INTERNEWS CENTER FOR INNOVATION & LEARNING

DIGITAL MEDIA PRESERVATION

WHY MEDIA ORGANIZATIONS CAN'T ARCHIVE*




(*And What They Can Do Instead)

WRITTEN BY
Anthony Bouch
March 2012

Internews

ABOUT THE AUTHOR

ANTHONY BOUCH is a senior systems architect with a focus on information assurance and pragmatic information systems. Anthony has spent most of his 18 years in IT working in the publishing and media industries, and is particularly interested in the preservation and dissemination of digital content. Originally a biology graduate, Anthony was drawn into all things 'systems' at an early stage in his career, and now holds a Master of Science in Information Security from Royal Holloway, University of London.

ACKNOWLEDGEMENTS

This study was commissioned and supported by the Internews Center for Innovation & Learning[1] – a project of Internews[2] that is dedicated to promoting a culture of learning and a spirit of curiosity and exploration across the field of media development.

For many of Internews' media partners and the media community in general, archiving is a challenging problem. There is little independent information or advice to guide organizations that want to create and maintain an archive system for their media content. This study was commissioned in an effort to help address this information gap by presenting a general introduction to the topic of archive management, as well as attempting to help build greater awareness and understanding of the topic as a whole.

The Internews Center for Innovation & Learning also acknowledges the assistance provided by AsiaWorks Television[3] in informing this study, and the contributions made by the nine Asian media outlets that were visited as part of the research process:

- Thai PBS (http://www.thaipbs.or.th/en) - Yothin Sitthibodeekul (Production Technical Manager), Anothai Udomsilp (Director of Academic Institute), Thepchai Yong (Managing Director)
- Madan Puraskar Pustakalaya (MPP - http://wwwmadanpuraskar.org) - Amar Gurung (Executive Director)
- Radio Sagarmatha (http://www.radiosagarmatha.org/) - Rajesh Ghimire (Consultant Journalist)
- Association of Community Radio Broadcasters Nepal (ACORAB - http://www.acorab.org.np/) - Rabi K.C (Technical Coordinator)
- Nepal Forum of Environmental Journalists (NEFEJ - http://www.nefej.org/) - Sushil Mainali (Director Audio-Visual Department), Laxman Upreti (President)
- Press Council Nepal - R.S. Bohara (IT Officer)
- Antenna Foundation Nepal (http://www.afn.org.np) - Rajendra Rijal (Chief Technician), Pramod Tandukar (Executive Director)
- Young Asia Television (YATV http://www.youngasia.tv) - Wasantha Gunawardena (ITManager), Hilmy Ahamed (Chief Executive Officer)
- Malaysiakini TV (http://www.malaysiakini.tv) - Maran Perianen (Program Director), Shufiyan Shukur (Senior Producer)

Special Thanks to Madhu Acharya, Project Director Nepal, Internews Network for coordinating the Nepal visits.

Many thanks to Internews' Asia team who provided advice, support and guidance throughout to the project.

Internews Asia Team:

Oren Murphy- Regional Director, Asia

Kullada Kirtsanachaiwanich- Finanace Manager

Siriporn Tongborrisut- Regional Program Accountant

Siriporn Sungkorn- Office Administrator

Sam de Silva- Innovation Advisor, Asia

DC Team:

Kathleen Reen- Vice President for Asia, New Media and Environment

Shannon York- Business Manager

Dorothy Dai- Program Officer

Internews Center for Innovation & Learning

Mark Frohardt- Executive Director

Eva Constantaras- Program Officer

Amnanda Noonan- Director of Research

Ericha Hager- Digital Media Coordinator

1 http://innovation.internews.org

2 http://www.internews.org

3 http://www.asiaworks.com

CONTENTS

EXECUTIVE SUMMARY

Over the past ten years, media organizations around the world have been migrating from traditional "tape-based" formats (both analogue and digital) to all-digital and completely "file-based" production systems. Such systems are giving content creators powerful production tools, more efficient workflows, and novel methods for distribution and content discovery.

However, "going digital" has also presented a unique set of challenges, at both organizational and operational levels, particularly for the safe storage and preservation of completed work.

This study focuses on digital archive management and will review the current "state of play" in archive management for content creators. It includes both academic and industry research, as well as the results of site visits to nine media organizations located in Thailand, Nepal, Sri Lanka, and Malaysia.

The purpose of this study is twofold: firstly, to present an introduction to digital archive management for media organizations and content creators, helping to build awareness and greater understanding of the topic as a whole; secondly, to provide guidance and practical advice, in particular for small- to medium-sized media organizations, as well as for those operating with limited resources, or under challenging environmental and political circumstances.

The findings of this study can be broadly divided into two areas. The first relates to the safe storage of content and the associated challenges faced by organizations attempting to store large amounts of data. The second is the effective cataloguing and management of digital archives, allowing content within an organization to be described, discovered and re-used.

Safe Storage

In terms of safe storage, it is essential to understand that digital media is dependent upon a chain of intermediary hardware and software components in order to be viewed, or played. These intermediary components are often subject to license restrictions, and, over time, may become obsolete, suffer mechanical failure or simply lose the data contained within them.

What's more, risks associated with technological obsolescence, combined with the limited data life expectancy of current digital formats, mean that no digital storage mechanism can be considered "archival" in the traditional sense. The best any media organization can hope to achieve is the longest possible period between rotations from one media format to another.

For smaller organizations, the problem of safe storage has been exacerbated by the rapid increase in the capacity of affordable hard disks (with terabyte-sized drives now available for less than 100 USD). The concentration of large amounts of data onto a single device substantially increases the risk that mechanical failure, or the loss of a device, will result in a correspondingly large loss of material.

Large capacity hard disks – with their ever-increasing volumes of data – have created a need for more effective backup and safe storage strategies, in particular for organizations without dedicated IT resources. Larger volumes of data have also made it more difficult to create backup strategies that include duplicate sets of data for off-site storage. While duplication and off-site storage of media are important practices for all organizations, they are particularly relevant for organizations with content that has educational, cultural, and social value, and especially for material that may be considered politically sensitive at both individual and organizational levels.

Catalogue and Archive Management

At its most basic, "cataloguing" a digital archive means creating a list that describes items in a collection. More sophisticated catalogue "management" systems allow items in a collection to be accessed from various contextual or referential perspectives, including the selection or filtering of a collection based on descriptive fields, or subject classification terms. A comprehensive electronic catalogue and archive management system also offers opportunities to discover and re-use digital content, as well as the opportunity to share, aggregate, or transfer content between systems and across organizations.

Effective catalogue and archive management strategies for organizations that create media for education, health and development purposes can also support and enhance the aims of such organizations by facilitating the dissemination and re-use of such material.

This study describes the cataloguing and archive management process from both a library science point of view, as well as from the practical point of view of a typical media organization.

From a library and information science perspective, cataloguing and archive management efforts typically require the use of formal methods and published standards designed to support the creation of scholarly and institutional archives. For such efforts, great importance is placed on the creation of verifiable and citable bibliographic records, as well as the provenance, accession, and authenticity of material in a collection.

From a media organization's point of view, cataloguing and archive management is typically focused on providing the minimum amount of information required to support the re-use of archival material in new projects. The priority for most media organizations is the efficient production of content. As such, innovation – including vendor support – has tended to focus on providing solutions for efficient digital production and workflow, with cataloguing and archive management solutions often poorly integrated into the rest of the production process.

However, as media organizations have accumulated more content, the need for more effective cataloguing and archive management strategies is becoming apparent for two reasons. The first is that the longer material is kept in an archive, the greater the likelihood that material will begin to acquire historical value. The second is that media reaching the end of its effective shelf life will require rotation onto new formats if it is to be preserved, and an effective catalogue and archive management system can help to support that process. It should be noted, however, that organizations with large quantities of shelf- and tape-based material will require substantial resources for the capture and transfer of media from tape, and that – even with the support of an archive management system – the cost, time and effort required to capture and catalogue such material means that few organizations have the resources required to do so.

While the needs of an institutional or academic archive may be very different from the needs of a production-focused media organization, this study concludes that media organizations can benefit from adopting some of the principles and open standards associated with library science-based archive management, helping to prevent "vendor lock-in" as well as creating additional opportunities for the discovery, re-distribution, and transfer of material within and across organizations.

Summary of Findings

The move from "tape-based" and other "playable" formats to "all-digital" file-based production systems has created significant challenges for small- and medium-sized media organizations, particularly for those without dedicated IT resources.

Noteworthy in the challenges of "going digital" is the fact that audio and videotape formats had previously provided a convenient format for safe storage (with up to a ten-year or greater shelf-life) as well as an easy to understand and use shelf-based catalogue management system. All-digital file-based systems, on the other hand, have introduced additional dependencies on software, hardware, as well as file management tasks, and have arguably increased the risk of material being lost as the result of the failure or loss of media-containing devices.

None of the nine Asian media organizations visited as part of this study (with the exception of Thai PBS) had effective backup strategies in place for their digital content. Nor were any of the organizations preparing duplicate or off-site data sets for safe storage. As a result, nearly all of these organizations had experienced data loss due to the mechanical failure of hard disks or the inability to read optical media.

It should be noted, however, that the challenges of implementing regular backup and safe storage systems are not unique to the organizations visited as part of this study. Anecdotal evidence suggests that most media organizations are struggling in this area, with many having also suffered significant data losses as a result. Nor are these challenges unique to the problem of archive management. The procedures and systems required to safely store archival material overlap with the procedures required for good data management practices in general.

As such, media organizations are in need of robust systems designed to prevent data loss, as well as assistance in coping with data in quantities that were previously only found in professionally run data centers.

Organizations with requirements to safely store material beyond the production lifecycle are also in need of effective cataloguing and archive management solutions – solutions that will offer the maximum possible period between media rotations, as well as offer the advantages of being able to search, select, and re-use material.

Ideally, such systems should also be based on published and open standards, and therefore able to support strategic initiatives – including the re-purposing, or re-distribution of content, as well as the transfer of content between organizations.

Education and awareness programs will also form an important part of any digital archive management strategy. So too will guidance and practical suggestions that media organizations of all sizes can benefit from when attempting to implement an archive management solution. Two solution-focused appendices have been provided at the end of this report, aiming to serve as signposts for the successful implementation of archive and data management systems for smaller organizations.

Opportunities also exist for media organizations that belong to a network of organizations to co-operate and share experience and knowledge in the area of archive management, perhaps even through the creation of centralized and shared deposit facilities for off-site storage and data safety.

This study also suggests that agencies that fund media development and content creation have an opportunity (and possibly even an obligation) to provide support in the area of digital archive management, in particular for public media organizations, and especially where material of educational and social value is being produced.

This report suggests that many organizations are finding it a struggle to create and maintain effective digital archive management systems. However, if outside entities should offer assistance in developing such systems, it is vital that organizations or individuals with appropriate skills and expertise provide ongoing support. Projects must be run ethically, making certain that material is handled with respect, and that appropriate measures are taken in order to prevent the accidental loss or damage of content.

A digital archive management project should also have clearly defined objectives, and make an important distinction between shorter-, longer- and long-term archive requirements. Any attempt, or claim, to support the long-term preservation of "digital heritage" must ensure that a standards-based approach is followed, and that the systems and infrastructure required for such an effort are available, either from supporting organizations or through partnerships with institutions capable of providing such services.

1

INTRODUCTION

We live in an age where remarkable changes are taking place – the age of computing, of all-things-digital, and of course the Internet. The Internet alone, with an estimated two billion users and rising [1], is providing us with novel methods of communication, and a degree of connectivity that would have been difficult to imagine as little as a decade ago. It has jet-propelled us into the Information Age [2], and its siblings - the Digital and Knowledge [3] Economies.

"Going digital" has also transformed the story-telling process, giving media organizations and content creators powerful new tools for the creation and distribution of content.

It might seem strange then, on the same page, to describe an equally profound and potentially negative consequence of the digital revolution: a phenomenon that is being referred to as "The Digital Dark Age" [4,5,6,7].

The problem stems from the fact that digital content, by its very nature, requires a coordinated chain of intermediary hardware and software components in order to be viewed, or played – and that these intermediary components are often subject to license restrictions and, over time, may become obsolete, or simply lose the information contained within them.

In what may turn out to be one of the greatest ironies of our time, printed matter, film, and other physical, non-digital formats of information may outlast many of their digital counterparts. In fact – even without the lofty goal of long-term digital preservation – the fragile chain of components required to store digital content in the short-term means that material can be, and is being, lost in larger quantities than ever before [5].

Of course, there are also significant advantages to all things digital, such as ease of duplication and distribution, as well as universal access to media from desktop and mobile computing devices. But these benefits come with the associated costs of required infrastructure, and as described above, a new set of dependencies on computing hardware, software and digital media formats.

From a media organization's point of view, the advantages of working with digital media have been easy to justify in terms of production costs and overall efficiency. But it's what happens after production where things become interesting. Whether for short- or longer-term preservation, large quantities of material (now often terabytes in size) have to be kept safely, or "archived" in such a way as to allow content to be preserved, found and re-used. It's here that the longer-term aspects of digital media preservation have conspired to create a unique set of challenges for content creators large and small, around the world.

This study focuses on the preservation of digital media via the successful creation of "working" digital archives, and will attempt to review the current "state of play" in archive management for media organizations in local radio and television production (including their Web-based equivalents).

