DESIGN FOR TOMORROW

LIBRARIES AND ARCHIVES

Managing media inventory

September 2020

Enabled by

Object Matrix

FOREWORD

"Media re-use." "Monetization." "Repurposing." "Operational efficiencies." "The cultural assets of a nation." "Getting the news out on time!"

These are just a few of the critical reasons I've heard over the years of why media libraries are critical to organisations. But, given how critical this area is how can we design our media libraries and archives to not only meet the criteria that we can know of today but also the criteria and challenges of an unknown future?

At Object Matrix these are questions we grapple with daily. We believe in principles such as don't have any single point of failure, don't get tied into closed vendor systems and plan for obsolescence; a good solution should always be able to easily migrate away from itself when its day is done. We believe that the metadata is as important as the data and we believe that it's your data – don't let anyone else take that away from you. And finally we believe it's not a case of will disasters, hack attacks or other challenges happen, but it's a case of planning for when they happen.

The DPP has brought together almost 50 contributors to the vexed question of how to plan for the future and this excellent report contains a wealth of resulting insights and guidance for everyone planning the long-term security and re-usage of their invaluable media libraries and archives.

Jonathan Morgan
CEO, Object Matrix Ltd.

SEPTEMBER 2020

INTRODUCTION

Design for Tomorrow is a project that explores what is required for a media company to be set up for success in a world of constant, rapid and unpredictable change.

The series began by establishing the **Key Design Principles** that every company should adopt if it is to meet its core purposes of protecting itself and its people; responding to challenges and opportunities; and achieving growth.

Those key design principles can be summarised as:

Safe

The business culture keeps employees, customers and the company safe.

Available

Teams have access to the systems, assets and people they need – wherever they are.

Flexible

The business has operational agility: it can respond rapidly to changing conditions and opportunities.

Informed

The business is led by data that enables it to understand itself and others.

Relevant

The best preparation for unexpected events, and dynamic new trends is to have a rich understanding of the customers you serve, and the context in which you operate.

But if those are the general principles that should underpin all companies, what does it mean to put them into practice in particular parts of the media supply chain?

This report continues this assessment by exploring the success factors required in building and managing libraries and archives. These repositories hold media and metadata – the inventory of the media business – whether for reuse and exploitation or for preservation. Their resilience and success is therefore crucial to the health of our industry.

The DPP would like to give special thanks to SDVI, Masstech and Object Matrix, who made this work possible, and have worked with us closely in shaping the output.

EXECUTIVE SUMMARY

- There is growing demand for access to archive content. This was heightened in 2020, but the growth of streaming has seen the value of back catalogues grow over a number of years.
- For many media companies, the archive is no longer an end point where content goes when it's complete; it is integral to the content supply chain.
- Archives therefore need to be more accessible and more flexible than ever before. New technologies and better connectivity have enabled new use-cases for the archive.
- Cultural preservation nonetheless remains as important as ever. These demands for flexibility, speed, and access must be balanced with longevity and stability.

There are six success factors required to design tomorrow's libraries and archives in this way:

1 **Focus on value**

Archives contain *assets*, which are items of value. Whether commercial or cultural, the goal of the archive is to maintain and maximise the assets' value.

2 **Well organised data**

Good metadata is a necessity for a useful archive. Future facing archives use both AI and human metadata generation to better find, describe, and use the content.

3 **Accessibility of content**

A connected archive allows users anywhere to find and retrieve content. Other systems can also access media in order to automate supply chains.

4 **Authorisation and integrity**

Accessibility is balanced by appropriate access control, keeping content safe. Auditability and integrity checking ensure the authenticity of the content.

5 Mitigation of risk

The archive is resilient to a variety of risks, including technical failure, loss of stored media, data corruption, and over-reliance on specific suppliers.

6 Adaptability to change

An archive must be scalable; it should allow for updates to metadata and content over time; and it must adapt to technology developments and obsolescence.

CONTRIBUTORS

The content for this report has been gathered through a series of workshops with almost 50 subject matter experts from across the industry, along with our expert sponsors, SDVI, Masstech, and Object Matrix. The input of these experts has been used to define and refine the Key Success Factors identified in this document.

It must be stressed, however, that while this report has been informed by our discussions with these experts, not everyone necessarily shares all the views presented here.

Noreen Adams
Head of Archive Services,
BBC Platform, BBC

Matthew Blakemore
Head of Product,
BBFC

Ben Blomfield
SVP Metadata Strategy & Operations,
Discovery

Dominic Brouard
Post Production Infrastructure Manager,
Vice Media

David Bryant
Senior Developer,
EditShare

Stewart Carter
Media Manager,
Amazon

Michael Clayton
Sales Account Executive,
BASE Media Cloud

Sandra Coelho
CEO and Executive Director,
Lola Clips and FOCAL International

Steve Daly
Head of Technology,
BBC Archives

Simon Eldridge
Chief Product Officer,
SDVI

Matt Fitzwalter
Archive Manager,
BBC Sport

Nik Forman
Director of Marketing & Partnerships,
Masstech Innovations

Andrew Gavaghan
Head of Archive,
ITV

Tim Gray
Senior Media Manager,
ITV Sport

Peter Guglielmino
Media and Entertainment CTO,
IBM

Tim Guilder
Technology Manager – ITV Daytime,
ITV

Sally Hubbard
Director, Media Management,
PBS

Abigail Hughes
VP Growth EMEA,
Premiere Digital

Jin Imaizumi
Deputy Managing Director,
NHK CosmoMedia Europe

Aditya Jha
Associate Vice President –
Client Solutions, Prime Focus

Chris Kelly
Product Manager,
Ross Video

George Kilpatrick
CEO,
Masstech Innovations

David Klee
VP Strategic Media Solutions,
A+E Networks

Bruno Langlais
Sales and Marketing,
Dalet

Nick Loizou
Library Manager,
IMG

Stephen McConnachie
Head of Data,
BFI

Matt McCue
Broadcast Operations Manager,
Channel 4

Ian Mclaren
Technical Director,
Reuters

Nicolas Moreau
Solutions Marketing Lead,
Sony Professional IMS

Jonathan Morgan
CEO,
Object Matrix

Ian Mottashed
Head of Product Marketing,
Imagen

Ken Murphy
Senior Director, Enterprise Metadata,
Discovery

Keiren O'Brien
CEO,
Filmlocker

Heather Palmer
Senior Project Manager –
Application Solutions, Media Asset
Management, CBC Radio Canada

