compute-less path from cloud data integrity verification data to the type of file-based fixity reports that are current industry standards.** Both techniques have the same goal - ensuring data durability, but they work on different storage concepts, and it is impossible to go seamlessly from one to the other.



This means, in particular, that file-based fixity in the cloud must be implemented using more traditional methods adapted for the cloud. Such as:

- Using cloud computing resources to implement a method to access the archive files (which will be visible as such to a cloud computing resource, since this is the layer at which they operate) and perform the hash calculation and comparison.
- Or potentially using cloud vendor-provided specialized services that perform the same thing (after validation).
- Of particular interest is the notion of serverless hashing, which is a minimal-footprint implementation of a process managed entirely by the cloud provider. Serverless computing can come with some limitations in terms of execution time and process footprint, but these are perfectly adequate for things such as computing a file digest as part of a fixity process.
- And for a bolder call to action, there might indeed be an intriguing possibility to build "archive specific cloud storage" that does offer the ability to resolve any storage integrity error detection into a file-based error indication akin to a fixity fail.

It is necessary to note that current wisdom and experience with the durability of the underlying storage medium, LTO tape, are the basis of current fixity verification implementations. However, one cannot assume that cloud storage will have the same characteristics because it cannot be expected to, and moreover, each cloud storage class (online, nearline, deep, multi-region, etc.) is, in fact, likely to have its own durability characteristics. Therefore, a proper study of each environment's specific durability characteristics will be necessary to determine the exact methodology and frequency required.

Section 2: Cloud Best Practices to Meet Archival Objectives

With any major technology shift, some of the most challenging changes to make are the ones that impact how staff performs their work. In the case of archives) these are the validation methodologies and deliverables required as continuing proof of archive integrity - some of which may be contractual requirements (insurance, for example).

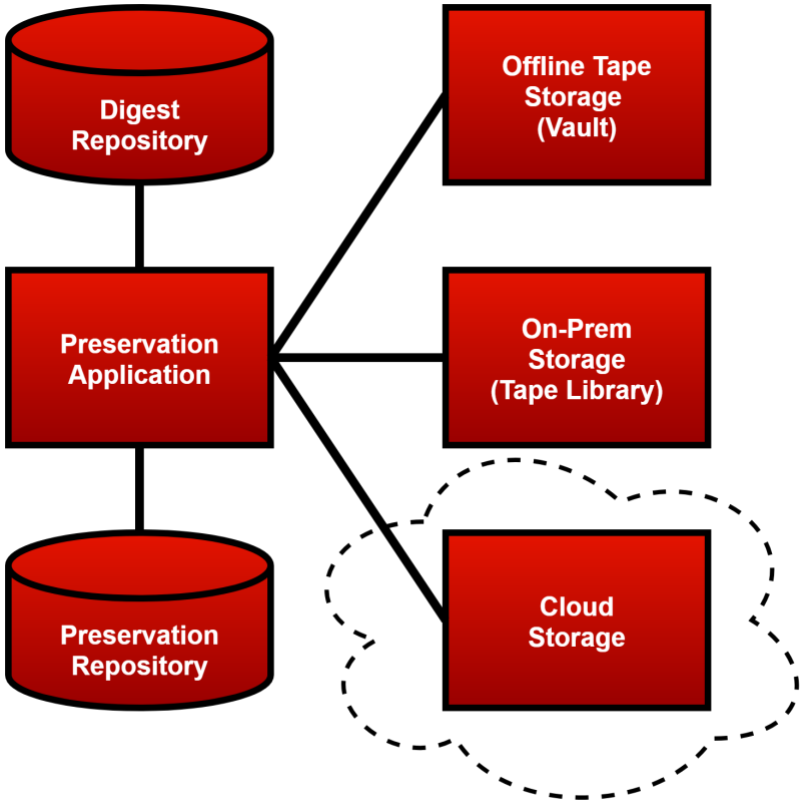
Given the importance of the preservation process, assuming an operation has an established preservation department using traditionally accepted strategies, these will, of course, continue to offer benefits leveraging cloud-based solutions as the curation and identification of what goes into the archive will continue to be more and more critical over time.