It includes the findings of site visits to nine media organizations in Thailand, Nepal, Sri Lanka and Malaysia with interests across a range of media activities, including news, documentaries, education and entertainment. The organizations visited represent a cross-section of content creators in Asia – each with their own set of challenges in terms of environment and resources.

The objectives of this study can be summarized as follows:

1.1 Study Objectives

1. Present an introduction to digital archive management for media organizations and content creators, helping to build awareness and greater understanding of the topic as a whole.
2. Identify best practices as well as provide signposts and guidance towards possible solutions in digital archive manage ment that will satisfy the diverse group of media outlets assessed as part of this study.
3. Suggest strategies for organizations that are attempting to preserve material that has social, historical and cultural value.

Chapter 2 will present background information on aspects of digital archive management, including a review of storage and cataloguing components.

It will also look at the move from tape-based audio and video formats, to all-digital, file-based systems – and the challenges it has presented for content creators.

Chapter 3 will introduce an Archive Maturity Model that will describe five states of archive management, providing a common vocabulary that can be used to describe and compare an organization's archive efforts.

Chapter 4 will present a summary of findings for the organizations visited as part of this study.

Chapter 5 will present the study's conclusions.

1.2 Definition of Terms

Throughout the remainder of this report, the term "digital archive" will be used to describe a storage system for completed work or projects that contains digital media files (audio files, video files, and supporting documents) that have either been "digitized" or "born-digital."

The terms "media organization," "content producer" or "content creator" will be used to describe organizations or individuals that are engaged in some form of structured content creation for broadcast or distribution – following general principles of story creation, planning and production, as opposed to casual or non-story-based media creation.

Any use of the term "longer-term" with reference to safe storage, cataloguing, or digital archive management will refer to solutions that are designed to safely store digital content for approximately ten years before requiring rotation.

Any use of the term "long-term" with reference to safe storage, cataloguing, or digital preservation will refer to solutions that are design to hold collections of digital material indefinitely.

2

BACKGROUND

2.1 Safe Storage

As part of an overall introduction to digital archive management, this section will briefly review several common digital media formats, and will compare each of them in terms of their suitability for use as a longer-term storage format for digital archives. Cloud storage facilities and their potential role in archive management will also be examined. The section will end with a brief look at two alternative approaches for long-term content preservation, before moving on to the topic of cataloguing and its role in archive management.

2.1.1 Optical Media

There are now three generations of commonly available optical media and devices: CDs, DVDs and Blu-ray discs. Each format is constructed and operates in a similar fashion. A polycarbonate disc supports an optical dye-based data layer, containing "pits" and "lands" that represent a series of channel bits, which in turn are converted into the binary values of zero or one [8]. "Pits" and "lands" are created with a laser or stamping machine, and are read from the disc when illuminated with a laser in an optical disc drive [9].

CDs, and later, DVDs, became popular as a consumer format for data backup because of the perceived durability and affordability of the media. However, there are no standards for the production of archival or long-term optical formats, and the quality of discs varies greatly between manufacturers.

UNESCO has published an excellent document [10], which summarizes the characteristics of CD and DVD optical media. The document also describes best practices in terms of maintaining an optical archive, including the use of master, working, and safety copies. The author of [10] also emphasizes the need to regularly test optical media. In practice – for smaller media organizations, and in particular those in developing countries where cost is a driving factor – it is typical to find inexpensive and single copy CD and DVD archives that are being stored in sub-optimal conditions without any media testing.

Research suggests that CD-Rs with a gold reflective layer and phthalocyanine-based dyes have the best shelf life [11] although, again, cost as well as user education are factors. Many organizations opt for the least expensive media, unaware that there may be differences in quality, and therefore shelf life, between manufacturers.

Some manufacturers, on the other hand, have made claims suggesting that optical media may have a recorded shelf life of between 50-200 years. According to the authors of [12], only a few studies exist that attempt to prove optical media longevity. Data from these studies suggest that CD-R media has a life expectancy beyond 15 years, while only 47 percent of recordable DVDs indicate an estimated life expectancy beyond 15 years, with some as short as 1.9 years.

Optical media in its present form and capacity is also impractical for the storage of video (as data discs), since transfer speeds along with the number of optical discs required would become onerous even for relatively modest-sized collections.

Figure 1 (below) illustrates the differences in capacity between optical media and hard disks over the past decade.

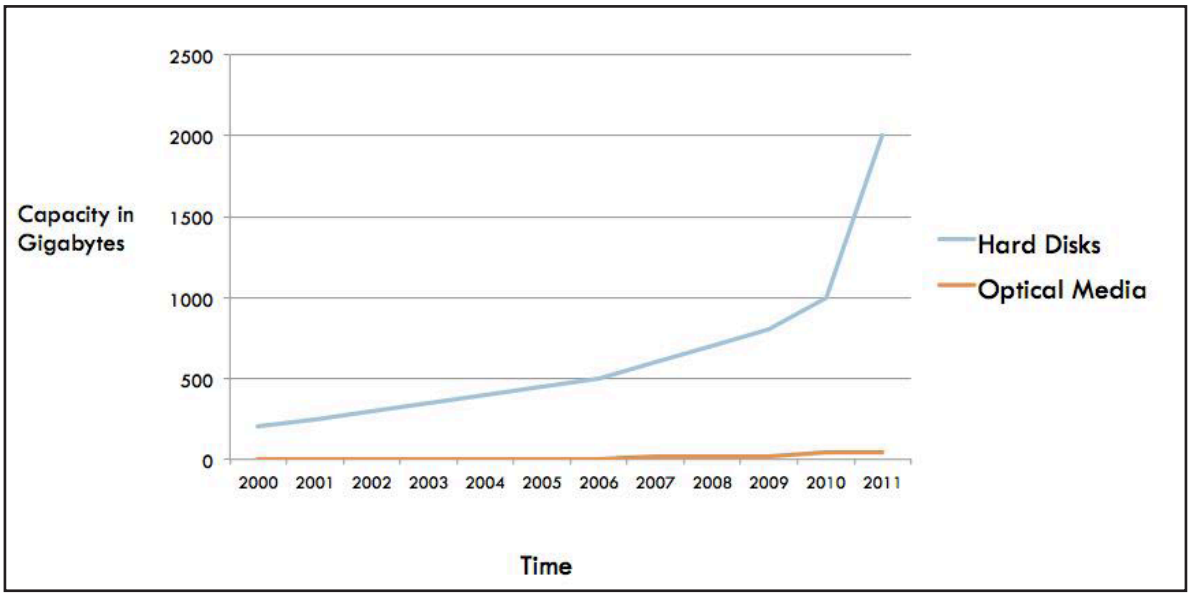

FIGURE 1 - CAPACITY OF HARD DISKS VS. OPTICAL MEDIA OVER TIME.

At the time of writing, dual-layer Blu-ray discs can store 50 gigabytes of data. When compared to a relatively inexpensive hard disk capable of storing up to 1 terabyte of data, 20 double-layer Blue-ray discs or just over 200 single-layer 4.7-gigabyte DVDs would be required to store an equivalent amount of data. What's more – assuming a DVD transfer speed of between 5x and 8x [13] – it would take from six to ten days to read and rotate 200 DVDs, and about 36 hours to read and rotate 20 Blue-ray discs (assuming non-interrupted transfer). The time required to rotate a medium-sized optical media archive, including operator time, will add significantly to the overall cost of maintaining an archive.

Low relative capacity, transfer speed, and significant variation in quality of media combined with a lack of standards suggest that optical media should be used with caution when attempting to store or create a digital archive. While optical media may be appropriate for smaller collections, the author of [10] recommends that more professional and reliable storage approaches be used where possible.

2.1.2 Hard Disks

Hard disk drives are a form of magnetic media. Their construction includes a rapidly spinning platter coated with magnetic material, combined with magnetic heads on an armature that can write and read data on the surface of the platter [14].

Hard disk capacity has risen dramatically over the past ten years with terabyte-sized drives now available at prices that put such drives within reach of consumers for personal use. However, as described by the authors of [15], the risk of data loss increases proportionally when storing such a large amount of data on a single device. The mechanical failure or loss of such a device will result in a correspondingly large loss of material. There are also anecdotal reports suggesting that external hard disks when "unplugged" and used for shelf-based longer-term storage may fade over time [16].

The advantages of hard disk drives include performance, capacity and cost. There is currently no other storage format capable of storing as much data per device, with performance only exceeded at the moment by significantly more expensive solid state drives (SSD).

Transfer time and rotation between external drives can still be an issue. Slower USB 2.0 interfaces at a maximum theoretical speed of 480 megabits per second (which in practice yields an actual data transfer rate of about 20-25 megabytes per second) would require approximately 11 hours to transfer a single-terabyte drive. USB 3.0 and the new Thunderbolt I/O interface allow larger disks to be

rotated in a much shorter period – in an hour or less.

It's worth highlighting that the recent advances in external interfaces and transfer speeds offer an order-of-magnitude improvement in data transfer rates. Improved hard disk transfer rates can help in reducing the overall time taken to rotate an archive, although the time taken for any media-rotation exercise will ultimately be determined by the slowest component or media format in the process.

It's also worth noting that affordable large hard disks are one of the reasons that the need for larger capacity archival formats has become so urgent. Organizations and individuals alike are now able to accumulate data in volumes that would have previously only been found in professionally run data centers.

In an attempt to mitigate the risk of single-device failures, devices that include a redundant array of inexpensive disks (RAID) have become popular and affordable. RAID arrays are typically composed of three or more hard disk drives in a single unit that presents itself to the host operating system as a single drive. Data is "striped" or duplicated across each disk, such that if a single disk fails, the unit can continue to function while the faulty drive is replaced [4]. Again, manufacturers' quality, design and reliability vary greatly.

With an unknown "unplugged" shelf life, and a typical mean time between failure (MTBF) of three to five years depending on use, hard disk drives should only be considered for longer-term storage requirements when combined with robust testing, backup and media rotation strategies.

2.1.3 Data Tape

Although data tape has never become a broadly accepted consumer format, it is the mainstay of professional data management and disaster recovery systems.

The main disadvantage of data tape is its linear format (as opposed to non-linear and random access devices). The tape player must forward or rewind to the section of tape being written or read. Forwarding or rewinding to find a location on tape can take seconds or longer, and so data tape access has a reputation for being slow. High capacity data tape cartridges also suffer from the risk of a single-device failure, or loss, as with large hard disks.

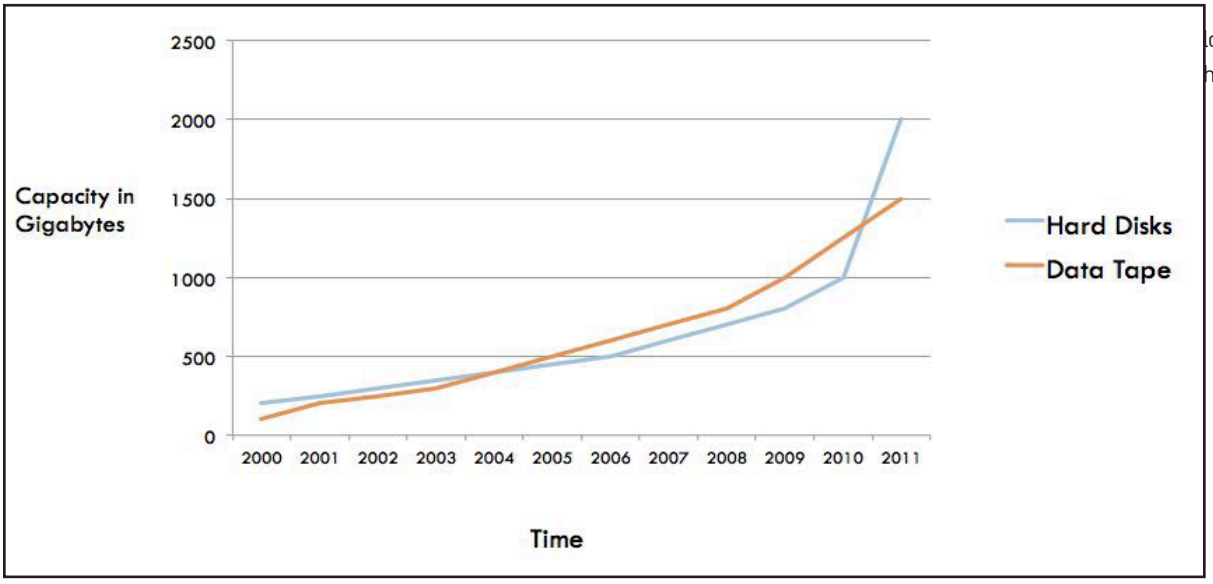

FIGURE 2 - CAPACITY OF HARD DISKS VS. TAPE MEDIA OVER TIME.

4 RAID devices can be configured according to several different "levels." The description above applies to RAID level 5, which is designed to protect against the failure of a single disk.

The latest generation of data tape drives and cartridges includes Linear Tape Open (LTO) devices. LTO was developed by a consortium with the goal of producing an interoperable and non-proprietary tape format [18,19]. There are currently five generations of LTO tape cartridges – with the latest, LTO-5, able to store 1.5 terabytes of data natively (or up to 3 terabytes via hardware data compression). The performance of LTO tape drives is also impressive, with a sustained read-and-write data transfer speed of up to 120 megabytes per second, allowing one terabyte of data to be written, or read, in about two hours.

Write-once-read-many (WORM) LTO cartridges are available from LTO versions 3 to 5. WORM technology may assist with "provenance" and data authenticity requirements in digital archives by providing a non-rewritable, non-erasable, unalterable format.