Nick Pearce
Marketing and Sales Director,
Object Matrix

Huma Qadri
Strategic Account Manager,
Google

Jakob Rosinski
Executive Architect,
IBM

Mitch Ross
Director, Technical & Channel Partnerships,
Masstech Innovations

Glen Sakata
Sales Area Manager US/CAN West,
Dalet

Mike Shaw
Founder,
MediaSaaS

Benjamin Shearer
Archive Engineer,
Discovery

Matt Shearer
Director of Product Innovation,
Data Language

Jacob Smith
Services & Reporting Team Lead,
BBFC

James Snyder
Senior Systems Admin,
Library of Congress

Mathews Thomas
Lead Executive Architect,
IBM

Matt Waldock
VP and Director of Business
Development, EMEA, Xytech

Imogen Wall
Enterprise Account Director,
Google

James Whitebread
CDO,
Masstech Innovations

David Wormstone
Archive Manager,
BBC News

DEFINING LIBRARIES AND ARCHIVES

A library or archive, in its broadest sense, is a collection of content. That collection will of course be supported by a technology stack and a physical infrastructure, but it is the collection of content itself that we consider when we refer to the archive or the library.

Such collections may be stored on film, on tape, on optical or magnetic disk, in the cloud, or on other forms of media. While there are still significant physical collections of media in many organisations, there is a general trend towards digitisation of such content, and of course the vast majority of new content is created wholly digitally. As such, this report primarily focuses on digital content repositories.

Even within this constraint, however, there are still many different uses for collections of content, and many varying ways of managing them. Different industries, organisations, and individuals have developed different terms and definitions to describe their own repositories and processes. It is therefore valuable to define the key terms we will use within this report.

For television and other video content, we can categorise most collections into one of two broad categories. While differing names do exist for these categories, we have chosen common terms that are familiar to a majority of our research participants:

- **Libraries** are collections of content designed for regular access or reuse. These may be finished programmes which can be exploited across new distribution platforms, or news content which can be rapidly repurposed and reused, for example.
- **Archives** are collections of content designed for longer term storage and preservation. They often exist as cultural records, and so safety and longevity may take precedence over speed or broad accessibility.

Many media organisations will have multiple repositories of content, falling into both of these categories. A common model for large broadcasters, for example, is that different production departments – such as news and sport – have their own working libraries, while the organisation maintains a single central archive of finished programme content.

They may interact with external organisations that provide library or archive services too. News agencies and stock footage libraries provide content which is used by many different clients for inclusion in their own programmes, while national preservation archives often store copies of programmes broadcast on public service broadcast channels.

The processes of archive management

Words like *archive* or *repository* may bring to mind images of a static and unchanging collection of artefacts, but any effective collection of media in fact requires a number of active processes to function effectively. The **Open Archival Information System** defines six key services, to which we have added some media industry specific context:

Ingest

The process of adding content to a collection can in fact be quite complex, with a number of sub-processes. Non-digital content might be digitised, while already-digital content might be normalised into a common format (sometimes called a house format). In some cases, content is subjected to quality control (QC) at the point of ingest, to ensure that the media in the collection is of high technical quality and can be readily reused. In other cases, QC is instead performed at the point of reuse.

Data management

The maintenance of the databases describing the content in the archive is key to ensuring that content can be found and reused. In the case of video, this means metadata – that is, information about the content. As will be explored in more detail later, excellent metadata is crucial to a well functioning archive or library. When content is ingested into the repository, it may be necessary to generate additional metadata, to validate the existing metadata, or conform it to a common format.

Archival storage

Once in the collection, the media must be stored and maintained. The technology or physical infrastructure will require maintenance, the content may be allocated or moved between storage devices, error checking may be performed, and so on.

Access

While libraries generally offer rapid and frequent access to media, any repository – no matter how long term – must of course offer access to its contents. Therefore, processes must be in place to browse and search, or otherwise locate content, and to request its retrieval. Once that content is retrieved, it may be viewed, reused, or exploited, as appropriate.

Preservation planning

Especially when managing long term archives, a strategy for preservation will be required, taking account of changing technologies and user needs. Active planning is required to ensure that content remains accessible and understandable over the long term. This includes keeping abreast of technology developments, planning for format migrations, and risk analysis.

Administration

Management of the day to day operations of the archive, development of policies and governance, and coordination of the other activities.

CONTEXT

For as long as it's been possible to record audiovisual content, there's been a desire to keep it, and to reuse it. But the demands for archive content, and the ways in which it is kept, have seen dramatic changes in the last few years.

Lockdown increases demand

The coronavirus pandemic of 2020 led to a dramatic rise in content viewing, as people were subjected to lockdowns and stay-at-home orders. But simultaneously, it created a scarcity of new content, as production ground to a halt, unable to create new programming while under lockdown or while maintaining physical distancing.

As a result, channels and VOD services scoured their archives to find programming that would meet viewers' demand. UK viewers turned to familiar TV as a source of comfort, with viewing of *Only Fools and Horses* up 20% and *Last of the Summer Wine* up 30% since 2019 (**Thinkbox**). Meanwhile, the BBC returned shows like *Spooks*, *Torchwood*, *The Missing*, and *Waking the Dead* (**iNews**, **Edinburgh Live**) to their iPlayer VOD platform to boost the catalogue.