In this section, we attempt to offer possible avenues to address the following dilemma:

Given the needs and requirements of audiovisual assets archiving, and what we know - and know that we do not know - about cloud storage (in particular, exact durability characteristics), what are possible avenues to introduce cloud storage in the archival landscape without compromising the essential objectives of an archival solution?

The Cloud as part of the archiving landscape

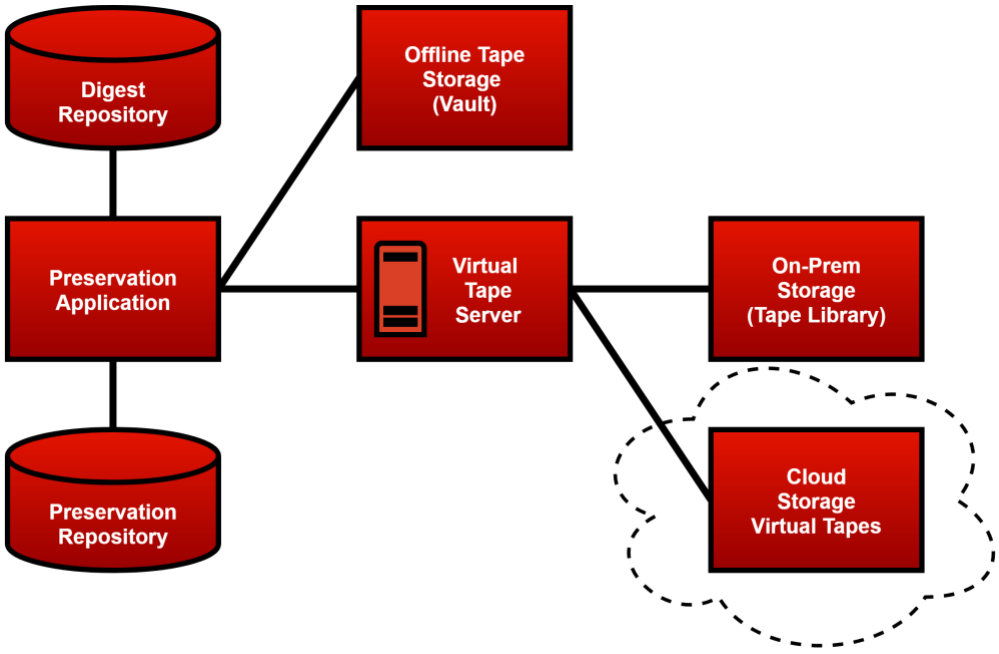
The introduction of cloud storage in an archival solution is not necessarily an all-or-nothing scenario. In fact, especially in the early stages, it could be an excellent approach to use the Cloud as one of the archive elements in a multi-element approach. Instead of making 3 tape copies, have 2 tape copies and 1 Cloud copy.



This can be done with the Cloud copy hidden behind a virtual tape library, but also this is a potential opportunity to move the Cloud portion of the archive to Cloud object storage, and work on addressing the specific challenges that come with that (namely, establishing a clear and stable connection between archive elements as they are known to the archivists, and Cloud objects).

As a general practice, a hybrid classic+Cloud archive solution represents a safe introduction of the Cloud within the archive, which allows for proper experimentation with the technology, without causing any risk to the overall archival strategy.

"Cloud storage as tape" - Virtual Tape Library as part of a hybrid storage approach



Virtual tape library (VTL) technology can be an excellent first step to introducing a Cloud storage layer as part of an existing archival solution based on physical tape libraries in a minimally disruptive manner.

Several solutions are available today that will allow the substitution of physical tapes with Cloud-based (or other storage class-based) "virtual tapes." With these solutions, one can maintain all of the front-end tape workflows and applications used in the organization. All the processes will work the same, and the tape abstraction layer can handle different formats transparently. Therefore, it is possible to have an archive deployment that implements both

physical and virtual tapes, with comparatively modest alterations from an original, "all physical tape" deployment. In particular, all notions of asset repository, multiple copy management, and integrity checking would still apply in the same fashion - the virtual tape copy is simply an additional asset copy.
Many CSPs now have services[4,5,6] that make archive workflow endpoint swap relatively straightforward. And numerous managed services have emerged that can aid in the process of migrating existing archives to Cloud-based alternatives.