The estimated data life expectancy of LTO media is quoted as being between 10 and 30 years (assuming the tape is read, or written to, a limited number of times, and is then stored in favorable environmental conditions), although as with other formats, obsolescence and the inability to access earlier-generation formats is likely to pose a greater risk than tape data life expectancy.

There is, however, one significant issue when considering tape as a format for digital archives. Until recently, tape drives required the use of dedicated tape backup software. And additional software adds another dependency to the already complex chain of dependencies in digital systems. What's more, most tape drive software in use today is proprietary, including proprietary methods of formatting and storing data on a tape. While it is conceivable that tape players and tape media will be available a decade from now, the risk that the proprietary software required to read the tape will no longer be available increases the likelihood that data on a tape may not be accessible even while media and players themselves are.

From a licensing perspective, it also seems strange that software, owned and licensed by another organization, may be required to access data that is owned by the content creator or holding organization. The problem is referred to as "vendor lock-in" and it poses a significant risk to the successful creation and maintenance of digital archives.

The risk posed by proprietary software used to control tape drives has not gone unnoticed. In 2010, IBM announced the Linear Tape File System (LTFS). LTFS is a specification that defines both the format and metadata file system of an LTO-5 tape cartridge. It also defines the specification for software that conforms to the LTFS standard [20,21,22]. LTFS allows tape cartridges to be "mounted," and presented to the operating system as if they were regular file systems and folders. Both IBM and HP have released free, single user, single station versions of LTFS software that would allow a single LTO-5 tape drive to be controlled by a computer, with files and folders copied onto an LTFS-formatted tape in a fashion similar to a regular disk drive.

Furthermore, the LTO consortium has open-sourced LTFS software under the GNU Lesser General Public License (LGPL 2.1) [23]. Although the copyright to LTFS remains with the consortium, an open-source codebase, combined with an interoperable and industry standard tape device, bodes well for the future of LTO/LTFS and its use as a component in archive management solutions.

At the time of writing, a standalone LTO-5 tape drive costs about 2,700 USD. LTO-5 cartridges with a native capacity of 1.5 terabytes cost 70 to 80 USD. The entry-level costs of an LTO/LTFS solution should be within reach of most small- to medium-sized organizations.

2.1.4 Solid State Drives

Solid State Drives (SSDs) are an interesting development in storage technology. Whereas conventional hard disk drives (HDDs) are electro-mechanical devices, prone to mechanical failure, SSDs have no mechanical moving parts. They are made from flash memory chips, and offer substantial increases in both data transfer speeds and mean time between failure (MTBF) when compared to traditional hard disk drives. There have been reports that the limited numbers of "write cycles" or the "write endurance" of flash memory may impact the MTBF of SSDs. The consensus at present is that SSDs will typically outlast their host device, and so "write endurance" is not an issue for the regular service duty of a solid-state drive [24]. An "unplugged" solid-state drive should theoretically outlast its "unplugged" HDD equivalent. SSDs are also less susceptible to external magnetic fields that could accelerate the degradation of a traditional HDD.

SSDs, at the time of writing, are considerably more expensive than HDDs, at just over 1 USD per gigabyte. This means that a 600-gigabyte SDD will cost approximately 600 USD – roughly ten times the price of its HDD equivalent. Although prices are predicted to fall slightly in 2012 [25], the current price of SDDs means that SSD technology is unlikely to be a suitable format for longer-term storage of digital media.

SSDs may be appropriate for use during the "ingest" stage of media capture or as working storage in part of an overall archive management system. Their typical use within the media and content creation industries is to provide fast and reliable working storage during "disk intensive" tasks such as non-linear video editing.

2.1.5 Cloud Storage

In classical risk management, there are three ways to manage a potential threat (or in risk management speak, "the risk of a threat being realized"). The first is to simply stop doing what it was that created the risk (risk avoidance). The second is to implement controls designed to mitigate the probability and impact of the risk (risk management). And the third is to give the risk to someone else (risk transfer).

Cloud storage represents an opportunity for the owners and curators of digital archives to transfer the risk of data loss to a third party. When combined with a published catalogue, cloud storage also represents an opportunity to provide broad access to digital archives via the Internet.

Cloud storage (including cloud computing) is defined by a set of services which include on-demand, pooled, broadly accessible, and measured services [26].

In reality, cloud storage means relying on the computing infrastructure and management systems of a specialized service provider. Specialized partners should be able to offer these services at rates that benefit from economies of scale, however, there are also significant issues related to trust, control, data location, and the guaranteed prevention of data loss.

Despite documented concerns, there is significant activity in this area.

The University of Southern California (USC), in partnership with Nirvana Inc., recently announced the establishment of one of the world's largest private storage clouds with 8.5 petabytes of digital archives spread over two sites [27,28]. The facility is being used to preserve and provide access to content housed at the USC Shoah Foundation Institute for Visual History and Education, which includes 52,000 Holocaust remembrance videos – digitized from 235,000 tapes [27].

The Library of Congress has also funded the pilot phases for the development of an open-source initiative called "DuraCloud." From the authors of [29]:

> "DuraCloud, a software platform being developed by the DuraSpace not-for-profit organization, will provide easy entry into the cloud infrastructure by offering data storage, data replication, and services to support data preservation, data transformation, and data access."

The DuraSpace initiatives will be discussed further in section 2.2 Cataloguing.

The implications of cloud storage are significant, and a complete review of economic, legal and infrastructure-related issues associated with cloud storage and computing is outside the scope of this report. From the digital archivist's perspective, cloud storage represents an interesting opportunity to combine or even replace the storage infrastructure required to support digital archives. For both short- and longer-term storage requirements, factors such as contractual agreements, guaranteed service levels and cost must all be examined closely before adopting cloud storage infrastructure as a component in digital archive management.

For long-term storage requirements (for example, where the storage of content is intended to outlive the organization or individual content creators) the continued cost of storage and maintenance combined with external institutional support should also be consid-

ered (as it should for any digital archive or holding designed to outlast its creators).

Described by some as a "first world" solution, organizations in developing countries, or where Internet access is limited or unavailable, will unlikely be able to consider cloud storage as a practical option – at least not without the assistance of an outside organization.

There is also the interesting philosophical question about whether cloud storage (in a constant state of media rotation) is the solution to the need for large capacity, long-term storage of digital objects, and whether this will supplant the pursuit and manufacture of new and larger digital media formats.

2.1.6 Other Formats and Initiatives

Two other archival media projects are worth briefly mentioning in this section.

The first is the Millenniata Project. Millenniata is a start-up venture that has created a new optical disc format that claims to offer a "permanent" digital recording [30,31]. The format can be read by any standard DVD player, but requires a special M-DISC writer in order to record to the media. Unlike regular DVDs, there is no reflective or die layer in M-Discs. Instead, during the recording process, a laser etches "pits" onto a substrate material that is said to be "rock-like." Millenniata claims that M-Discs will be readable for "hundreds of years." Millenniata has partnered with Hitachi-LG to produce M-DISC drives, and expects to have units available for purchase in early 2012. The M-DISC is an interesting development in digital storage technology, although it remains to be seen whether the product will succeed commercially. As with regular optical media, an initial storage capacity of 4.7 gigabytes and relatively slow transfer speeds mean that M-DISC in its current form would only be appropriate for small digital archives. More information can be found on the Millenniata web site [5].

The second project worth mentioning is the Rosetta Project [32]. Although not directly applicable to the problem of digital preservation for content creators and media organizations, the Rosetta Project is an interesting example of a project designed to create a "very long-term archive" as well as address the problem of digital obsolescence.

The archival method chosen by the Rosetta Project is to microscopically engrave analog images onto a 2.8" nickel disc. And for its first project, it is creating a disc that will serve as a kind of "decoder ring" for over 1,000 human languages – hence the name of the project, with its parallel to the real "Rosetta Stone" [33]. The disc will contain 13,000 micro-etched pages of language documentation.

The following quotation is from the "Technology" page of the project:

> "For the extreme longevity version of the Rosetta database, we have selected a new high density analogue storage device as an alternative to the quick obsolescence and fast material decay rate of typical digital storage systems."

After eight years of development, five of the Rosetta Project discs were produced in 2008, with three of them distributed to their new owners. Images and a description of the process required to create the discs, including an interactive presentation, can be found on the Rosetta Project site [6] and blog [32,34].

2.1.7 Safe Storage Summary

In summarizing this section on safe storage, the following quotation from the author of [11] is particularly relevant:

5 http://millenniata.com/

6 http://rosettaproject.org/

> "No computer storage medium can be considered archival, irrespective of its physical longevity: technological obsolescence is inevitable and all media have limited life spans. For the foreseeable future, the need to periodically refresh electronic records onto new media is therefore inescapable. Careful selection of storage media can maximise the periods between refreshment cycles and simplify the refreshment process, in addition to ensuring that data is as secure as possible."

The author of [11] also recommends evaluating storage media based on selection criteria that include:

- Longevity – the media should have a data lifespan of at least 10 years.
- Capacity – the media should have a storage capacity appropriate for the quantity of data to be stored.
- Viability – the media should support robust error-detection and verification.
- Obsolescence – the media technology should be well established, and ideally based on open standards.
- Cost – cost should be evaluated based on media, in cost per gigabyte, as well as on the cost of purchasing and main taining the necessary equipment.
- Susceptibility – the media should be resistant to damage, including tolerant of as wide a range of environmental conditions as possible, and any measures required to counter environmental risks should be affordable and achievable.

We can see that a number of factors must be taken into consideration when choosing storage media for digital archives. For small collections (up to 250 gigabytes of data), high-quality DVDs, or Blu-ray discs (and perhaps eventually M-Discs) may be appropriate, although every effort must be made to ensure that the device and media used are of the highest quality. For medium to large collections (250 gigabytes and above), LTO and LTFS as open standards represent a good choice, and would score well against the selection criteria above with a high probability of reaching a target rotation period of ten years. Hard disk drives, with attractive cost, performance and capacity features will form an important part of any archive management solution. However, with a relatively short MTBF and an unknown "unplugged" shelf life, hard disk drives should only be considered for longer-term storage when combined with robust backup, testing and media rotation strategies. Lastly, cloud storage facilities will play an increasingly significant role as a storage format for digital archives – although important questions relating to access, trust, longevity, and the guaranteed prevention of data loss need to be answered when considering the cloud as a component in an overall archive management solution.

2.2 Cataloguing

It's one thing to safely store your digital content. It's another thing to be able to find it.

This section will briefly examine the topic of catalogue creation, with a view to creating systems that will allow media organizations and content creators to both organize (arrange), and find content that has been prepared for inclusion in a digital archive.

The topic of cataloguing is a large one, especially from a library science point of view, although at its simplest, creating a catalogue means creating a list that describes items in a collection.

2.2.1 From a Library and Information Science Point of View

It's worth describing a few aspects of catalogue creation from a library and information science point of view, at least to provide some of the basics.

To start with, there's an important distinction between "describing" items that are in your archive, and "classifying" items in your archive.

"Describing" items in a catalogue refers to the process of recording data about the object (data about data is referred to as "metadata"). Data in this case would include things like the date the object was created, the title of the object, the type of media or format the object exists in, the object size (or extent), copyright information, etc. The process of recording descriptive information about objects in an archive is referred to as "descriptive cataloguing."

"Classifying" an object, on the other hand, requires a "subject" understanding of the object – for example, whether the object belongs to the category of education, science or politics.

Take for example the modern digital camera. When an image has been captured, nearly all modern digital cameras will record information in the Exchangeable image file format for digital still cameras: Exif [35]. Exif data embedded in a picture can tell you the date of the picture, aperture and shutter speed, the size of the image, etc. – all examples of "descriptive" metadata. What the camera cannot do is tell you what the picture is about. That's a classification problem.

Classification can be performed using controlled and uncontrolled vocabularies. Controlled vocabularies constrain the choice of classification terms to a defined taxonomy – typically, a hierarchy of terms. The advantages of a controlled vocabulary are consistency in terms, and in the case of "thesaurus-like" vocabularies, the ability to subdivide terms into broad and narrow meanings. This, in turn, allows a catalogue to be searched or divided into broad or narrow lists, depending on the selected term. Published vocabularies in the form of taxonomies and thesauri are available from the Library of Congress Authorities [36], UNESCO [37] and others, including specialized collections of terms for specific subject matter.

Uncontrolled vocabularies, like a "tagging" scheme, are not constrained to a list of terms in a defined taxonomy. They are typically user-defined terms that can be assigned to a catalogue entry on an entry-by-entry basis. Tagging or social classification schemes have become popular because of the Web, and have allowed user-generated content to be classified in whatever way the community of "taggers" sees fit.

Both controlled and uncontrolled classification schemes have their own strengths and weaknesses – although the goal of each is to allow a catalogue to be viewed in different ways, and to assist in the discovery of catalogue items.

"Indexing" is a term that is sometimes used to refer to aspects of catalogue creation – both descriptive cataloguing as well as classification.

Standards exist for the process of cataloguing itself, which is particularly important in scholarly collections where a complete bibliographic record is important for both finding material, as well as correctly citing it.

There are also standard schemes available which define a set of data fields, or metadata, to be used in descriptive cataloguing. These include:

Machine-Readable Cataloging (MARC). MARC is one of the oldest and most widely used metadata standards for cataloguing information in libraries of books and other material. MARC was developed at a time when computing power was relatively low, and system memory was precious, and so uses a three-digit numeric code (from 001-999) to identify each field in a record. For example, "245" for title, "100" for name, or '"300" for physical description, etc. [38].