Archive content was also remixed to create new programmes. Disney+, for example, commissioned Arrow Media to produce *Magic of Disney's Animal Kingdom* using archive material, while ESPN's *The Last Dance* relied heavily on archive sports footage (**Parrot Analytics**). Contributors to this research reported similar experiences within their own organisations, including sports federations looking to buy back content from commercial libraries in order to maximise their own exploitation of historic content.

Streaming the back catalogue

Long before the pandemic, however, we were experiencing a surge in demand for archive content. The growth of SVOD services has created a desire for catalogues filled to bursting with content, driven especially by availability of full series.

The growth of SVOD services has created a desire for catalogues filled to bursting with content

With fierce competition between services, a large content catalogue can be an important differentiator, and well known series from the back catalogue can often be as much a selling point as high budget originals. Without the scarcity of supply caused by linear channels' limited hours in the day, there's no reason for content providers not to fill every niche and serve every audience they can, with extensive libraries of programming.

This summer, Nielsen launched its weekly streaming viewing list, reporting the top ten most viewed titles on key streaming services. It clearly shows the value to audiences of library content. For the week commencing August 3rd, for example, just one of Netflix's top ten shows was an original, with the other nine being back catalogue acquisitions (Nielsen via **Variety**). Its catalogue of originals is growing fast thanks to huge investment, making up 51% of shows on the platform that were under one year old as of December 2018, however these originals still constitute just 11% of the overall library, leaving the majority to acquired back catalogue (Ampere Analysis via **Broadband TV News**).

The value of archives

For content owners, the value of their archives has been on something of a rollercoaster over the last two decades.

The value of archives has been on something of a rollercoaster over the last two decades

In 2006, **Variety** described investors' "fascination" with film libraries, but by the early 2010s, declining DVD sales meant that their value was falling dramatically. MGM's net receipts from DVD sales tumbled from $140 million in 2007 to $30 million in 2010 (**Variety**), while Disney sold Miramax in 2010 for $660 million, having once valued the studio's library at anywhere up to $2 billion (**Reuters**, **Deadline**).

However, those values have risen again sharply in recent years thanks to the growth of streaming. Chris Ottinger, MGM's president of worldwide television distribution and acquisitions was quoted in 2019 by **The Hollywood Reporter**, explaining that, "our library has never been older and it has never generated as much revenue as it does today. I think libraries are really really undervalued at the moment – shockingly undervalued, to be honest." The company saw double-digit growth in licensing sales for their back catalogue across the previous twelve months.

The clear conclusion is that holders of rights for library content must have the ability to retrieve and exploit that content if they are to take advantage of the commercial opportunities available to them.

More than programmes

While movies and television series sit at the heart of most content providers' libraries, the types of content that are stored – and the products that can be produced from the archive's raw material – have diversified.

What was once 'secondary' content can now be exploited in new ways. Sports broadcasters and federations are using footage of interviews, friendly games, training sessions and more to create tailored fan experiences.

What was once 'secondary' content can now be exploited in new ways

Entertainment and factual programmes are often supplemented by additional content created for social media or the web, which could be in new editorial formats (such as ultra short form) or technical formats (such as vertical video). Archives may wish to retain not just these additional video assets, but graphic or even text elements that are shared on social media.

Meanwhile, existing archive programmes may be reused and remixed in new ways. One broadcaster explained how they'd recently retrieved a 1970s series from their archive in order to share short clips on social media, due to their pertinence to contemporary events.

This all creates challenges for the archivist. Which content should be kept? Is a tweet valuable in the long term? Or an Instagram Story? If a news story has been repackaged for 20 different outlets, should each version be retained, or just the broadcast master? Many today treat the broadcast output as canonical, but as viewing moves online, such distinctions become less and less clear.

Blurring lines

As explored above, there are two main drivers for the existence of libraries and archives: reuse and cultural preservation.

Some organisations' use-cases fall clearly into one of these two camps. National archives, for example, are long-term guardians of content. Indeed, they may not own rights to exploit the assets they preserve. Stock footage libraries, meanwhile, exist first and foremost to generate revenue from their assets.

For many content owners, however, the lines between these use-cases are blurring. Exploitation of the archive is becoming more important than ever. Where archives were once considered a repository at the end of the process, they are now central to the content supply chain.

Where archives were once considered a repository at the end of the process, they are now central to the content supply chain

Even where older content may not have direct commercial value, there is increasing desire to open up access to the public. The BBC, for example, maintains a physical archive of over 15 million assets, and is undergoing a major programme of digitisation (**BBC**). As the corporation approaches its centenary in 2022, it intends to provide public access to a greater proportion of the 10 million hours of audio and video content it holds (**Radio Times**).

Technology enablers

These blurring lines are increasingly enabled by the development of new technologies. As physical archives are digitised, content becomes easier to access and reuse. As storage and computing power grow, it has become more and more possible to deliver scalable, accessible repositories of content, in which access and reuse is dictated by data and policies, not by physical separation.

It would be impossible to talk about scalable storage systems without considering the cloud. And indeed significant archives of content are now held in public cloud storage, with benefits of remote access, speed, and flexibility.

However, archives tend to be more conservative areas of the business, often with good reason. Reliance on a single external provider may be considered risky when the time horizons under consideration are measured in decades rather than months or years. As such, some organisations prefer to own and control the physical infrastructure of their archive, or to use hybrid options involving private cloud or multiple clouds.

A significant driver for some archives to use the cloud is the availability of artificial intelligence and machine learning to generate metadata. Human curation is increasingly supplemented by – or enabled by – automatic generation of metadata by AI algorithms. With greater volumes of content than ever before, and a stricter focus on efficiency, use of these technologies seems only set to increase.

Key Success Factors

It is clearer than ever before that archive content has great value, and it is also clearer than ever before that unexpected events can disrupt any part of our industry. An approach is needed that protects the valuable assets of the archive against such disruption, while making them available for reuse and exploitation wherever possible.

While each organisation of course has its own nuances, challenges, and advantages, there are common criteria that can be used to measure success. These criteria provide a blueprint to aim for – whether building an archive for the near or long term, and whether primarily for monetisation or preservation.