One potential pitfall of using a virtual tape library is that it is an emulation layer by its very nature, leading to added complexity in the overall architecture. Furthermore, a VTL deployment is somewhat limited. For instance, a VTL may make using CSP-provided checksum services optimized for object storage more challenging. So, the fixity processes themselves need to be virtualized or containerized to run in the Cloud to perform a checksum validation, or the virtual tape's data needs to be retrieved to where the fixity validation process resides.

Virtual Tape Library as a fixity testbed

A particularly interesting feature of a hybrid, physical+virtual tape library implementation is that it is still possible to run the same fixity checks and fixity reports currently used to validate tape-based archives - on both sides of the archive (Cloud and physical tape). To the archive application, this is all tape.

Therefore, with a choice of a tape work unit on the Cloud side that matches what is used on the physical side, it is possible to establish an exact apples-to-apples comparison testbed for a comparative study of fixity in the Cloud vs fixity on an LTO-based archive, as described above. This would address a particularly crucial need in terms of validating Cloud storage as an archive - providing solid experimental data on Cloud storage, properly tied to the same data entities currently used in fixity verifications.

---

[4] "AWS Storage Gateway project - backup services and costs." https://aws.amazon.com/getting-started/hands-on/replace-tape-with-cloud/services-costs/
[5] "Connect Veeam to the B2 Cloud Using StarWind VTL - Backblaze." 20 Feb. 2018, https://www.backblaze.com/blog/hybrid-cloud-example-veem-vtl-cloud/
[6] "How to replace your tape infrastructure - Azure Backup." 30 Apr. 2017, https://docs.microsoft.com/en-us/azure/backup/backup-azure-backup-cloud-as-tape

The drawback of this method is that it is, by definition, a fairly drawn-out process. Both LTOs and Cloud storage have good durability characteristics, and the expectations are that data faults are going to be infrequent and take a while to emerge. So, this approach would be very much a multiyear process, as it would be an implementation of the longitudinal study described above.

A more Cloud-centric approach to archival

A bit further than simply duplicating a tape-based archive in the Cloud, is to adopt a more Cloud-centric approach to the archive - using object storage instead of a tape emulation layer, but still doing fixity checks.

Again, this process allows for more experimentation with the Cloud, which is strategically important because it will facilitate the integration of archive with any production taking place in the Cloud, as it will happen in the same storage space and layer. At this point, there is no need to organize the archive in tape-size sets of data. Instead, the organization of the archive can be made around logical sets, mirroring the organization of the files in production. And this has the potential to make the archive more intuitive, easier to understand, easier to search through, and ultimately easier to use.

Packaging and asset repository

While this approach has the potential of facilitating a potentially more seamless integration of archival with the rest of the production workflow, it comes with at least two specific challenges of note:

- Asset packaging needs careful consideration, as it is no longer following the "tape-sized work unit" grouping of the previous process.
- This requires the introduction of an asset repository properly connecting Cloud objects and assets as defined by the archival application. This was one of the main topics of the previous ETC Archive Working Group and was described in great detail in their work product.

Multi-cloud approach

Moving beyond the concept of having a single Cloud archival instance of every asset, having multiple copies with different Cloud vendors accomplishes an even greater goal of geographic separation and infrastructure availability, in addition to addressing the risks associated with single-vendor dependency. However, the multi-cloud approach puts increased demand and requirements on the notion of a central repository containing metadata/integrity data and location pointers for all instances of an asset.

Using the Cloud to the Archive's advantage

Using the Cloud as a source for tape archive migration is a good way to get value from it. There are, in fact, many benefits to a Cloud archive that can be explored at this stage.

Fixity checks using Cloud computing

With assets in the Cloud, there is a definite operational benefit to remaining in the Cloud when performing fixity checks. Fortunately, Cloud storage is almost always adjacent to readily available Cloud computing resources, some of which also offer optimized processes to perform checksum calculations - an essential component of fixity checks. It is therefore a natural fit to migrate fixity checks to Cloud computing once sufficient validation of the process has been achieved.