Metadata Object Description Schema (MODS). MODS is a modern metadata format that was designed to be simpler and more user-friendly than MARC, yet provide a richer metadata standard than simpler schemes, such as Dublin Core (see below). MODS uses descriptive field labels, like "titleInfo," "name," "genre," or "typeOfResource," etc. [39].

Dublin Core. The Dublin Core Metadata Element Set is part of the Dublin Core Metadata Initiative (DCMI) and contains fifteen field elements or terms (for simple Dublin Core) used in resource descriptions, including date, title, creator, rights, etc. [40].

Encoded Archival Description (EAD). EAD is a metadata standard for the descriptive cataloguing of archives and special collections.

As well as regular metadata fields (that can be mapped to MARC and other standards), the standard contains a descriptive section that can include the full inventory of collection items, each with its own complete descriptive record, which can include title, date, name, size, container type, etc., including any restrictions that may apply to the collection as a whole, or to individual items. EAD would be a suitable metadata format for a collection of scholarly manuscripts or even for a collection of oral history recordings [41].

Visual Resources Association (VRA). VRA Core is a metadata standard designed to describe works of visual culture, as well as the images that document them. VRA Core distinguishes between a "work," and an "image of a work" and a "collection" and allows relationships to be maintained between them [42].

PBCore. PBCore is a metadata standard for audiovisual media, and was developed by the US public broadcasting community (funded by the Corporation for Public Broadcasting). PBCore is based on Dublin Core, with additional media specific elements [43]. An example of a large PBCore-based and publicly accessible archive is the WGBH Open Vault project [7].

Metadata standards are important not only for discovery purposes within an organization, but also as a way to share, aggregate, or transfer collections between systems and across organizations. The PBCore website has a nice summary of the advantages to using machine-readable metadata standards for describing items in a catalogue:

> "Your investment of time will pay off as you develop your PBCore–based media catalogue or metadata system. Your media assets will be findable, reusable, and shareable to whatever extent you wish to make them. Your collection will be able to interoperate with a growing number of other media software systems and archives, including the American Archive. You will be able to publish media online with detailed metadata, including linked data, allowing users to discover assets that previously were hard to find or inaccessible. You'll have a digital media catalogue you can repurpose for any intended use, from internal media asset management, discovery, and reuse, to stock footage sales and eCommerce, K-12 curriculum content, public website access, offline archives, and media preservation repositories."

Clearly there are advantages to creating a standards-based catalogue. However, the resources required for creating and maintaining such catalogues can be significant. Training and expertise is required in both descriptive and classification-based cataloguing tasks, and certainly not every media organization requires "scholarly" level catalogues in order to manage its digital archives. That said, where possible – and ideally as close to the content creation stage as possible – a standards-based approach to catalogue creation can provide unique opportunities for the discovery, distribution and sharing of resources across a wide range of distribution channels and audiences.

2.2.2 From a Content Creator's Point of View (DAM, MAM)

Cataloguing in the media industry – for content creators, production companies, and broadcasters – is typically described in terms of "digital asset management." And so the solutions used to support workflow as well as archive management are referred to as digital asset management (DAM) or media asset management (MAM) systems.

The distinction between workflow, and archive management in this context is particularly important.

Workflow or production-focused DAM solutions are designed to support the process within which content is created. Systems in this category range from small and fairly simple to large, fully automated production control systems – such as those designed to receive and manage digital assets from "ingest" to post-production, and on to "play-out" via integrated live "on-air" broadcast and programming systems. Within the production (or post-production) environment, content is being manipulated and changed frequently. Completed work, finished stories, or edited packages – having been played, delivered or distributed – are then ready to be archived.

For many media organizations, the focus is on production and production-related costs. What matters is getting the work out. As a result, vendors have also tended to focus on creating solutions that support efficient production workflows. Archive management solutions are often "bolted" on to the back of production systems, resulting in a poorly integrated solution dependent on proprietary

7 http://openvault.wgbh.org/

software and formats (for example, the now defunct Apple Final Cut Server product, integrated with backup and archive management solutions from Archiware PressStore).

There is another reason why archive management has played second cousin to its production-focused equivalent. Until fairly recently, audio and video tape formats, including high-quality digital video tape formats like DigiBeta, DVCAM, Mini-DV, etc., were still (and in some cases, are still) being used to record and store completed projects. It turns out that having a "playable" format (tape in this case), with a pretty good shelf life (from 10-15 years), offered content creators a convenient way to catalogue and access their archives. There are some obvious disadvantages to a tape-based archive, but as an archival "unit of currency," tape was something that nearly every content creator understood, and could manage. Cataloguing tape (both descriptive cataloguing and classification) could be as simple as writing a label on a tape box cover, or creating an entry in a printed journal or spreadsheet.

With a convenient and portable archive format at hand, production companies and vendors understandably focused on the intermediary steps in the production process, where significant cost savings and efficiency could be found in a file-based, all-digital workflow, before returning to tape for final storage and archive.

The major downside of an audio or video tape-based catalogue, of course, is what happens when it comes time to rotate expiring media. Large tape libraries are extremely expensive to capture and transfer into digital file formats. Catalogue information on album or box covers needs to be captured and recorded, and the actual transfer of content requires dedicated operators, playback machines, and the "real time" playing of every tape in the collection. Thousands of tapes will require thousands of hours to rotate.

2.2.3 Open Source Cataloguing and Archive Management Solutions

In both the production-focused world, as well as in the academic and library science world, there are a number of open-source initiatives for DAM and digital archive management systems. A helpful list along with reviews can be found at the Open Source Digital Asset Management website [8].

Particularly interesting for archive management are the DSpace and Fedora Commons projects. Both of these projects are now part of the DuraSpace initiative [9]. DuraSpace is a registered not-for-profit organization that sponsors three projects related to the long-term preservation of digital assets. The first, DSpace, is a turnkey application for digital repositories, including standards-based metadata support for digital repository management and discovery. The second is the Fedora Commons Project, a framework for building digital repositories that can be tailored to suit specific repository requirements and standards. Fedora Commons is the repository being used in the WGBH Open Vault project. The third is the DuraCloud project, which has been described previously in the cloud storage section 2.1.5, above.

Using open-source software doesn't automatically guarantee better quality software, or software that is any less vulnerable to becoming unavailable or unsupported. However, a standards-based open source project with good community support, combined with a permissive license and the support of a respected foundation (as copyright holder) may offer a cost-effective solution, while simultaneously supporting the objectives of both short- and long-term preservation of media.

A complete review of open source solutions for digital archive management is outside the scope of this paper; suffice it to say that interesting work is happening in this area, and that open source and standards-based solutions should be considered in any effort to implement a complete digital archive management solution.

2.2.4 Cataloguing Summary

Ultimately, the effort and emphasis placed on catalogue creation will be determined by the purpose, focus, and value of the digital

8 http://www.opensourcedigitalassetmanagement.org

9 http://www.duraspace.org

archive. If the archive's only purpose is to support short-term operational use, then simple non-standards-based and proprietary systems may be suitable. However, if the cataloguing exercise is focused on creating records that are designed to support long-term preservation – including the possible transfer of items to an institutional partner or repository – then a more formal approach, including the use of published metadata and cataloguing standards, is essential.

The ideal scenario would be to use tools that support both approaches, allowing media organizations to manage and arrange their digital archives such that the archive, in part, or in whole, can be positioned for different uses, transferred to other organizations, or made generally available for search and discovery by the public.

It's also important to remember that catalogues (at least electronic catalogues) are a digital resource just like the digital objects they describe, and so all of the management and preservation issues previously mentioned apply equally to the preservation of catalogue data itself.

3 AN ARCHIVE MATURITY MODEL

In this chapter an informal archive maturity model will be defined. The model will serve as an aid in the assessment of archive management efforts. It will also help to establish a shared vocabulary that can be used to describe or discuss archive management systems in general. The model should not, however, be considered a generally accepted method of benchmarking archive management systems.

It should also be noted that a low benchmark level should not be interpreted as a criticism of the organization being measured. At this point in the report, it will hopefully be clear that effectively managing large amounts of digital material is a challenging task for organizations of all sizes, and that the scope of such efforts should be proportional to the value of the material being stored. In the case of a collection that has limited value, a low benchmark would make perfect sense.

The model will be composed of the following five levels, with definitions for each:

1. No Archive: This is the lowest level of the model. The organization has no structured storage, shelving, or catalogue management system. Items are found on an ad-hoc basis, relying solely on the knowledge held by individuals (or individual and unshared systems) within the organization.

2. Basic Archive: In a basic archive, items are stored in a labeled and structured manner, such as a shelf-based system or some form of structured electronic storage, but without a structured catalogue. Items are found on an ad-hoc basis, relying on the knowledge of individuals (or individual and unshared systems) within the organization.

3. Catalogued Archive: This archive is based on a catalogue system. Items are stored in a labeled and structured manner, such as a shelf-based system or some form of structured electronic storage, and a comprehensive physical or elec tronic catalogue exists.

4. Advanced Archive: In an advanced archive, a policy-based archive management system is in place. Items are stored in a structured manner, and a comprehensive electronic catalogue management system exists, including advanced search facilities that support the discovery of catalogue items.

5. Standards-Based Archive: A standards-based archive has all the features of an advanced archive, as well as a system that is based on published standards for metadata description and classification – supporting the aggregation, or transfer of the archive between systems and organizations.

The following illustration shows the progressive maturity levels used to assess an archive according to this model.

FIGURE 3 - AN ARCHIVE MATURITY MODEL.

4

CASE STUDIES

This chapter presents a summary of site visits to several media organizations. It includes a description of the challenges each organization faces in digital archive management, as well as an indication of the state of each archive when compared to the archive maturity model.

4.1 Thai PBS

Bangkok, Thailand - Friday 2nd December 2011 - 11:00am-4:00pm

Thai PBS is Thailand's first "free-to-air" national public broadcaster. It began broadcasting in 2008, and is funded via taxes on tobacco and alcohol amounting to 2 billion baht (65 million USD) annually. Functioning under the umbrella of Thai PBS is the Academic Institute of Public Media. The Institute conducts research and evaluations, as well as promotes the knowledge and dissemination of public media in general. Included in the Institute is a learning center, within which a library and museum have been created to curate a growing news archive and a collection of public media materials.

FIGURE 4 - KHUN YOTHIN SITTHIBODEKUL IN THE THAI PBS FILE-BASED PRODUCTION CENTER WITH ITS RACK-MOUNTED SERVERS AND STORAGE SYSTEM.

Thai PBS has recently moved to modern new headquarters equipped with state-of-the-art high-definition broadcasting, production, and studio facilities. It is also the first media organization in Thailand to have a completely file-based all-digital workflow management system, including automation, play-out, storage and archive components. The system was procured through a vendor partnership, and includes equipment manufactured by Harris Broadcast Communications, Isilon (an EMC2 company) and Quantum.

The systems at Thai PBS are divided into News Production and Program Production (each with 600 hours of working storage capacity), as well as On-Air Automation.

The news and program departments each maintain their own archive of material that existed prior to the fully automated system being installed. These archives include shelf-based tape libraries and other departmental systems.

The new system provides an automated LTO-5 tape-based archive management system (contained in a Quantum near-line automated tape changer) with room for 400 LTO-5 cartridges, capable of storing up to 20,000 hours of broadcast material. The system includes a catalogue management application with support for rich metadata-based description and classification cataloguing tasks.

Thai PBS represents the high end of the spectrum in traditional broadcast and production facilities. However, the systems and media stored within the archive at Thai PBS are based on a proprietary Harris format, and therefore cannot be described as a Standards-Based Archive. It's unclear at this stage what strategy Thai PBS will adopt in order to provide open access to its archive. However, the Academic Institute of Public Media will begin a separate archiving exercise, using content selected from the Thai PBS production system, which will almost certainly be stored and managed in an open and standards-based manner.

FIGURE 5 - THAI PBS ARCHIVE MATURITY MODEL: A HIGH-END ARCHIVE MADE VULNERABLE BY BEING BASED ON A PROPRIETARY FORMAT, IT FALLS INTO THE CATEGORY OF AN ADVANCED ARCHIVE.

4.2 Madan Puraskar Pustakalaya (MPP)

Kathmandu, Nepal - Wednesday 14th December 2011 - 10:00am-12:00pm.

Madan Puraskar Pustakalaya ("Pustakalaya" meaning "library" in Nepalese) is the oldest library institution in Nepal, formed before the national library and national archives about 60 years ago. Although not a content creator per se, MPP was an interesting organization to visit in Nepal as it is attempting to deal with many of the physical and digital safe-storage and cataloguing issues associated with managing a large collection of material.

The library's initial mandate was to collect Nepali language material, and it now holds the largest collection of Nepali language texts anywhere, with 31,000 books and monographs, and over 6,000 periodicals. MPP began as a collection center, and then slowly developed a catalogue management system in order to index and manage its archive. The book collection at MPP has now been completely catalogued, with periodicals about 50% complete. MPP is also actively engaged in an open access program. Approximately two years ago, it began a migration from its in-house catalogued management system to the use of a MARC 21-based open source solution using the Koha Integrated Library System. As records are migrated into the new system, they are also made available publicly via the MPP website, where bibliographic records can be searched online.