Our consultation with experts suggest there are six factors that are key to successful libraries and archives:

1 **Focus on value**

2 **Well organised data**

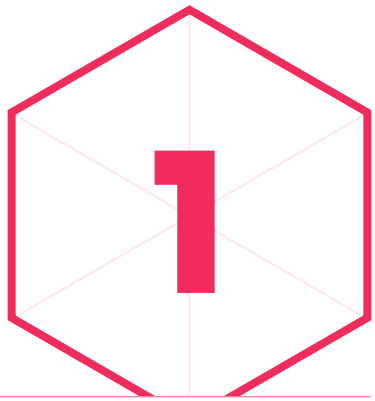
3 **Accessibility of content**

4 **Authorisation and integrity**

5 **Mitigation of risk**

6 **Adaptability to change**

Focus on value



- Archives are collections of assets, which must have value. Whether commercial or cultural, the value to your organisation must be understood
- The commercial value of content is dictated by the rights available to reuse it. While they can be complex, understanding them is an important goal
- In general, broadcasters and platforms should keep a copy of all content they distribute. Raw material such as rushes requires more nuanced policies
- In order to maintain the value of the content, the archive must be well maintained, while operational costs are minimised

An archive is a collection of assets; that is to say, useful or valuable items. So perhaps the most important of our success factors is the maintenance of a clear focus on the value of the contents of the archive. However, value can be difficult to define and predict, so this may also be the most challenging success factor to achieve.

An archive is a collection of assets; that is to say, useful or valuable items

Put simply, it costs money to store, manage, and maintain an archive or library. As with any function of the business, those costs should be aligned to the value that the function delivers. But the measurement of that value can differ depending on organisational priorities.

For a commercial broadcaster, for example, it may be direct monetary value gained by exploiting the content rights. For a national archive, the value is in cultural preservation. The latter is no less important, but can be harder to quantify.

Predicting future value

Even when the value delivered is measurable in dollars and cents, predicting the future value of content can be challenging. The data explored in the *Context* section demonstrates that, much like any investment, the value of your content may go down as well as up.

Much like any investment, the value of your content may go down as well as up

It can be very difficult to predict which content viewers will want to watch in six months, let alone six years, or six decades. The premise of the *Design for Tomorrow* series is that we must structure media businesses to thrive in an uncertain future, and this reminds us that any decision made today about the future value of content is a calculated risk.

The costs of keeping content in the archive must be balanced against the risk of not having that content in future. While there are most certainly individual organisational judgements to be made, some general principles emerge which apply to most circumstances.

With growing demand for back catalogue content and ever reducing storage costs, it is almost always desirable to retain all finished programme material. Multiple contributors to this work spoke of lost commercial or creative opportunities caused by not having content in their archives that they wished they'd kept.

News organisations most often keep the master package for every story produced, as it provides an ability to refer to history and context when reporting future stories. This provides a clear editorial value to their output.

In general, raw material such as production rushes may have less value in the long term. There are exceptions – such as interviews with notable individuals – which may have isolated value, but these are the exception rather than the rule. Nonetheless, retention of all production material for a shorter period may be worthwhile in order to enable uses such as reversioning. Sometimes it is difficult to judge the value of raw material immediately after it has been produced, so one useful approach can be to retain this content for a period of time, such as a few years, and then have in place a review process.

Understanding rights

In order to exploit content now or in the future, you need to own the rights to do so. It is therefore important to understand the rights to a piece of content, and to have these clearly documented.

This can be more difficult to achieve than it may at first seem. In many cases, the production entity is different from the commissioning body, and rights agreements can be complex. Each organisation can retain varying rights to distribution in different territories, over different time windows, or on different platforms.

As one contributor explained, rights are often not black and white. They may be affected by a company lawyer's interpretation of contractual agreements, or the organisation's tolerance for risk. A rights contract written in the 1970s, for example, is unlikely to be explicit about rights for reuse on VOD platforms!

A rights contract written in the 1970s is unlikely to be explicit about rights for reuse on VOD platforms

For raw footage or other non-programme material, the approach here will again depend on content type. Rights management processes tend to be stronger in active and managed libraries such as those in news and sports organisations. In these cases, differences are well understood between self shot content and that acquired through news agencies or from sports federations and teams. Long form production can be more challenging, as production teams are more likely to be made up of freelance individuals who come together for a fixed period to make a programme. In these cases, it is unlikely to be valuable to keep such material in perpetuity.

More reasons to archive

It may on occasion make business sense to retain material to which you do not own the rights for reuse. Broadcasters and platforms may be obliged to keep a copy of everything they have broadcast or made available, even when their rights expire; or they may simply find value in doing so for their own historical records.

Additionally, if there is a possibility of negotiating new rights windows in future, it may be valuable to keep a copy of a programme for this eventuality. It can then be reused when the rights are updated, avoiding the need to retrieve the content from the original producer or the owner of the underlying rights. Indeed, multiple broadcasters discussed during this research that they routinely keep such content as they cannot always trust the underlying rights owner to have the content available. In this case, there is value to the broadcaster retaining a copy of the content.

Reducing cost

Of course, the other way to ensure maximum return on the investment of retaining an archive is to keep the investment as small as possible. As one contributor put it, "keep what you can afford to keep, in a way you can afford to keep it".

Keep what you can afford to keep, in a way you can afford to keep it

A number of our contributor organisations described how they store different content in different forms based on its value. One example is storing the transmission file for all content that is broadcast, and retaining a higher resolution master file only for content which is deemed to have a higher likelihood of reuse or a higher cultural value.

Cost management also means keeping value in mind throughout the interpretation of the rest of these Key Success Factors: increasing efficiency of data management by using AI, balancing the speed of access for content against the cost of the storage tier, and taking a balanced approach to risk mitigation in line with organisational priorities.