Another interesting aspect of a Cloud archive is that it can, in some cases, be more accessible than a tape archive (albeit with the assumption of a Cloud implementation and access that offers bandwidth comparable to what is available with an on-prem deployment, which while available, is not always present everywhere). This is especially true of older tape formats (not even that old, really, as LTO format and drive obsolescence is quite rapid) where access to drives can be a problem, but it can also simply be a matter of time to access being faster with Cloud storage - although that is not always the case, as reading a tape from a local library can be faster than Cloud retrieval from the deepest of deep storage, although this is, at present, a rather intense competition for which is the fastest access method.

In any case - conditional on a solid level of trust in the integrity of Cloud storage for archive purposes, which can be reached through experimentations described above - it is quite possible to envision using the Cloud, instead of an existing tape set, as the source for archive migration to a new tape set, in scenarios when a tape copy is required.

By the same token, instead of retrieving physical tape archive elements to perform fixity checks prior to migration to new tape media, an organization can use the Cloud archival elements as a source.

Asset identification and curation

The single most common issue with archive and long-term preservation is searchability. It is difficult enough to know what a given asset's details and relevance are at the time of its

creation. So, it should come as no surprise that it only gets more challenging to understand this after any significant period.

It is generally no longer "good enough" to have a naming convention for files. What good is it to know that you have 12 versions of an asset if you don't know which version was approved? Are you sure the latest version in a collection was the version used for final output? Which behind-the-scenes footage did the talent approve for use? Who is the owner of the assets?

Depending on how a preservation store is structured, the details of the decisions made during the lifecycle of a piece of content may even be necessary when going back to a preservation asset repository. As such, depending on your production and its associated strategy and goals for preservation, it is essential to think about the details you will need to provide.

Ultimately, for an archive to be complete, you either need to store assets whose names provide all of the details one might ever need to use them or otherwise store the necessary information required to understand the data contained.

Types of Cloud archive approaches

Cloud - Cloud only archives, if relied on as a sole archive platform, with no other backups or providers, will ideally be stored in multiple data regions to take advantage of the added protection offered by geographic separation. Most Cloud vendors offer a variety of geographically separated storage options with a good range in terms of the geographic spread, from state-wide to planet-wide. One interesting feature of these redundant storage options is that the replication across the various storage regions is a behind-the-scenes part of the product - something that the end user never has to worry about (although there are, depending on the implementation, some ways to validate it). This is a huge convenience factor, as everyone who's implemented geographically separated storage architectures knows that maintaining the various sites in perfect, low-latency sync, can be quite a challenge.

LTO + Cloud - is a practice whereby content is archived on LTO ***and*** the Cloud in a similar fashion to traditional archival, except that if/when the content is to be migrated to a new LTO, it is possible, depending on level of trust in the integrity of the Cloud archive, to use the Cloud

archive as the source which streamlines the process considerably (and generally provides better access/read times than tape-based access). This method allows for a traditional fixity process to take place at the time of the migration.

LTO + Polycloud - is a similar practice to LTO + Cloud allowing for traditional fixity processes, adding one or more Cloud archives to further increase data durability potential, while also preventing Cloud vendor lock-in scenarios. It is also expected that this process would allow for the traditional fixity process to occur at the time of the migration, as described above.

Polycloud - eschews traditional archival methodologies, instead relying 100% on the data mechanisms of the Cloud vendors employed and relies on the fact that the odds of losing integrity the same bits from the same files are VERY low.

One interesting challenge posed by all the multi-entity storage solutions above (LTO + Cloud, LTO + Polycloud, Polycloud) is the need to keep the various entities properly in balance with each other, not only from the data sync standpoint but also to have established proper location equivalences between the various instances of a given asset, as, for example, the path of a given file may not necessarily be the same between two Cloud vendors (depending on their file naming convention limitations), and it certainly will not be the same between Cloud and tape (as it will at least require a tape index in addition to a path).

This brings up the notion of a central archive asset metadata repository, which was discussed in the previous Working Group from the standpoint of fixity but also expanded in this Working Group to add the notion of asset metadata and being a sort of multi-architecture location clearinghouse. Delving into the details of such a repository was unfortunately also beyond the scope of this Working Group, but would definitely be a very valid effort for an upcoming session of the Archive Working Group.