FIGURE 6 – THE MADAN PURASKAR PUSTAKALAYA OFFICE IN KATHMANDU, NEPAL, WHERE STAFF ARE BUSY DIGITIZING AND CATALOGUING MATERIAL IN THE LIBRARY.

MPP's collection activities have expanded to include books, monographs, periodicals and ephemera concerned with Nepal as a whole.

MPP began an ambitious newspaper digitization project, attempting to collect 200-250 titles on a daily basis – digitizing all papers with a digital camera-based copy stand. DSpace was used as the cataloguing system, and the idea was to host the entire collection online. However, the minimum file size required for legible captures per page was 15 megabytes, and meant that for the general population with limited bandwidth, access to the archive would be impossible. Further complicating the project was the fact that there is no practical optical character recognition (OCR) software for the Nepali language, meaning that the text of newspapers could not be searched. MPP is also involved in software localization programs, and attempted to create a Nepali OCR solution. However, after two years, the best it could achieve was a solution that recognized two fonts, at two different sizes (with about 90-95 percent accuracy).

MPP also has an extensive microfilm program, some of which has been digitized.

MPP currently employs 20 staff and is funded by philanthropic donations from its chairman (who is also the chairman and co-owner of the media organization that publishes the prestigious Himal magazine and the Nepali Times), as well having previously received project-based funding from the Canadian IDRC, the EU, and the British Library Endangered Archives Programme. Since 2010, the government has also provided MPP with 1 million rupees per year (12,000 USD). Other regular but small grant funding is received from the South Asia Union Catalogue Project run by the University of Chicago.

MPP notes that for physical material, the environmental conditions in Kathmandu are better than most in South Asia, with an average humidity of 75% and average low and high temperatures that range from 2 degrees Celsius in the winter to 28 degrees Celsius in the summer. The main hazards for physical storage in Kathmandu are dust and pests (including silverfish). The microfilm collection at MPP is regularly quality tested for signs of degradation.

Digital material is currently stored in the disk subsystem of an office server, with approximately 3 terabytes of storage. Checksums are performed against digital files (mostly scanned images) in order to verify the integrity of such files once they have been transferred between devices. Interestingly, MPP indicated that it would like to continue with its microfilm program, as it is seen to be a good archive format however, MPP does not currently have the resources to do so.

The MPP photo collection is currently being made accessible via Flickr [10]. Although MPP is not entirely sure that this is the medium it ultimately wants to use for its photo archive, it has received a positive response to the material that has been published to date.

MPP's use of standards-based catalogue and archive management solutions means that it is at the "Standards-Based Archive" level in the archive maturity model.

FIGURE 7 - MPP ARCHIVE MATURITY MODEL: THE USE OF A MARC 21-BASED OPEN SOURCE SOLUTION VIA THE KOHA INTEGRATED LIBRARY SYSTEM FULFILLS THE REQUIREMENTS OF A STANDARDS-BASED ARCHIVE.

4.3 Radio Sagarmatha

Kathmandu, Nepal - Wednesday 14th December 2011 - 1:00pm-2:30pm

FIGURE 8 - A PORTION OF THE SHELF-BASED ARCHIVE AT RADIO SAGARMATHA.

Radio Sagarmatha is an independent community radio station based, and broadcasting, in Kathmandu. During the 1990 democracy period, the government decided to open print media to independent media organizations. Radio and television were not yet open to independent organizations; however, the Nepal Forum of Environmental Journalists (NEFEJ) along with three other organizations began the process of applying for the first independent radio station license. Five years later, in 1997, they obtained permission to begin broadcasting on FM 102.4 MHz as Radio Sagarmatha. Shortly after launching, the station was reorganized under the sole control of NEFEJ, which mandates that programming at Radio Sagarmatha should be politically neutral.

Radio Sagarmatha has an 18-hour daily program composed of music, news, entertainment, including talk shows (with call-in), and some community-created content. The station is funded through advertising as well as some project- and donor-based funding. News at Radio Sagarmatha is mostly a review of news from other sources, with a small amount of station-produced news content.

Although one of the original and earliest independent radio stations in Kathmandu,

10 http://www.flickr.com/photos/mpp_flr

Radio Sagarmatha now occupies a crowded radio space that is shared with 40 other radio stations operating in Kathmandu.

The station is mandated by law to keep at least a one-month history of broadcast content. As such, the daily broadcast (split on an hourly basis) is recorded to a computer in the archive room and stored in MP3 format.

Radio Sagarmatha employs 20 staff, two of whom are tasked with managing the station's library. The library is composed of a shelf-based audio archive, containing close to 10,000 audio CDs and audiocassettes – including an extensive collection of popular music that spans the lifetime of the station. A programming archive (as well as music) is stored on a PC in the library room. Winamp (a software application for managing audio libraries) is used to manage its PC-based collection of MP3 files.

A particularly popular program at the station is the Once Upon a Time (Uhile Ka Bajeka Palama) series. A program that was initially designed for an older audience, the series is narrated by an elderly member of NEFEJ, who recounts his personal life history and experiences living in Nepal. The program is broadcast weekly for 30 minutes and has been running for 15 years. Part of the series has been transcribed and published as an oral history book. The station believes that the complete series is stored on the computer in its library, although it has not been officially catalogued.

Another popular, daily, half-hour social and cultural program about life in Kathmandu was broadcast for nearly five years; however, it was not successfully preserved.

There has not been a consistent archive management strategy at Radio Sagarmatha – due in part to political and organizational changes that have occurred within the station over the past 15 years. The Radio Sagarmatha website contains a live stream from the daily broadcast as well as a section for the station's program archive; however, at the time of writing, neither the live stream nor archival content was available from the site.

A shelf-based library combined with a partially organized collection of data stored on a PC, without a comprehensive catalogue, mean that the archive at Radio Sagarmatha is a "Basic Archive" according to the archive maturity model.

FIGURE 9 - RADIO SAGRAMATHA ARCHIVE MATURITY MODEL: WITH A SHELF-BASED AUDIO ARCHIVE AND A PC-BASED COLLECTION WITHOUT A COMPREHENSIVE CATALOGUE, IT MEETS THE CRITERIA OF A BASIC ARCHIVE.

4.4 Press Council Nepal

Kathmandu, Nepal - Wednesday 14th December 2011 - 3:00pm-4:00pm

As with MPP, Press Council Nepal is not a content creator; however, it is attempting to manage a large newspaper archive, and as such represents an organization of interest to this study insofar as it is dealing with issues of safe storage and cataloguing.

The Press Council is a government organization, funded by the Ministry of Information and Communication (MOIC). It is tasked with enforcing the Press Council Act 2048 B.S (1992 A.D.) including a journalists' code of conduct. The Press Council employs 28 staff (including contract staff), with three members of staff assigned to archive management. The Press Council is also responsible for the categorization and circulation audit of Nepalese newspapers, dividing them into categories (A, B, C, etc.), based on format and size of circulation.

Since 1974, it has maintained a record of all newspapers published in Nepal. Its initial preservation efforts centered on binding printed newspapers into books for shelf-based storage. In 2001, it began a digital scanning process and has scanned 2.3 million pages to date. It started with two large-format Contex document scanners, which have recently been replaced by a single large-format Océ scan-

ner. Newspapers are scanned flat – including facing pages, which are then cropped and separated into individual pages – after which adjustments can be made for brightness and contrast. A link is then established between the scanned image and the publication page before writing the scanned images to CD or DVD.

FIGURE 10 – IT OFFICER RAM SHARAH BOHARA WITH THE SHELVING SYSTEM FOR THE CD ARCHIVE AT THE PRESS COUNCIL.

For about one year now, the archive system has had a dedicated server, and newspaper scans have been stored on hard disk as well as on master DVDs. However, during our visit, the server could not be demonstrated, having recently suffered a disk subsystem failure. A data recovery operation was underway in an attempt to restore six months' worth of scanned material.

The cataloguing system at the Press Council is based on a system of labeled CDs and DVDs in custom-built CD shelving systems with labeled drawers. Microsoft Office Word documents are used to record the key to the catalogue, linking publication titles to a particular drawer in the archive.

The Press Council staff admitted to having concerns about the data life expectancy of their optical media, since they have no way of knowing how long CDs and DVDs are going last. A proposal has been submitted for improved infrastructure and network storage in order to support the archive. Limited resources and little support for additional capital expenditures mean that funding is an issue at the Press Council.

With a shelf-based CD/DVD library and a catalogue stored in Microsoft Word format, the archive at Press Council Nepal is placed at the "Catalogued Archive" level in the archive maturity model.

FIGURE 11 - THE PRESS COUNCIL NEPAL ARCHIVE MATURITY MODEL: A SHELF-BASED LABELED CD/DVD ARCHIVE COMBINED WITH DOCUMENTS THAT CATALOGUE THE LOCATION OF MATERIAL MEETS THE CRITERIA FOR A CATALOGUED ARCHIVE.

4.5 Nepal Forum of Environmental Journalism (NEFEJ)

Kathmandu, Nepal - Thursday 15th December 2011 - 1:00pm-2:45pm

The Nepal Forum of Environmental Journalism (NEFEJ) was established on June 1st 1986. It was initially registered as a private company (as the Nepal Forum of Environmental Communicators), since at the time, there was no legal entity that allowed individuals to register as an association or to organize freely in Nepal. The constitutional changes during the democratic period allowed NEFEJ to re-register as a non-profit in 1990.

NEFEJ's overall goals have been to raise public awareness on the topic of sustainable development, as well as to lobby and advocate for environmentally friendly public policies. As of May 2011, NEFEJ had 118 members, composed of journalists as well as experts in the field of environmental science. It has been producing environmental print, radio (and later video) documentaries for 25 years.

NEFEJ does not receive regular funding, and most of its members are volunteers with regular employment elsewhere. It attempts to co-operate with like-minded organizations in other countries, occasionally receiving project-based funding as such. Staff size there-

fore expands and contracts based on levels of project activity.

FIGURE 12 - EDITING AN EPISODE OF AANKHI-JHYAL ON DVCAM TAPE AT NEFEJ.

NEFEJ began a weekly 30-minute video production series in 1994 titled Aankhi-Jhyal (which translates to The Peacock Window and means that through this window we can see outside from inside – but not inside from the outside). The show is broadcast on Nepal Television (NTV) as well as by other broadcasters across the country. A total of 742 episodes have been produced, reporting on social and environmental development issues throughout Nepal. The program is very popular, and represents a unique collection of material that describes social as well as environmental issues across the country – including footage and reporting from many remote and areas of Nepal that are difficult to access.

NEFEJ has an extensive shelf-based videotape library, including VHS, Hi-8, U-matic, Betacam, and DVCAM material. As an organization, NEFEJ was concerned about the complexity and risks associated with digital systems, and until now has not begun a large-scale digitization effort. Perhaps wisely instead, it has chosen to temporarily rotate expiring tapes back onto fresh digital DVCAM tapes. Most of the U-matic master tapes for Aankhi-Jhyal have been rotated to DVCAM. Raw footage, however, still remains on U-matic cassettes and NEFEJ no longer has a working U-matic player with which to rotate the remaining material.

FIGURE 13 - A PORTION OF THE TAPE LIBRARY AT NEFEJ.

The catalogue at NEFEJ is composed of lists for some master tapes, which describe the name and duration of the master copy of the program. Some logging information exists, although for the most part the catalogue at NEFEJ is dependent upon the information included on tape album covers for its shelf-based cassettes.

As a shelf-based archive, the library at NEFEJ is very well organized. However, since it lacks a comprehensive catalogue, it is a "Basic Archive," according to the archive maturity model.

FIGURE 14 - NEFEJ ARCHIVE MATURITY MODEL: AN EXTENSIVE AND WELL-ORGANIZED SHELF-BASED TAPE ARCHIVE, BUT WITHOUT A CENTRAL CATALOGUE, THE COLLECTION AT NEFEJ MEETS THE CRITERIA OF A BASIC ARCHIVE.

4.6 Association of Community Radio Broadcasters Nepal (ACORAB)

Kathmandu, Nepal - Thursday 15th December 2011 - 3:00pm-4:30pm

ACORAB was established in 2002 as an umbrella organization for community radio in Nepal. ACORAB's objective is to promote and support the development of community radio stations, and thereby to develop an informed citizenry and encourage freedom of expression through an inclusive, non-partisan medium for local and community audiences.

The association has grown from an initial membership of 19 to its current level of more than 200 community radio stations located throughout Nepal. ACORAB has received strategic funding from the BBC World Service Trust and the Danida Human Rights and Good Governance Advisory Unit (DANIDA/HUGOU), and has 22 staff members based in its Kathmandu office.

In 2009, ACORAB established the Community Information Network (CIN) – a satellite network connecting all of its regional community stations. ACORAB in Kathmandu transmits via fiber optic cable to a Kathmandu-based earth station, which in turn transmits to satellite. Local stations are able to receive the transmission from the Kathmandu office via K-band satellite receivers.

FIGURE 15 – BROADCASTING VIA SATELLITE FROM THE STUDIO AT ACORAB IN KATHMANDU.

Remote stations receive a morning news transmission from the central station, as well as compile and submit news and programming to the central office during the day. The central station prepares an evening satellite broadcast composed of the daily regional submissions, allowing regional news and programs to be broadcast across the country.