Well organised data

- An archive without good metadata has extremely limited value. Curating and protecting metadata is as important as the content itself
- Individual metadata structures for specific types of information are more usable than monolithic schemas, and can be related together in knowledge graphs
- Universal unique identifiers are crucial to understanding content and relating it to metadata and other related information
- Storing minimum metadata directly with assets enhances resilience and portability, but more extensive or dynamic metadata should be kept in dedicated databases
- Human curation of data remains important, but the use of AI generated metadata will grow as its value and reliability increases

Content has no value if it cannot be found. The archive must therefore be well organised, in order that content can be searched and browsed.

Content has no value if it cannot be found

Generating, organising, and maintaining good metadata is fundamental to a well functioning library or archive. Ideally, content would arrive with first class metadata in tow, but in reality this is rarely the case, so excellent metadata processes are required at the point of ingest.

Finished programmes usually have at least a base level of metadata – that which is required for an electronic programme guide (EPG) or streaming interface. This will include title, synopsis, genre, and perhaps cast information. In these days of increasingly rich VOD user interfaces and recommendation systems, it may also have detailed tagging and labelling.

Production material such as rushes is often a different matter. There is rarely good logging metadata, and so it can be difficult to identify material, and even more difficult to reuse it. There is a risk that the response to this is to keep everything 'just in case', but unlike with finished programmes, there are few instances where this has proven useful, and the large volumes of media can be costly to store.

Finally, historic material that already exists in the archives can be a challenge for some organisations. Contributors told stories of content that couldn't be identified at all due to a complete lack of data. In a modern archive or library, this cannot be allowed to happen; any content added to the repository must be correctly identified for its value to be maintained, and significant efforts can be required to update or generate metadata for legacy content.

Metadata structures

Modern database and search technologies have brought new capabilities to manage and search unstructured data, such as documents and arbitrary 'blobs' of data. This means that, in the simplest case, it is feasible to build an archive simply by ingesting whatever data you have in whatever form it arrives.

This stands in stark contrast to a more classical archive methodology, in which a highly curated, somewhat rigid data model is used to classify and describe each artefact. Such approaches often looked towards a single all-encompassing master

metadata schema, which may work well in highly structured and curated long-term archives, but can be unwieldy, and less well suited for libraries that are frequently updated, used, and exploited.

A middle ground provides the best approach for the majority of use-cases. Choose appropriate schemas for different types of content and different classes of data, and link these together where required. In this way, each use-case has a manageable set of data requirements that can be adapted and evolved as required, but data can be connected together to form an overall understanding of an asset or a collection.

Integrating data in this way means linking different data entities together. Connections are made between them in a data structure known as a *Knowledge Graph*.

Whichever metadata form is used – and especially in the knowledge graph approach – it is vital to have strong identifiers.

Universal identifiers

An identifier is simply a name or label that identifies an entity – in this case a media asset. It can be as simple as a filename or a programme name, but a more structured approach is strongly recommended.

A unique identifier allows information to be unambiguously related to the media file

An identifier system allows a piece of content to be individually identified, and connected to related information. For example, detailed metadata may exist in one database, while rights information may exist in another. A unique identifier allows these pieces of information to be unambiguously related to the media file itself, and to each other.

Identifiers should be unique; that is to say that no two pieces of content should use the same identifier. There are a number of ways to achieve this, including an organisation-wide numbering scheme, or a generic system such as a Universally Unique Identifier (**UUID**) or Uniform Resource Identifier (**URI**). However, there are media-specific, managed, global identifier systems such as **EIDR** and **ISAN**,

which – at least for finished content – are strongly recommended. They enable clear identification of a piece of content not just within your organisation, but as it moves through the supply chain between organisations.

There can be a challenge in deciding which level of the content hierarchy you should assign identifiers to. A series will have an identifier, as will each of its episodes. Different versions of the episode – such as different language, compliance versions, or those with access services such as sign language – should usually have their own identifiers. However, different renditions – such as the same video encoded at different bitrates – may not. And it is unlikely that each piece of raw footage for an episode is assigned an external identifier.

This challenge has grown in recent years, as an increasingly international media market means that more versions of an episode may be created than ever before. The important thing is to be consistent, and to ensure that each different editorial entity can be uniquely identified. After all, you wouldn't want the explicit version of your content to be sent to a family friendly VOD platform, or the Spanish dubbed version sent to a French speaking country.

An increasingly international media market means that more versions of an episode may be created than ever before

Referenced and embedded data

There are of course multiple ways in which metadata can be stored. It could be embedded directly within the media file (as in an AS-11 DPP file), stored on the archive storage alongside the media file, or placed in a separate database.

The advantage of storing the metadata directly with the media is that it becomes self-describing. Should the database fail, or become disconnected from the media, the assets can still be identified and described. Advantages of separate databases include better search functionality and easier updates, which can be important as the associated data may change over time while the audio video content itself remains static.

In general, it is useful to store a minimum metadata set directly with or beside the media file as a reference of last resort, ensuring that the archive is self-describing, portable, and resilient. The best way to achieve this is to use open standard formats such as **AXF**. Meanwhile, separate more complex data into dedicated databases. In general the authoritative descriptive metadata would be in one database, while rights information that can be highly dynamic and temporal may be in a separate specialised system.

The most important data to embed in or with the asset is of course the universal identifier which allows it to be related to these external data stores.

AI and automation

It is increasingly common to use machine learning and other automation to create, add to, or refine metadata. Artificial intelligence can now be used to recognise people, objects, text, music, and other entities within the content. It can also be applied to scripts and other related assets to further understand the content.

It is increasingly common to use machine learning and other automation to create, add to, or refine metadata

Currently, the challenge is usually that such systems generate too much information, and it can be hard to process. In addition, because accuracy falls below 100%, they can generate incorrect information. Nonetheless, the time and cost savings are potentially significant, and so the use of these systems will continue to grow. The real advantage of deep learning systems, of course, is that the more they are used, the better they become.

