Planning the digitization, storage and access of large scale audiovisual archives

Matthew Addis[1], Freddy Choi[1], Ant Miller[2]

[1]IT Innovation, 2 Venture Road, Chilworth Science Park, Southampton, SO16 7NP, UK
{mja,fc}@it-innovation.soton.ac.uk, http://www.it-innovation.soton.ac.uk
[2] BBC Information and Archives, Windmill Road, Brentford, Middlesex, TW8 9NQ, UK
ant.miller@bbc.co.uk, http://www.bbc.co.uk

This paper presents ongoing work in PrestoSpace on how broadcast archives can plan large-scale, long-term digitization and storage projects. In our approach, carrier decay, technical obsolescence, and rapidly falling costs of mass storage are represented as a series of statistical and predictive models. The models include ongoing migration within a digital archive. The objective is to allow archive managers to investigate the trade-offs between how many items to transfer, the cost of transfer and storage, how long it will take, what quality can be achieved, how much will be lost, and what digital storage solutions to adopt over time. The process and models are based on digitization projects conducted by large broadcast archives that are currently migrating their collections into digital form. Whilst our focus is on broadcast archives, our findings should be readily transferable to other scenarios where there is a need to store large volumes of digital data over long periods of time.

1. Introduction

The PrestoSpace [1] Project (www.prestospace.org) plans to push the limits of the current technology beyond the state of the art to provide products and services for bringing effective automated preservation and access solutions to Europe's diverse audiovisual collections. Part of the work being done in the project concerns cost models for long term preservation of audiovisual archives. A large audiovisual archive, e.g. as held by a national broadcaster, typically contains several million discrete items stored on tens or even hundreds of kilometers of shelving. Material has often been accumulated over fifty years or more (longer in the case of film), on a range of carriers (film, tape, discs), uses a range of formats (analogue and digital), and will be in varying states of decay (vinegar syndrome, binder hydrolysis, mould etc.). PrestoSpace conducted a survey [2] of 31 institutions across 20 countries, which between them hold an estimated 20 million items of audio and video material. UNESCO estimate the worldwide figure to be over 200 million hours, much of which is at risk of being lost and hence is the subject of international appeals for preservation [3]. Technical obsolescence and physical decay are major concerns with a large percentage of material being fragile, damaged or difficult to play [4]. This is causing significant and irrevocable loss of our cultural memory. This situation is not

improving. PrestoSpace found that, for the archives it surveyed, annual investment into preservation amounts to 30MEuro, which is enough to preserve 1.5% of holdings each year [6]. However, for tape based material with an average 20 year life expectancy [5] this will result in the loss of 40 % of existing material by 2045 in the best case scenario, and at worst 70% by 2025 [6]. For archives that want to maintain the ability to use their content, e.g. broadcast archives, migration into a sustainable digital form is the only answer, and the need for large-scale, long-term planning of digitization and storage projects is clear. The problems described above are not exclusive to broadcast archives, but apply much more widely to any institution archiving large volumes of data on media (e.g. tapes and optical disks) where concerns exist for long term safety and obsolescence.

2. Cost models for audiovisual archive planning

Cataloguing of audiovisual archives is often incomplete or is done at a level that does not allow the specific content and condition of each media item to be identified. As a result, archives do not know exactly what content they hold, how many items they have, or what condition those items are in [6]. The first stage of planning preservation (migration and storage in a new digital archive) is to create a map of the current archive, both from a technical (carriers, formats, conditions) and content (genres, value) perspective. The sheer volume of items means that statistical sampling is often the only viable approach. The statistical map is then used to assess the urgency and cost of preservation of the archive as a whole. Since a statistical approach is taken, the condition and content of each individual item is not known until the execution of the preservation project itself, i.e. when items are taken off the shelf and inspected prior to transfer. Therefore, the preservation workflow necessarily involves an assessment and sorting stage that determines what action is taken on an item by item basis. The migration process is shown in Figure 1

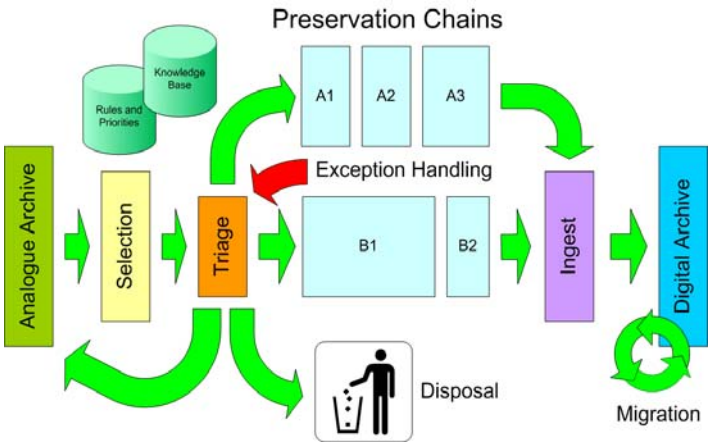

Figure 1 Diagrammatic view of the migration process.