As with all media organizations in Nepal, ACORAB has a legal requirement to keep a recorded copy of its complete broadcast schedule for one month. As such, all program content is recorded to hard disk, and then periodically migrated from hard disk to DVD for longer-term storage in MP3 format. The DVDs are labeled and stored by date. The remote stations also record content to hard disk, although few remote stations are migrating from hard disk to DVD. ACORAB reports that occasionally DVDs in its archive cannot be read, and that the content in remote stations is vulnerable to PCs infected by viruses and malware. The ACORAB office in Kathmandu uses a licensed version of Kaspersky anti-virus software to protect its PCs and content before migration to DVD.

As a relatively young organization, the archive collection at ACROAB is not yet large, although it does have a complete, shelf-based, DVD archive for all broadcasts from 2009 until time of writing. The technical team at ACROAB manages the archive.

As a shelf-based archive, the library at ACORAB is well-organized. However, since it lacks a comprehensive catalogue, it is a "Basic Archive," according to the archive maturity model.

FIGURE 16 - ACORAB ARCHIVE MATURITY MODEL: AN ORGANIZED, SHELF-BASED, DVD ARCHIVE; HOWEVER, WITHOUT A SEARCHABLE, CENTRAL CATALOGUE, THE COLLECTION AT ACORAB MEETS THE CRITERIA OF A BASIC ARCHIVE.

4.7 The Antenna Foundation Nepal (AFN)

Kathmandu, Nepal - Friday 16th December 2011 – 4:00pm-5:30pm

AFN was established in 2002. Its mission is to create positive social and political change through the production of high quality radio content. AFN's mission is underpinned by its principled approach to content creation, including its advocacy for human rights, good governance, integrity and independence.

AFN has a production program that includes radio magazines, radio drama, live radio recording and television content – including a popular docudrama on youth involvement in community conflict resolution and peace building called Naya Bato, Naya Paila. Production topics also include HIV & AIDS, women's and children's health, water management, and general health – all supporting the foundation's objective of raising social awareness and changing audience behavior.

Content produced by AFN is made available to a countrywide network of roughly 300 radio stations (both public and commercial) for inclusion in local programming schedules. AFN also helps to support local media and capacity development through a program of training and support, including the development of a network of 250 local community reporters.

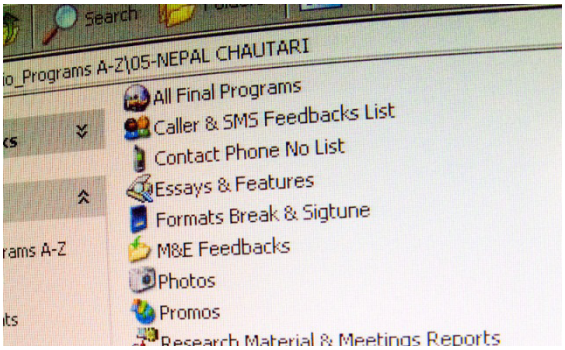

FIGURE 17 - AN EXAMPLE OF THE DIGITAL FOLDER STRUCTURE USED TO STORE COMPLETED PROJECTS AT AFN.

AFN employs 24 people and is funded by regular as well as project-based contributions from donor agencies, including The Asia Foundation, Search for Common Ground, SNV and Internews.

The foundation has attempted to keep archival copies of all produced content since it began operations in 2002. Material is stored on both internal and external hard disks (including a 1.5-terabyte external shelf-based hard disk) as well in CD and DVD archives. AFN reports that some material has been lost due to hard disk failures.

AFN does not have a dedicated archive function, with archival tasks performed as part of its regular operational and technical support activities. AFN uses a software application called Media Monkey to help with some of its audio management tasks. A structured, digital folder system is also used to store and locate produced content, including supporting material such as research documents, scripts, photos and other production-related assets.

AFN content, including a selection of archived programs, is also available from the foundation's website [11].

While visiting the foundation, this researcher also contacted two regional radio stations and asked them about media management, backup and archive strategies. In both cases, backup and archival actives were ad hoc, with material being stored on local internal and external hard disks. In some cases, material that was perceived to be of particular value would be taken home by local producers and stored on their personal computers for safekeeping.

Although well-organized, the archive at AFN is composed of a mixture of both internal and external hard disks, as well as a shelf-based CD/DVD collection. AFN does not have a comprehensive archive catalogue and so its archive is a "Basic Archive," according to the archive maturity model.

11 http://www.afn.org.np/

FIGURE 18 - AFN ARCHIVE MATURITY MODEL: A COMBINED HARD DISK AND SHELF-BASED CD/DVD ARCHIVE, WITHOUT A COMPREHENSIVE CATALOGUE, THE COLLECTION AT AFN MEETS THE CRITERIA OF A BASIC ARCHIVE.

4.8 Young Asian Television (YATV)

Colombo, Sri Lanka - Wednesday 21st December 2011 – 10:00am-4:00pm

YATV originally began as an international effort to create a youth-focused global production and broadcast network with operations in 22 countries around the world. In 2002, YATV was reorganized, and now focuses primarily on the production of content in Sri Lanka. It has a fully equipped production facility and studio in Colombo with 96 people on staff.

FIGURE 19 - A PORTION OF THE WELL-ORGANIZED TAPE LIBRARY AT YATV, WHICH IS CATALOGUED USING LIBRARY MANAGEMENT SOFTWARE.

Work at YATV includes documentaries, visual magazine series, short films and talk shows covering a range of social, political and developmental issues. Productions are aimed at promoting diversity and fostering greater communication and understanding amongst communities. Programs cater to Sinhala, Tamil and English audiences – both at home and abroad – and are designed to encourage constructive discussion and dialogue. Production at YATV includes several well-received and popular programs, such as Connections (an interactive dialogue between Sri Lankans both on the island and elsewhere in the world, with the objective of contributing to the reconciliation process in post-war Sri Lanka), The Interview (a guest interview and discussion program with topics including peace and politics, business and development, society and the environment, and culture and the arts), and Development Diaries (a weekly series discussing the process of moving on from war and conflict to peace and development).

YATV also undertakes commissioned work related to a variety of social issues, partnering with development institutions, including UN agencies and international NGOs, to produce and broadcast advocacy programs on matters of relevance to their various fields of work.

The archive at YATV is extensive, with close to 8,000 hours of material, including 500 one-hour programs on information and education, and over 6,000 30-minute edutainment and infotainment programs. The archive is tape-based and is stored in a dedicated tape library room with environmental controls in place. The shelf-based tape library is catalogued and managed using a software database application called Media Library Organizer Pro. YATV also has a dedicated librarian, whose responsibilities include maintaining the physical cassette library, tagging and labeling cassettes, keeping the catalogue up to date, as well as issuing and collecting tapes.

In 2009, YATV launched an ambitious online repository for content distribution called YaWaves [12]. The site features subject-classified content from the YATV archive, as well as services to support the contribution of content from other organizations and partners.

12 http://yawaves.com/.

As part of the YaWaves project, YATV began a digitization program in 2008, whereby approximately 80% of the tape library has been digitized and is stored online as well in a shelf-based set of external hard disks. However, the tape library is still considered the "master" library for YATV.

YATV also has an excellent technical and engineering department with skilled staff able to service and maintain legacy audio-visual equipment. With its shelf-based tape library and library management software, the archive at YATV is an "Advanced Archive," according to the archive maturity model.

FIGURE 20 – YATV ARCHIVE MATURITY MODEL: A WELL-ORGANIZED TAPE LIBRARY WITH A SEARCHABLE CATALOGUE APPLICATION MEANS THAT THE ARCHIVE AT YATV MEETS THE CRITERIA OF AN ADVANCED ARCHIVE.

4.9 Malaysiakini TV

Kuala Lumpur, Malaysia - Friday 6th January 2012 – 10:00am-4:00pm

Malaysiakini TV is a news video production desk at Malaysiakini.com. Malaysiakini.com is an online news portal and one of the few such portals that is funded by a successful, subscription-based revenue model. The editorial focus of Malaysiakini.com is 90% political news and events. The organization promotes journalistic ethics and is a non-political, non-partisan entity. Its independence from mainstream media in Malaysia means that it is able to publish articles and stories without interference from political parties, or those with a vested interest in the political process. Malaysiakini.com has 18 journalists publishing in English, Chinese, Malay and Tamil.

Malaysiakini TV currently employs four people, and first began in late 2006 with a grant and a mission to report on Indian issues in

FIGURE 21 – HARD DISKS CONTAINING EDITED MATERIAL AT MALAYSIAKINI TV.

FIGURE 22 – A PORTION OF THE MINIDV COLLECTION AT MALAYSIAKINI TV.

Malaysia. At the time, the parent .com site did not have a Tamil language section, and it was thought that an audio-visual format might be suitable for the dissemination of Indian-focused content in Malaysia. As well as online access, content from the site is downloaded and distributed to regional areas via NGOs operating within Malaysia. Tamil education, displacement, state workers, housing

issues, etc. all form part of the editorial content at Malaysiakini TV.

Funding at Malaysiakini TV has been provided by its parent organization, Malaysiakini.com, as well as The Friedrich Naumann Foundation, International Center for Journalists, Internews and others. Malaysiakini TV occasionally receives requests from other networks and broadcasters for content that is sold on a per usage basis.

In addition to content produced in-house, Malaysiakini TV receives stories for approval and publication to the Malaysiakini TV website from a network of trained citizen journalists.

Edited news stories average between six and eight minutes in length, with approximately four news stories published every day during non-parliamentary periods, and approximately six stories per day during parliamentary periods. Citizen journalist (CJ) contributions vary, with between one and two CJ stories published per day. The frequent publication schedule means that post-production and editing is kept to the minimum required to publish a story online.

The archive at Malaysiakini TV is composed of a shelf-based collection of over three thousand MiniDV tapes. Edited and digital material is stored in a shelf-based collection of hard disks. The archive is catalogued by publishing a directory listing for edited packages, which are organized by disk label and date. Malaysiakini TV has expressed concerns about the safety of its archive, and would ideally like to have a completely digital archive that can be duplicated and stored safely off-site.

A focus on the daily production of news and content, combined with limited resources, means that archive management at Malaysiakini TV meets the criteria of a basic archive according to the archive maturity model.

FIGURE 23 - MALAYSIAKINI TV ARCHIVE MATURITY MODEL: A COMBINED HARD DISK AND SHELF-BASED TAPE ARCHIVE, WITHOUT A COMPREHENSIVE CATALOGUE, THE COLLECTION AT MALAYSIAKINI TV MEETS THE CRITERIA OF A BASIC ARCHIVE.

5 CONCLUSION

One of the objectives of this study is to provide an overview of digital archive management, and in doing so, make clear that "going digital" has presented media organizations with unique challenges in the creation of effective archive management solutions.

Media organizations can't archive – not because they don't want to, but – because at the moment, there is no such thing as an "archive" in the traditional sense where digital media is concerned. The best any media organization can hope to achieve is the longest possible period between rotations from one media format to another. As such, organizations are in need of practical advice, tools and systems that will allow them to rotate, as well as safely store, find, and re-use content, for daily operations, as well as in an effort to catalogue content for longer- or long-term preservation.

This study divides archive management into two main components: namely, the safe storage of digital content and the creation of cataloguing systems that allow content to be organized, found and re-used.

The safe storage of digital media has turned out to be particularly challenging for small- and medium-sized organizations. A rapid increase in the capacity of affordable hard disks without a corresponding increase in generally available and practical backup solutions has meant that many organizations are struggling to manage volumes of data that previously would have only been found in professionally run data centers.

As a result, none of the organizations visited as part of this study (with the exception of Thai PBS) had effective backup strategies in place for their digital content. Nor were any of the organizations preparing duplicate or off-site data sets for safe storage. Nearly all organizations visited had experienced data loss due to the mechanical failure of hard disks or the inability to read optical media.

It should be noted, however, that the challenges of implementing regular backup and safe storage systems are not unique to the organizations visited as part of this study. Anecdotal evidence suggests that most media organizations are struggling in this area, with many having also suffered significant data losses as a result. Nor are these challenges unique to the problem of archive management. The procedures and systems required to safely store archival material overlap with the procedures require for good data management practices in general.

The cataloguing of digital content is described in this study from both a library science perspective, as well as from the practical point of view of media organizations, concluding that media organizations are understandably "production-focused" and that cataloguing and archive management solutions are often poorly integrated with the rest of the production lifecycle. However, as organizations accumulate content or completed work, the need for effective archive solutions becomes more apparent, especially as media approach the end of their effective shelf live, or as content begins to acquire historical value.

Archive management solutions for media organizations should ideally also attempt to incorporate some of the principles and open standards associated with the field of library science, allowing media organizations to manage and arrange their digital archives such that the archive, in part, or in whole, can be positioned for different uses, transferred to other organizations, or made generally available for search and discovery by the public.

Education and awareness programs will also form an important part of any digital archive management strategy. So too will guidance and practical suggestions that media organizations of all sizes can benefit from when attempting to implement an archive management solution. Two solution-focused appendices have been provided at the end of this report, aiming to serve as signposts for the

successful implementation of archive and data management systems for smaller organizations.

Opportunities also exist for media organizations that belong to a network of organizations to co-operate and share experience and knowledge in the area of archive management, perhaps even through the creation of centralized and shared deposit facilities for off-site storage and data safety.

This study also suggests that agencies that fund media development and content creation have an opportunity (and possibly even an obligation) to provide support in the area of digital archive management, in particular for public media organizations, and especially where material of educational and social value is being produced.

Another stated objective of this study is to suggest strategies for organizations that are attempting to preserve material that has social, historical and cultural value, or has in some way been designated as "digital heritage."