The extent to which AI generated and human generated metadata will be used is of course a function of the business requirements. A long term preservation archive may have very high standards of metadata management for the core data, and may supplement this with additional computer generated information. A library that regularly supplies content to premium VOD services may require human curation of all user-facing metadata to ensure a consistently high quality user experience. But an archive ingesting untagged rushes or historical content with poor metadata may find automated data to be extremely useful indeed.

Accessibility of content

- As with most systems in the modern media supply chain, the archive is most valuable and flexible when it can be accessed at any time from anywhere

- It is still the case that different content may be stored in different storage tiers based on value, but this can be abstracted in a unified *content lake*

- Even where costs or business needs dictate the use of slower storage for master content, metadata and viewing copies should be available instantly online

- A network connected library, accessible via APIs, enables other processes and systems to access content directly, increasing supply chain efficiency

The term *archive* may once have conjured up images of dusty shelves in an inaccessible vault, but a future facing archive could not be more different. Of course many organisations do still have miles of shelving filled with video tapes, film, and data tapes; but almost without exception they are working hard to digitise them.

And while digitisation may offer certain benefits of future proofing the storage technology, the real prize is accessibility. The events of 2020 have shown us more than ever the need for remote access to our content and tools. By enabling users within the organisation to find, view, and retrieve content, you enable its value to be unlocked.

By enabling users within the organisation to find, view, and retrieve content, you enable its value to be unlocked

The need for speed

Excellent accessibility is essential to libraries that are built for reuse and exploitation. It is no coincidence that the commercial content providers among our contributors were the most likely to have moved their libraries to the cloud. Such stores are built to be rapidly accessed, and also to be updated over time.

The extreme case for rapid access to content was demonstrated by one broadcaster that faced a race against the clock during a live current affairs programme. The presenter referenced some comments made in a previous episode, and wanted the clip called up to replay. The production team were unable to retrieve the content fast enough from their LTO archive, before the presenter got frustrated and took matters into his own hands. He found a copy of the video online on his phone, holding it up to the camera to show it. Had the production team had rapid access to the archive at their fingertips, they could have retrieved the required video within minutes.

This would have come at a cost, of course. Retaining a large archive on fast storage – whether on premise or in the cloud – costs money. While the cost of storage is reducing, the old trade-off between access time and cost is as relevant as it ever was.

The old trade-off between access time and cost is as relevant as it ever was

For preservation archives, it is usually sufficient to keep the full resolution content on slower access storage, whether that's LTO tape on premise or archival tier cloud storage. For most libraries, some tiering of content will be required; that which is frequently accessed or deemed high priority should be kept on faster storage, while older or lower priority content is stored on archive storage to save cost. For libraries

where speed is key, it may be worth the cost of keeping everything on fast storage. In such cases, good systems and processes are required for Hierarchical Storage Management (HSM).

There are, however, some areas in which requirements are similar for every archive or library. Metadata should be searchable instantly, and lower bitrate viewing copies (proxies) of the content should be available to view immediately. If high resolution content cannot be accessed right away, an automated process should be in place to retrieve it.

A connected lake of content

It is beginning to become common to bring together all types of archive or library content into a single *content lake*, that allows users to search and find all content from one interface. There may be multiple tiers of storage underpinning this from different vendors or using different technologies, and the lake may contain different content types which are subject to different access and retention policies. But by having a single point of aggregation, all content can be accessed by users and by other systems.

By enabling other systems access to the content repository, intelligent automation and orchestration can be deployed to improve downstream processes. If content can be retrieved, transcoded, and delivered directly from the library or archive, the efficiency of the supply chain can be increased. If the content lake is easily accessible to systems such as linear playout or VOD packaging, then the movement of content between systems can be reduced. For this reason, those that have moved their supply chains or playout to a cloud are more likely to benefit from moving their archive there too.

If content can be retrieved, transcoded, and delivered directly from the library or archive, the efficiency of the supply chain can be increased

Wherever it is stored, a content lake can also be made available to all users across the organisation, provided it is accessible on the network. In doing so, maximum reuse and exploitation of the content is possible. For example, a news library can be integrated with the Newsroom Computer System (NRCS), allowing journalists to search and find content for their news stories within their primary editorial interface.

Of course, this may all be subject to access policies, as we'll explore in the next section. But such restrictions should be a function of business needs, not of technology limitations.

Access restrictions should be a function of business needs, not of technology limitations

Accessibility for the long term

It would be easy to think that accessibility of content is useful only for reuse and exploitation. But in fact, ensuring that content is readily accessible is also important for maintaining the archive in the long term.

Many content companies work with key partners for functions such as media services and fulfilment. In an agile business, these partners may change as needs evolve, or functions may be brought in house or outsourced. In order to be able to change providers, the content must be accessible to the new provider.

If the content is stored on offline LTO tapes, those tapes may have to be shipped to a new provider, that will need an associated infrastructure in order to read and manage them. If the content is in a cloud content lake, or other network accessible store, the new provider can simply be granted access to it.

As will be discussed later, technology migration is also a key feature of long term archiving. And if content cannot be readily found, retrieved, and read, it cannot be easily migrated. Speed may be less important here, but content must be portable, which requires it to be accessible.

Authorisation and integrity

- An archive contains valuable company assets, and so access should be carefully managed using policies and role based access control
- Strong authentication and authorisation systems reduce the need for manual approval processes, balancing frictionless access with security
- Logging of access and changes enables auditing, and helps maintain integrity and authenticity of the content

If content is all available on networked storage with fast access and easy search and browse, then there are of course associated risks. There are very few organisations where every member of staff would be allowed unlimited access to all content, and so access must be managed.

The reasons to restrict access to content are manifold, whether ensuring that unreleased content is only accessed by those involved in its production and distribution, or protecting content to which you don't own the rights from accidental reuse. Or of course, guarding against malicious access by cyber criminals or opportunist content thieves.

There are two key aspects to such protection: authorisation, and auditing.

Authentication and authorisation

In the days of physical archives, archivists and librarians acted not just as curators, but also as gatekeepers. A modern, connected archive replaces this aspect of their role with automated access control.