Section 3: A Real-life Cloud Archive Story: Montreux Jazz Festival

One of the challenges of the Working Group has been to find solid examples of current use of Cloud technology for archival, from an entity with the freedom to get into the details to present their experience. Fortunately, as it happens, one of the group members was able to call on a contact he met at the Montreux Jazz Festival during his time working with a storage manufacturer.

Alain Dufaux is Operations and Development Director of the École Polytechnique Fédérale de Lausanne (EPFL) Metamedia Center[7], supporting the Montreux Jazz Festival. And he was kind enough to share some of his experiences with us spanning the last 10 years developing and supporting the festival's preservation efforts.

Taking place for two weeks every summer in Switzerland, on the shores of Lake Geneva, the Montreux Jazz Festival was originally created in 1967, and has since become an event of incredible cultural significance, hosting iconic performances by artists including Nina Simone, Miles Davis, Aretha Franklin, Ella Fitzgerald, Marvin Gaye, Prince, Leonard Cohen, David Bowie, Elton John, and Stevie Wonder. In fact, Deep Purple and Prince even wrote the legend of the festival into their songs, and David Bowie and Freddie Mercury went to live in the region and record albums.

Recognizing the importance of preserving these incredible performances early on, Claude Nobs, the festival's visionary founder, undertook significant efforts to preserve them, recording everything in the highest quality possible at the time of the performances.

Ultimately amassing hundreds of thousands of photos and about 11,000 hours of video recordings and 6,000 hours of audio recordings in a multitude of formats, the Montreux Jazz Festival archive was officially recognized in 2013, inscribed in the UNESCO Memory of the World Register[8].

[7] "Cultural Heritage & Innovation Center - EPFL." https://www.epfl.ch/innovation/domains/cultural-heritage-and-innovation-center/

[8] "The Montreux Jazz Festival: Claude Nob's Legacy | UNESCO." https://en.unesco.org/memoryoftheworld/registry/597

In 2007 the Claude Nobs Foundation[9] initiated the Montreux Jazz Digital Project[10] with the École Polytechnique Fédérale de Lausanne (EPFL[11]), a public research university located in Lausanne, Switzerland, specializing in natural sciences and engineering. The project aimed to digitize and preserve the incredible collection for coming generations, whereby the EPFL would digitize the assets and maintain the preservation efforts in exchange for access to the data. This was quite timely, as the oldest tapes in the collection (as is often the case) were starting to deteriorate.

For the EPFL, this represents an innovation platform, a unique database made available to numerous researchers and laboratories working in acoustics, audio/video signal processing, artificial intelligence, neuroscience, musicology, sociology, and many others.

While carrying on in the spirit of the festival's visionary founder, and collaborating with numerous partners to help develop and promote cutting-edge technologies, the EPFL began leveraging the same techniques as the film industry in their preservation strategies - Multiple LTO tapes (in this case, in 2010 they were using LTO-4, then later on moving to LTO-6), separated geographically.

In 2010, just as the digitization effort was officially underway, one of the EPFL's partners gave them 1PB (Petabyte) of what was then experimental S3 storage. The project quickly integrated the storage in their overall effort, and immediately proved to be a game-changer. In particular it has enabled the EPFL to augment their festival workflows to create rich preservation archive elements within a few minutes of a performance's completion, providing immediate access for the public and EPFL researchers. This is because it is possible to have better continuity of how the storage is structured, between production and archive, as the assets in the Archive storage do not have to be chunked into tape-size segments, but can be moved in more or less the same structure that was used for production - the only reorganization that needs to take place is solely what is dictated by the requirements of best archival practices, without work unit size limitations.

[9] "Claude Nobs Foundation." https://www.claudenobsfoundation.com/
[10] "Montreux Jazz Digital Project – Domains of innovation - EPFL." https://www.epfl.ch/innovation/domains/cultural-heritage-and-innovation-center/montreux-jazz-digital-project-2/
[11] "École polytechnique fédérale de Lausanne - EPFL." https://www.epfl.ch/en/

In fact, the EPFL have come to trust the erasure coding and geographic segregation of the storage instances so implicitly over the years that they use the S3 storage as the data source when migrating to new physical media backups (still LTO-based) when required, which is a major milestone in trusting the S3 storage for archival integrity. Though to this day, they have not needed to go back to an LTO for anything. The entire collection of preservation level assets exists on S3 storage.