Respected national and international organizations are actively working to provide guidance and support in this area. The International Federation of Library Associations and Institutions has published the IFLA Manifesto for Digital Libraries, Bridging the Digital Divide: making the world's cultural and scientific heritage accessible to all [44]. The manifesto highlights the importance of access to information as a means to support universal goals of health and education. The manifesto also emphasizes the importance of open standards and protocols, and their role in facilitating dissemination and access.

The International Association of Sound and Audiovisual Archives (IASA) has also published an excellent guide titled Ethical Principles for Sound and Audiovisual Archives [45]. The guide describes the importance of specialist skills and infrastructure required by an organization attempting to hold audio and visual content for long-term preservation. In particular, the guide states that audio and video recordings, along with associated materials, shall be treated with appropriate respect and that mishandling by unskilled operators should be avoided.

Clearly the goals of access to information as a means to support development and education are valid at operational levels, as well as for the long-term preservation of digital heritage. However, it is particularly important for any organization claiming to offer practical assistance in the area of digital preservation to distinguish between activities that are designed to support the immediate objectives of an organization, and those that are designed to support the long-term preservation of content. For long-term preservation, organizations must evaluate their capabilities carefully and ethically, and where appropriate, seek partnerships with institutions or organizations that have the required skills and resources to support such efforts.

In summary, an organization attempting to preserve and catalogue its content must consider the resources it has available – with the level of investment in archive management ultimately determined by the purpose, focus, and value of the material to be preserved. This study examines the various issues surrounding the creation of digital archives, and provides guidance based on formal as well as practical approaches to such efforts. It also provides an archive maturity model to help evaluate the state of an organization's archive. Using this model, the case studies illustrate that most media organizations are in need of practical advice, tools and systems in order to preserve their content. Employing best practices, as embodied in the Standards-Based Archive model, is the most effective method for gaining the maximum benefit from a working archive, and is a requirement for any attempt at the long-term preservation of material.

BIBLIOGRAPHY

[1] Miniwatts Marketing Group. Internet Usage Statistics. http://www.internetworldstats.com/stats.htm (accessed Jan 10, 2012).

[2] Wikimedia Foundation, Inc. The Information Age. http://en.wikipedia.org/wiki/Information_Age (accessed Jan 13, 2012).

[3] Wikimedia Foundation, Inc. Knowledge Economy. http://en.wikipedia.org/wiki/Knowledge_economy (accessed Jan 10, 2012).

[4] Wikimedia Foundation, Inc. Digital Dark Age. http://en.wikipedia.org/wiki/Digital_dark_age (accessed Dec 1, 2011).

[5] K. D. Bollacker. Avoiding a Digital Dark Age. http://www.americanscientist.org/issues/pub/avoiding-a-digital-dark-age/1 (accessed Dec 1, 2011), 106.

[6] M. Patterson. Digital Dark Age Ahead? http://www.americanthinker.com/2011/01/digital_dark_age_ahead.html (accessed Dec 1, 2011).

[7] P. Ciciora. 'Digital Dark Age' May Doom Some Data. http://news.illinois.edu/news/08/1027data.html (accessed Dec 1, 2011).

[8] G. Nijhuis. Optical CD Code. http://www.laesieworks.com/digicom/Storage_CD.html (accessed Jan 15, 2012).

[9] Wikimedia Foundation, Inc. Optical Disc. http://en.wikipedia.org/wiki/Optical_disc (accessed Dec 31, 2011).

[10] K. Bradley, Risks Associated with the Use of Recordable CDs and DVDs as Reliable Storage Media in Archival Collections - Strategies and Alternatives; UNESCO, 2006.

[11] A. Brown, Selecting Storage Media for Long-Term Preservation; The National Archives, UK, 2008.

[12] The X Lab™. Optical media longevity. http://www.thexlab.com/faqs/opticalmedialongevity.html (accessed Jan 16, 2012).

[13] Optical Storage Technology Association. Recording Speed. http://www.osta.org/technology/dvdqa/dvdqa4.htm (accessed Jan 18, 2012).

[14] Wikimedia Foundation, Inc. Hard Disk Drive. http://en.wikipedia.org/wiki/Hard_disk_drive (accessed Jan 16, 2012).

[15] M. Lu and T.-c. Chiueh, Challenges of Long-Term Digital Archiving: A Survey; Stony Brook University, Stony Brook, NY-11794, 2005.

[16] L. Jordon. Hard Disk Warning! http://www.larryjordan.biz/hard-disk-warning/ (accessed Nov 30, 2011).

[17] D. Robb. Big Data Drives Rebirth of Tape Market. http://www.enterprisestorageforum.com/backup-recovery/big-data-drives-rebirth-of-tape-market.html (accessed Jan 17, 2012).

[18] Wikimedia Foundation, Inc. Linear Tape-Open. http://en.wikipedia.org/wiki/Linear_Tape-Open (accessed Jan 17, 2012).

[19] Hewlett-Packard, IBM and Quantum. LTO Background. http://www.ultrium.com/About/background.html (accessed Jan 17, 2012).

[20] Wikimedia Foundation, Inc. Linear Tape File System. http://en.wikipedia.org/wiki/Linear_Tape_File_System (accessed Jan 17, 2012).

[21] Hewlett-Packard, IBM and Quantum. LTO-5 Media Partitioning & Linear Tape File System (LTFS) Highlights. http://www.ultrium.com/technology/ltfs.html (accessed 3 May, 2012).

[22] IBM. IBM Linear Tape File System. http://www-03.ibm.com/systems/storage/tape/ltfs/index.html (accessed Jan 17, 2012).

[23] Free Software Foundation. Lesser General Public License (LGPL 2.1). http://www.gnu.org/licenses/lgpl-2.1.html (accessed Jan 19, 2012).

[24] Z. Kerekes. SSD endurance summary. http://www.storagesearch.com/ssdmyths-endurance.html (accessed Jan 19, 2012).

[25] M. Ricknäs. Report: SSD prices will fall below $1 per GB in 2012. http://www.macworld.com/article/164691/2012/01/report_ssd_prices_will_fall_below_1_per_gb_in_2012.html (accessed Jan 19, 2012).

[26] R. Metz. Cloud Computing Explained. http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/CloudComputingExplained/206526 (accessed Jan 19, 2012).

[27] E. North-Hager. USC Launches Cloud Archive. http://uscnews.usc.edu/science_technology/usc_launches_powerful_cloud_archive.html (accessed Jan 19, 2012).

[28] D. Raffo. USC Digital Repository uses cloud archiving for 8.5 PB of video. http://searchcloudstorage.techtarget.com/news/2240111428/USC-Digital-Repository-uses-cloud-archiving-for-85-PB-of-video (accessed Jan 19, 2012).

[29] M. Kimpton and S. Payette. Using Cloud Infrastructure as Part of a Digital Preservation Strategy with DuraCloud. http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/UsingCloudInfrastructureasPart/206548 (accessed Jan 19, 2012).

[30] L. Mearian. Start-up to release 'stone-like' optical disc that lasts forever. http://www.computerworld.com/s/article/9218881/Start_up_to_release_stone_like_optical_disc_that_lasts_forever (accessed Jan 20, 2012).

[31] Millenniata, Inc. M-DISC. http://millenniata.com/ (accessed Jan 20, 2012).

[32] The Long Now Foundation. The Rosetta Project. http://rosettaproject.org/ (accessed Jan 20, 2012).

[33] Wikimedia Foundation, Inc. Rosetta Stone. http://en.wikipedia.org/wiki/Rosetta_Stone (accessed Jan 20, 2012).

[34] A. Rose. Macro to micro etching. http://blog.longnow.org/2008/11/03/macro-to-micro-etching/ (accessed Jan 20, 2012).

[35] Japan Electronics and Information Technology Industries Associaton. JEITA Standards - Digital Cameras. http://www.jeita.or.jp/english/standard/html/1_4.html (accessed Jan 15, 2012).

[36] Library of Congress. Library of Congress Subject Headings (LCSH). http://authorities.loc.gov/ (accessed Jan 23, 2012).

[37] UNESCO. UNESCO Thesaurus. http://databases.unesco.org/thesaurus/ (accessed Jan 23, 2012).

[38] Library of Congress. MARC 21 Format for Bibliographic Data. http://www.loc.gov/marc/bibliographic/(accessed Jan 23, 2012).

[39] Library of Congress. Metadata Object Description Schema (MODS). http://www.loc.gov/standards/mods/ (accessed Jan 23, 2012).

[40] The Dublin Core Metadata Initiative Limited. Dublin Core Metadata Element Set, Version 1.1. http://dublincore.org/documents/dces/ (accessed Jan 23, 2012).

[41] Library of Congress. Encoded Archive Description - EAD. http://www.loc.gov/ead/ (accessed Jan 23, 2012).

[42] VRA Core Oversight Committee. VRA Core. http://www.vraweb.org/projects/vracore4/index.html (ac-cessed Jan 23, 2012).

[43] PBCore Public Broadcasting Metadata Dictionary Project. PBCore. http://pbcore.org/ (accessed Jan 23, 2012).

[44] IFLA. IFLA Manifesto for Digital Libraries. http://www.ifla.org/en/publications/ifla-manifesto-for-digital-libraries (accessed Jan 25, 2012).

[45] IASA. Ethical Principles for Sound and Audiovisual Archives. http://www.iasa-web.org/ethical-principles (accessed Jan 25, 2012).

[46] University of California. NOID. https://wiki.ucop.edu/display/Curation/NOID (accessed March 1, 2012).

[47] University of California. BagIt File Packaging Format. https://confluence.ucop.edu/display/Curation/BagIt (accessed March 04, 2012).

[48] Lbrary of Congress. NDIIPP Partner Tools and Services Inventory. http://www.digitalpreservation.gov/tools/#b (accessed March 04, 2012).

APPENDIX A- A SIMPLE DIGITAL ARCHIVE SOLUTION

This section describes a simple archive solution that can be used as a guide for the creation of an archive management system suitable for small- to medium-sized organizations.

The solution is based on a repository model, with a bundle- or story-based folder structure for archive items. It is ideally suited for produced content, including ancillary items and supporting documentation.

The design goals of the system are as follows:

1. To create a portable archive repository format that is not dependent on a specific vendor solution or platform.

2. To create a self-contained archive that can use any format (or carrier) or any combination of formats and devices to store archive items.

3. To simulate a shelf-based, labeled archive for easy retrieval of archival items.

4. To allow the cataloguing of the archive to be accomplished with readily available software and tools such as word processing documents, or spreadsheets, as well as optionally supporting a simple, relational database model for the development of a catalogue application.

5. To optionally support the integrity checking of archive items, for archive testing, or the safe transfer of archive items from one device to another.

The key to the system is the assignment of two identifiers, or IDs, to individual archive items and to devices (in this case this report will use the term "device" to refer to any carrier or format used to store archive items).

In a simple, folder-based system, the name of top-level folders can be represented by the name of a series, or production, or even the title of a story. However, by using an ID, one can create a normalized repository that will lend itself to more efficient use of media, as well as support one of the stated design goals – the optional use of a database application.

Identifiers, or IDs, come in several different flavors, including numerical, mixed alphanumerical, and those with no meaning at all, such as surrogate or opaque identifiers [46]. For this solution, this report will use an identifier that has both semantic and sequential components. The exact scheme used can be left up to the implementing organization to design.

An identifier with the form ORGANIZATION-LOCATION-NUMBER will be used, where the number is a sequence starting from 1.

For example, the identifier for an archive item of an imaginary media organization called ULSHMedia with an office in Kathmandu might look like this:

ULSH-KTM-0000001

We'll create a similar identifier for a device. In this case, "XHD-000001" will be used as the identifier and label for an external hard disk. A physical label would be placed on the device, ideally using a label printer with label media that is resistant to fading or degradation over time. (NOTE: Since an LTO/LTFS formatted tape cartridge emulates a regular folder and file system, this device could just

as easily be an LTO cartridge, with archive items stored in exactly the same as they would be stored on any other folder and file-based device. As an example, a label on a shelf-based LTO cartridge might be "LT0001").

Here is the folder structure for archive item ULSH-KTM-0000001 that will be stored on XHD-000001

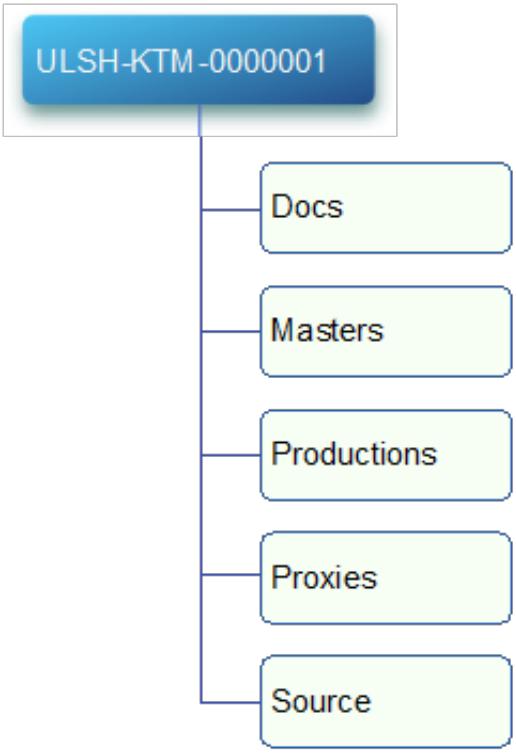

FIGURE 24 - A SIMPLE FOLDER STRUCTURE FOR ARCHIVE ITEM ULSH-KTM-0000001.