As with any networked system containing important business data, access to the library or archive should be authenticated, meaning that each user is individually identified before they are granted access. It is no longer sufficient to rely on basic controls like access to a particular network; individual authentication is crucial for logging and auditing.

It is no longer sufficient to rely on basic controls like access to a particular network; individual authentication is crucial for logging and auditing

Once authentication validates a user's identity, authorisation ensures they are granted only the appropriate privileges. This means controlling the type of operations they can perform: search, browse, viewing metadata, viewing proxies, updating metadata, retrieving full resolution content, and so on. It can also involve controlling which assets they can access, perhaps hiding some altogether, or restricting the operations that can be performed on others.

This type of role based automated access control is the preferable solution because it is efficient, and adds little or no additional delay or friction to the process of accessing content, beyond the initial login process. In some circumstances it may also be necessary to implement a manual approval process that allows a user to seek higher authority to access protected content. However, this increases complexity and reduces agility, so should only be implemented where required.

Auditing and integrity

Even with carefully controlled access, there is a small chance that things can of course go wrong. A rogue staff member or a cracked password can unfortunately lead to content getting into the wrong hands. This is one of the many reasons why access control should be supplemented with logging and auditing.

The level of logging required may depend on the business risks involved. For some, it may not be necessary to record each search operation or proxy view. But in a repository storing a billion dollar blockbuster before it is released, even these simple actions will need to be logged. Any archive or library will benefit from tracking each access to the high resolution content.

It is also important to record changes to content and metadata. This will guard against accidental errors, but also more malicious interventions. This is especially pertinent in a world of deep fakes, where the authenticity of content is crucial.

Auditing is especially pertinent in a world of deep fakes, where the authenticity of content is crucial

For news organisations or other high profile content owners, it is important to know with confidence – and to be able to demonstrate – that content has not been tampered with. Integrity checks such as content hashes can help protect against both accidental or malicious changes to content, as well as data corruption. Even content which has not been doctored can cause controversy if taken out of context, especially when dealing with footage of high profile public figures, and so it may be important to trace this back to the context of the original material.

Higher risk cases might warrant encrypting content 'at rest' when it is stored in the archive, and in other instances there are simpler measures available, such as checksumming and access auditing.

Standard cyber security best practices must also be followed. These are not unique to an archive, but when bringing together the company's most valuable assets into a common repository, such best practice becomes more important than ever.

Mitigation of risk

- Risk cannot be eliminated, but it can be dramatically reduced through active mitigations. These come at a cost, so must be aligned to the value of the archive

- Multiple copies of content must be kept; at a minimum in different sites or cloud zones. Offline copies additionally protect against human error and cyber attack

- When using service providers or clouds, the mitigation of the risks of reliance on single vendors should also be considered.

Content in an archive or library must, of course, be secure and resilient against a variety of risks. Nothing can ever be 100% safe, but the correct mitigation measures can reduce risk dramatically, albeit at a cost. Therefore, the level of mitigation put in place will depend on the value of your content, and your business' appetite for risk.

Authorisation and integrity discussed some of the appropriate measures to protect against unauthorised access, and the final success factor, *Adaptability to change*, explores the risks of technology obsolescence. But there are a number of other events which should be considered and planned for.

Failure of technology

Perhaps the most obvious potential failure is that of the technology supporting the archive: the storage medium, the data on it, or the underlying infrastructure.

The most obvious potential failure is that of the technology supporting the archive

The first response to the potential of such failure tends to be redundancy; that is, storing multiple copies of the content. An oft-quoted guideline is the 3-2-1 rule: keep three copies of the content, on at least two media types, with one offsite. This is certainly wise, though in an archive which is regularly updated, thought must be given to how updates are propagated across instances.

We tend to consider catastrophic loss of storage, but it is also worth noting that most storage media can also degrade gradually. As a result, proper maintenance and testing is required for any self-managed storage, even when there is in-built error detection using tools such as checksums.

It is also easy to focus on the content itself while forgetting the metadata. As has been discussed, good metadata is key to finding and exploiting content, so databases must be resilient too.

It is common in many organisations to outsource management of the physical aspects of the archive to a specialist provider, or to put the archive in the cloud. This abstracts you from some of the challenges above, and the cloud in particular gives great flexibility to duplicate content across regions and availability zones seamlessly. However this approach does come with its own associated risk, in greater reliance on key third parties.

Reliance on suppliers and partners

There are significant advantages to operating an archive in the cloud, whether public, private, or hybrid. So how much risk is associated with storing all your content with one vendor?

Of course the tolerance of this risk is individual to each business. Preservation organisations that are concerned with time frames of decades or more tend to see greater risk in such reliance on suppliers, whereas producers, broadcasters, and content platforms may derive greater benefit from storing their content in the same environment they run their supply chain.

Some companies choose to take a multi-cloud approach, splitting their content across different clouds (for example, news workflows in one and long-form in another) or storing content in two clouds at once. The content lake approach described previously may allow these collections to be aggregated into a single interface where needed, though there are likely to be some additional costs associated with moving content between clouds or storing content twice.

A final approach is to use a hybrid solution in which content is stored in the cloud, along with metadata stores and access proxies, but a copy is also kept in an on-premise store. A storage management system is usually employed to provide a single point of access and management. The on-premise copy may be a backup of last resort, containing the media and reference metadata, disconnected from the network (such as an LTO tape placed on a shelf). An offline copy has the significant advantage of being resilient against human error such as accidental deletion of media from the primary archive, or cyber attack on the main archive.

In a hybrid solution, content is stored both in the cloud and in on-premise store

Geopolitical considerations

We operate in an increasingly global content marketplace, but national boundaries still have significance. One important consideration is the territory in which content and data is stored. In order to ensure long term access to, and control of the data, it may be necessary to store it in particular jurisdictions. This can be especially important if the metadata contains personally identifiable information, which could be subject to additional regulation. Some see this as an argument for self-hosting their archive, although cloud providers and many global media services companies now offer contractual options to store data in particular physical locations.