The Montreux Jazz Digital Project experience presents a good microcosm of the gradual adoption of what is essentially Cloud storage as a core component of an archive initiative - starting with simply using it for its convenience, but gradually evolving towards greater trust of its integrity for archival purposes, to the point where it actually becomes the "golden copy" used for migrations.

Conclusion and Future Directions

It was very clear during the meetings of the ETC Working Group that the question of data integrity and the fixity gap remain major topics of concerns - largely due to the challenge of reconciling established fixity practices with Cloud integrity mechanisms. This conversation overrode most other concerns and topics, mainly because of its fundamental importance.

After thorough examination of the topic, the conclusion reached by the group is that there is no "low-friction" path to derive fixity data from Cloud integrity processes. While they are both effective data integrity processes, they operate at different layers in the storage stack and there is no computational path to resolve the former from the latter that is simpler than doing the fixity calculation itself.

Therefore, the next generation of Cloud-based archive deployments will likely need to be hybrid setups incorporating both on-prem tape storage and Cloud storage. Beyond simply meeting a primary preservation/archive verification goal, this will also allow, over time, for proper data collection for a well-reasoned comparative evaluation of various implementations of Cloud storage with regards to their data integrity characteristics - as well of course as allow for their evaluation against the current storage technologies. For this purpose, it would appear that a

virtual tape-based implementation will likely offer the most rigorous, apples-to-apples comparison.

While fixity and archive maintenance remained the main topic of this session of the ETC Archive Working Group, there is much left to be explored, both related to this concept but also to the wider implication of practical applicability of Cloud technologies to digital asset preservation:

- A future challenge, and potentially a direction for a future Archive Working Group, stems from the fact that the current full file-retrieval, checksum-verification fixity algorithms are quite resource intensive. With the ever-expanding rates of media generation, there is a very real risk of excessive growth of resource requirements for proper digital archive maintenance - in the Cloud or otherwise. This might require the evolution of other archive verification methods that have higher efficiency. One such direction may be the exploration of Proof of Reliability (POR) methods applied to archive storage.
- Considering a world where the fixity issue is properly characterized and solved within the Cloud, one could envision an evolution of the archive taking better advantage of Cloud technology - proper distribution among Cloud storage tiers based on item importance and use cases, for example. For this purpose, however, there is a lot of room for exploration of proper packaging and organization that would allow for proper structuring of an archive to take advantage of both this new infrastructure, and also the increasing use of Cloud technology in the production workflow.

References

Kalas, A., Levenson, S., ETC Adaptive Production Archive Working Group, & The Entertainment Technology Center at The University of Southern California. (2019). *Guidelines For The Preservation Of Digital Audio-visual Assets In The Cloud*. https://drive.google.com/file/d/1tQOUPCtQ6UWgJB0A-uUTGqRYWEyUsts1/view

The Science and Technology Council of the Academy of Motion Picture Arts and Sciences. (2007). *The Digital Dilemma*. https://www.oscars.org/science-technology/sci-tech-projects/digital-dilemma

The Science and Technology Council of the Academy of Motion Picture Arts and Sciences. (2012). *The Digital Dilemma 2*. https://www.oscars.org/science-technology/sci-tech-projects/digital-dilemma-2

AWS Storage Gateway project - backup services and costs. https://aws.amazon.com/getting-started/hands-on/replace-tape-with-cloud/services-costs/

Connect Veeam to the B2 Cloud Using StarWind VTL - Backblaze. 20 Feb. 2018, https://www.backblaze.com/blog/hybrid-cloud-example-veem-vtl-cloud/

How to replace your tape infrastructure - Azure Backup. 30 Apr. 2017, https://docs.microsoft.com/en-us/azure/backup/backup-azure-backup-cloud-as-tape

EPFL - *Cultural Heritage & Innovation Center.* https://www.epfl.ch/innovation/domains/cultural-heritage-and-innovation-center/

UNESCO - *The Montreux Jazz Festival: Claude Nob's Legacy* https://en.unesco.org/memoryoftheworld/registry/597

EPFL - *Montreux Jazz Digital Project – Domains of innovation* https://www.epfl.ch/innovation/domains/cultural-heritage-and-innovation-center/montreux-jazz-digital-project-2/