The containing folders for ULSH-KTM-0000001 are described as follows:

The "Docs" folder can be used to store production-related documentation, including planning, treatment, and post-production documents such as scripts, transcripts, translations, etc.

The "Masters" folder can be used to contain any finished media – that is, the master or completed audio or video file that will be used for broadcast or distribution. There can be multiple master files, for example language versions or long and short versions, etc.

The "Productions" folder can be used to store the edited project, including any software editing project files, sequences, and clips, etc.

The "Proxies" folder can be used to store low-resolution versions of the contents of the Masters folder, suitable for upload to online systems, or for use as previews in catalogue management applications. Having a structured folder system with master files in a known location also lends itself to the automated production of proxies using batch or automated proxy generation software.

The "Source" folder holds raw audio or video material that has been captured from recording devices like video cameras and audio recording equipment. The folder can be subdivided into card or capture numbers, for example folders "01," "02," "03" - each containing the complete contents of an AVCHD SD card from a recording camera.

Additional folders such as “Art,” “Communications,” “Output,” “Email,” etc. can be added as required.

With a simple labeling scheme for both archive items and devices, a catalogue can now be created to record information about archive items, including which device they are currently being stored on.

Here’s an example spreadsheet, with a single entry for the archive item and device above.

	A	B	C	D	E	F	G	H
1								
2	ID	Date	Title	Description	Subject	Copyright	Series	Location
3	ULSH-KTM-0000001	2012 03 03	Connections Episode 1	An interview with x, y, and z on the impact of devleopment in the region.	Politics, Environment, Heath	ULSHMedia	Connections	XHD-000001

FIGURE 25 - A SIMPLE SPREADSHEET-BASED CATALOGUE.

From our description of cataloguing in section 2.2, it should be clear that this is a simple and non-standards-based catalogue, however, the spreadsheet above could be adapted to support a metadata scheme such as Dublin Core. A controlled vocabulary of subject terms (like the UNESCO thesaurus) could also be used for subject classification.

The advantages of this simple system include being able to search the spreadsheet for catalogue items, determine the location of a catalogue item, and being able to create and manage the system without any dedicated catalogue management software. Storage devices can also be shelf-based, satisfying the design goal of being able to create an easy-to-use, shelf-based library.

The disadvantages of this system include the lack of data entry validation, the need to manually increment the sequence number of archive item IDs, and the limited space available in a spreadsheet for more descriptive fields such as a “notes” or a “synopsis” field. Another significant drawback to the system is that it is a single-user system, and so care must be taken to ensure that only a single user updates the document at any given time.

Another design goal of this system is to support the optional use of a database application. A complete design for such an application – including system architecture, authentication, authorization, search, etc. – is outside the scope of this report; however, a simple entity relationship model is provided below that illustrates the many-to-many relationship between the two core system objects – items (depicted as catalogue “entries”) in red, and “devices” in blue.

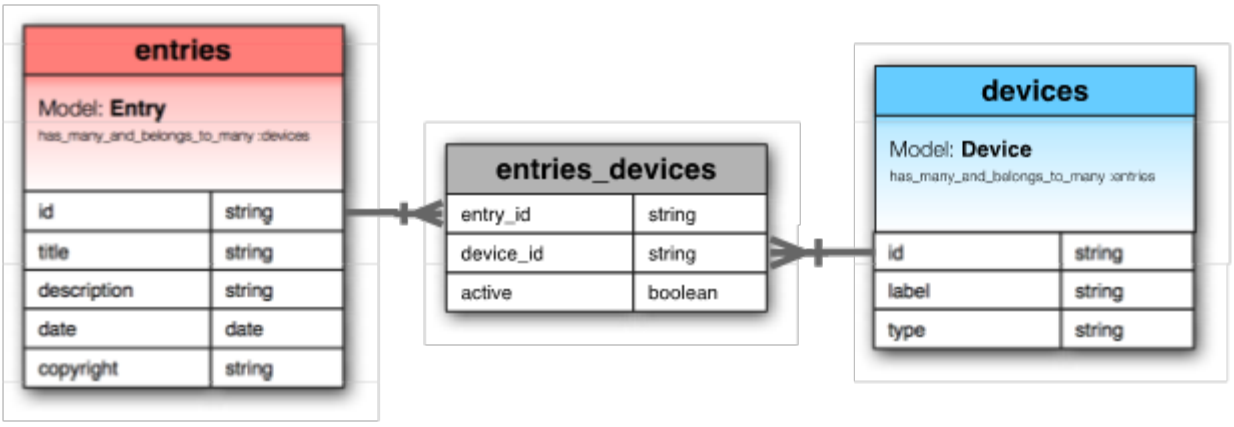

FIGURE 26 - AN ILLUSTRATIVE OBJECT MODEL FOR CATALOGUE ENTRIES AND ARCHIVE DEVICES.

While a catalogue entry can be associated with many devices, only one device would be considered the current, or active, device at any given time. This supports a device history, indicating where an archive item may have been previously stored, as well as a rudi-

mentary versioning system in cases where an archive item is retrieved from the archive, updated and then placed on a new device.

Care should be taken when considering any software development exercise, in particular the cost of ongoing support and maintenance of any custom-developed system. In the case of an archive management application, all of the issues highlighted previously in this report concerning data life expectancy and technical obsolescence apply equally to the safety and preservation of an archive database, and are the reason that projects like The Fedora Commons Repository exist – helping to safeguard access to archival data.

One of the last design goals of this system is an optional integrity test that would allow the integrity of the contents of an archive folder to be checked, in particular when archive items are transferred from one device to another, or from one organization to another. Fortunately, there is a published standard for such tests against a hierarchical folder structure, including software that can assist in both manual and automated verification. The standard is called BagIt and it defines a hierarchical file-packaging format for the exchange and verification of digital files [47,48].

Archive items that implement the BagIt system would require a slightly different folder structure, with all of the data containing folders (like "Docs," "Source," "Masters," "Proxies," etc.) placed in a subdirectory called "data." The root level folder for the archive item would then contain the BagIt manifest and BagIt version text documents. A BagIt manifest contains a digital checksum that has been calculated for every file in the data directory, and can be used by the receiving system to verify that all of the files have arrived intact.

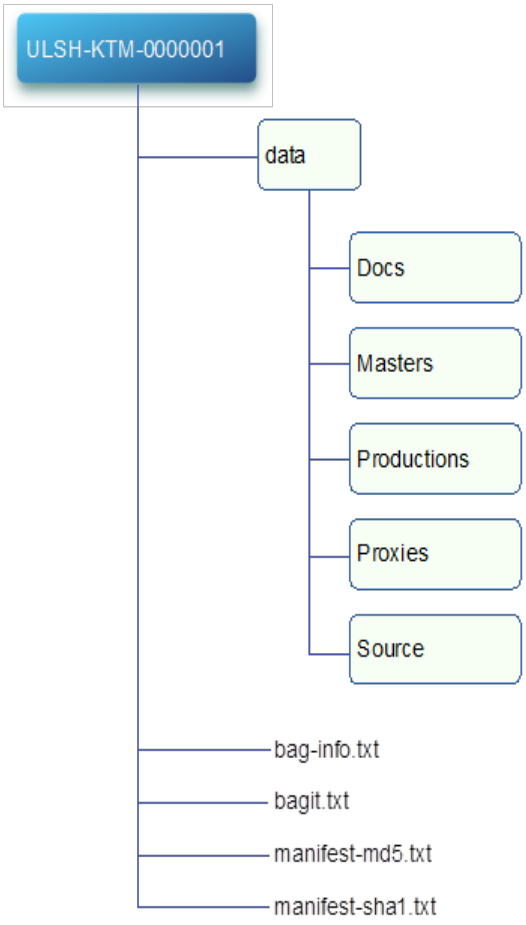

FIGURE 27 - AN UPDATED ARCHIVE ITEM FOLDER STRUCTURE TO SUPPORT CONTENT INTEGRITY CHECKING USING THE BAGIT SPECIFICATION.

APPENDIX B- A MODEL ARCHIVE STATION

This section will describe a single-user ingest and archive station, suitable for the capture and digitization of legacy tape formats, as well as for the preparation (or staging) of archive items before cataloguing and migration to archival media.

The system – at the time of writing – represents a state-of-the-art station in terms of storage, capacity and transfer speed. One of the key design goals of the system is the highest possible transfer speed and throughput in order to reduce operator time, and so the system described below will make use of the new Thunderbolt I/O interface from Intel. The system will also demonstrate how an LTO/LTFS tape system can be used without any dependency on proprietary tape backup software.

1. 17" Apple MacBook Pro with Thunderbollt ™
2. Pegasus R4 RAID 8TB Storage with Thunderbolt™
3. Blackmagic UltraStudio 3D Thunderbolt ™ Capture Device
4. HP Ultrium 3000 LTO Tape Device
5. Magma ExpressBox 3T Thunderbolt ™ to PCI Express Chasis or Magma ExpressBox 1
6. ATTO ESAS-H680 SAS HBA

FIGURE 28 - A MODEL INGEST AND ARCHIVE STATION.

An Apple MacBook Pro was selected for this solution since, at the time of writing, Apple is the earliest adopter of Thunderbolt technology. LTO/LTFS software (including command-line utilities) for LTFS/LTO tape drives is also being released by HP and IBM for the Linux and Mac OS X platforms before Microsoft Windows.

Legacy archive formats from tape players or older tape formats can be captured using the Blackmagic UltraStudio capture device (including RS422 controlled decks). Material that is already in a digital format can be transferred to the system using regular Ethernet, Firewire-800, USB or optical interfaces.

The Pegasus R4 RAID device can be configured as RAID 5 for short-term storage and staging of archival material before transfer to

removable media such as data tape, external hard disks or optical media. Incremental backups should be performed to protect staged material stored on the RAID device.

HP has released a free, single-user version of StoreOpen LTO/LTFS automation software, which allows LTFS formatted LTO-5 cartridges to be presented as a collection of folders for easy transfer to data tape. However, at the time of writing, LTO tape drives that natively support Thunderbolt are not available, so the Magma Thunderbolt to PCI Express or Magma ExpressCard to PCI Express adapter must be used in order to connect to the LTO tape drive via a Serial Attached SCSI (SAS) interface. Command-line LTO and LTFS utilities that do not require the StoreOpen application can also be used to format and transfer files to the tape drive.

Archival material can be checked, organized and catalogued according to the simple archive management scheme described in Appendix A. When 1.5TB (or less) of archival material has been organized into ID-based folders, it can be transferred to a labeled LTO-5 data tape for long-term storage. At the time of writing, a free LTO label generator is available from http://tapelabels.librelogiciel.com/ and weatherproof polyester OL173 LTO labels could be purchased from Online Labels at http://www.onlinelabels.com/OL173.htm.

Master and duplicate data tapes should be created for local as well as off-site storage (this should apply to any format used, including external hard-disks or optical media).

For larger organizations, the roles of ingest, material organization, cataloguing, and transfer to archival media would typically be divided among multiple stations, roles and departments. However, for smaller organizations, the solution described above will comfortably manage audio and video collections ranging from a few hundred gigabytes to many terabytes in size.

When combined with a scheme similar to the one described in Appendix A, the system can be used to create a shelf-based, labeled and easy-to-access archive with material safely stored both on- and off-site.

ABOUT THE INTERNEWS CENTER FOR INNOVATION & LEARNING

The Internews Center for Innovation & Learning supports, captures, and shares innovative approaches to communication through a creative program of research and development worldwide. Founded in 2011, the Center seeks to strike a balance between local expertise and needs and global learning in order to develop a comprehensive approach to understanding and catalyzing information exchange.

In Internews' 30-year history of promoting independent media in more than 75 countries around the world, the last five years have arguably seen the most changes in the global media and journalism environment. Across all Internews programs, adoption of cutting-edge technology is integral to advancing the work of the journalists, bloggers, citizen reporters, scholars and others who provide a vital interpretive role for their communities. The Internews Center for Innovation & Learning deepens and enhances our capacity to link existing expertise to research that helps define, understand and monitor the critical elements of changing information ecosystems and to pilot projects that apply and test the data, platforms and digital tools to meet information needs of specific communities. This is far from a solo endeavor. A network of partners, ranging from technologists to academics to activists is critical to creating and sustaining a dynamic and iterative collaborative space for innovation.

Internews Administrative Headquarters
PO Box 4448
Arcata, CA 95518 USA
+1 707 826-2030

Internews Washington, DC Office
1640 Rhode Island Ave. NW Suite 700
Washington, DC 20036 USA
+ 1 202 833-5740

www.internews.org
info@internews.org
www.facebook.com/internews
www.twitter.com/internews

Internews is an international media development organization whose mission is to empower local media worldwide to give people the news and information they need, the ability to connect, and the means to make their voices heard.

Through our programs, we improve the reach, quality, and sustainability of local media, enabling them to better serve the information needs of their communities.

Formed in 1982, Internews is a 501(c)(3) organization headquartered in California. Internews has worked in more than 70 countries, and currently has offices in Africa, Asia, Europe, the Middle East, and North America.

The Internews Center for Innovation and Learning aims to harness the potential of digital technologies and innovative approaches to better meet the information needs of communities around the world. The Center serves as a hub to inform and engage others in the fields of media, information technology and development.

Internews
Center for Innovation
& Learning