Adaptability to change

- An archive must be scalable, both in the amount of content it can store and in its ability to process incoming and outgoing content
- As storage formats evolve, or business needs change, media will need to be migrated from one format or store to another
- The archive must be adaptable to new formats of media and metadata, accounting for evolution in technology
- Archives are not static; content can not only be viewed and retrieved, but its metadata can be updated and enhanced over time

There are characteristics of flexibility and agility that an archive must share with all aspects of the content supply chain. It must be able to scale up, in order to accommodate growth of content over time. To be truly flexible to business requirements, it ought to be able to scale down too. In addition to scalable storage, the ingest and outgest capacity must be able to scale for peaks in demand.

And as explored previously, the archive should be flexible to allow access from anywhere. But there are also considerations that are particular to libraries and archives. The most important is migration.

Managing technology change

Over time, the representation of the archive will change, at any and all levels; the storage technology will become obsolete, the databases will be replaced, the media encoding will no longer be state of the art, and so on.

As one participant put it, "get comfortable with migration". Another spoke of a culture of continual obsolescence. Unfortunately, without continual maintenance of the content, it may become unusable over time.

Without continual maintenance of the content, it may become unusable over time

In the case of self-managed archive infrastructure, the physical storage media will need to be replaced. This is a familiar process to users of technologies such as LTO data tape. In general, it is more cost effective in the long term to keep up to date with these migrations every three to seven years, rather than have a difficult mega-migration every 20 years, for example.

Of course, you can abstract your organisation from this problem by using the cloud or another managed storage service provider. However, it is still important to keep up to date with new technologies and pricing models; the storage tiers available in the cloud today are not the same as they were ten years ago. And those business relationships may not last as long as an archive needs to; migration from one supplier to another could be required due to changing business context as well as for technical reasons.

The encoding and packaging of the media must also be considered. Today's codec, media wrapper, or metadata format may be obsolete in a few years' time. An archive will need to adapt to new formats, and old content may need to be migrated.

For long term preservation, consideration must be made of whether the content will be readable in many years' time. Some organisations prefer standards based formats and tools for this reason, though the most important factor is whether the decoder specifications are published and stable, in a way that can be used in perpetuity, irrespective of whether it was developed in an open or a proprietary way.

In cases where content is normalised into a house format – or digitised from a physical format – it is often useful to retain the original, as a source for future migrations.

A living archive

Technology and infrastructure are not the only things that will change. The content itself – and especially its metadata – may also evolve over time.

The content itself – and especially its metadata – may also evolve over time

The context in which the archive exists is not static. One content owner explained how they've recently had to revisit a huge amount of content in their archive which is now seen as culturally insensitive around issues such as race. There is no desire to erase history, but it is important to label this content so that audiences understand its context, and that means viewing, identifying, and updating the metadata for these programmes.

As the use of streaming services has grown, so too has the demand not just for content, but also metadata. Whether to drive more feature-rich user interfaces, or better recommendation algorithms, requirements for rich metadata have increased. Content owners often need to update their metadata, so they need immediate access to databases and to viewing copies of the content. An adaptable archive will enable this change, so that content maintains the maximum value possible.

CONCLUSION

There is a great variance in the requirements of libraries and archives; from those wholly focussed on long term retention and preservation, to those with a clear requirement to deliver shorter term commercial return. Perhaps more than in any other area of the *Design for Tomorrow* project, we see a split between these two different use-cases.

Yet there is as much that unifies them as divides them. Any archive wishes to protect its content. Any archive wishes to make its contents as accessible as possible. Any archive needs excellent data about its contents, to make them searchable, understandable, and usable. Any archive must adapt to changing business and technical context.

An effective archive organisation understands the value it provides – whether its assets have commercial value or cultural value – and seeks to maximise that. It sees technology as an enabler not just of storing content, but of delivering value from it.

The major change in recent years is the rise of the connected archive. Interfaces to search and browse the archive, view the content, and (where access rights allow) retrieve it, are now essential. Connections to other systems allow archive content to be integrated into creative workflows, and delivered to new platforms and outlets that enable new commercial opportunities.

With the possible exception of dedicated preservation organisations – such as national archives – no company can now see the archive as somewhere that content is deposited when it's finished with. An archive is the beating heart of a media company; the place where its most valuable assets are stored. It deserves respect and care, and if managed correctly, it will deliver great value back to the organisation.

An archive is the beating heart of a media company; the place where its most valuable assets are stored

THE DESIGN FOR TOMORROW SERIES

These reports can be downloaded by DPP members here as they become available

This report was written by **Rowan de Pomerai** and designed by **Vlad Cohen**.

Workshops for *Libraries and Archives: Managing media inventory*, were led by **Rowan de Pomerai**, and organised by **Abdul Hakim** and **Anh Mao**. Background research was by **Alex Fenton**.

Management of the *Design for Tomorrow* project is by **Abdul Hakim**, with support from **Jayne de Ville** and **Anh Mao**. Content for the project is led by **Mark Harrison** and **Rowan de Pomerai**.

About the DPP

The DPP is the media industry's business network. It is a not-for-profit company with an international membership that spans the whole media supply chain, covering global technology companies, production companies, digital agencies, suppliers, service providers, post production facilities, online platforms, broadcasters, distributors and not-for-profit organisations. The DPP harnesses the collective intelligence of its membership to generate insight, enable change and create market opportunities. For more information, or to enquire about membership visit

thedpp.com

About Object Matrix

Object Matrix is the award winning software company that pioneered object storage and the modernisation of media archives. It exists to enable global collaboration, increase operational efficiencies and empower creativity through deployment of MatrixStore, the on-prem and hybrid cloud storage platform. Their focus on the media industry gives them a deep understanding of the challenges organisations face when protecting, processing and sharing video content. Customers include: BBC, Orange, France Televisions, BT, HBO, TV Globo, MSG-N and NBC Universal. For more information, please visit **www.object-matrix.com**.

dpp®