Efficient operation of this workflow is essential to minimizing costs and typically consists of allocating items to one or more transfer chains each of which are matched to a specific combination of carrier type and condition (e.g. tape in good condition, or tape that is 'sticky' and hence needs baking). These transfer chains then feed into the digital archive. Since discrete digital media suffers from many of the same problems of discrete analogue media, then to ensure no further loss, the digital archive is operated with an inbuilt and relatively short term (3-5 year) migration strategy. The 'triage' of items before transfer is an essential step of the process and is underpinned by a preservation strategy developed through cost modeling. This strategy is implemented as a set of rules that determine what action is taken, and is supported by a knowledge base that allows the cost of preserving each item to be estimated based on easily observable features such as visible condition or chemical markers. The rules that determine whether to transfer an item or not include the business priority for transfer based on assessment of the value of an item. The two major factors that affect the cost of transferring items into a form that can be digitally archived are degradation (e.g. chemical decay, or wear and tear) and technical obsolescence (media players, spare parts and skilled operators are not available). There are no hard and fast rules for calculating life expectancy due to decay (for example, tape condition depends on manufacturer, production batch, storage conditions, frequency of handling). Therefore, we model degradation as a matrix with transitions between different media condition states. This is shown in Figure 2. This describes the probability of an item being in a condition Y in year n+1 if it was in condition X in year n. The use of a matrix allows more than one degradation process (e.g. chemical decay, wear and tear) to be modeled simultaneously. The matrix is populated using statistics built up over time (e.g. by monitoring the archive) or by using the experiences of other archives with similar media.

Condition	*Future Condition*				
	Playable	**Dirty**	**Fragile**	**Damaged**	**Unrecoverable**
Current Condition	*% of condition*	*% of condition*	*% of condition*	*% of condition*	*% of condition*
Playable	90%	10%	0%	0%	0%
Dirty		80%	15%	5%	0%
Fragile			70%	20%	10%
Damaged				60%	40%
Unrecoverable					100%

Figure 2 Condition change matrix for modeling media degradation

Each observable media condition (e.g. a tape splice being good, dry, sticky or detached) is mapped to an operation required for transfer (e.g. cleaning, re-splicing) which can then be associated with a cost (e.g. 80Euro per hour of tape). In this way, observable conditions in an archive sample are converted into costs for transfer or repair. The estimated cost of transfer for the archive as a whole is then projected over time using the condition change matrix. Other time dependant factors such as technical obsolescence and inflation are modeled as affecting the cost of transfer or repair, e.g. the obsolescence lifecycle of a particular tape player can be modeled as a cost multiplying factor which changes each year (e.g. when a player is first discontinued, when manufacturer support is no longer available, when the second hand market is ultimately depleted of players etc.).

The digital archive ingesting the items produced by migration is modeled as a series of mass storage devices (tape robots, hard disk arrays etc.) that contain digital media

(tapes, disks). Archive experience reveals that mass storage migration typically takes place on a 3-5 year timescale, with interim refreshing of media (e.g. moving to higher capacity tapes) taking place more often. This process is shown in Figure 3.

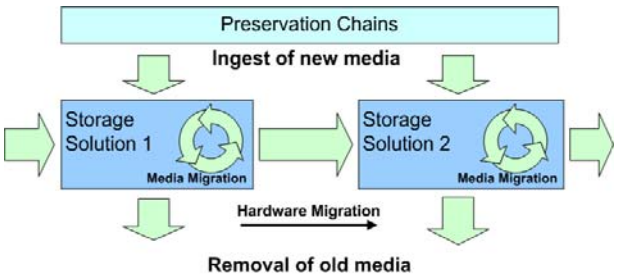

Figure 3 Ingest, migration and disposal model for a digital archive

The rapid and continual fall in cost of digital storage along with the relatively short lifetime means that it makes sense to purchase storage 'on demand' rather than as a single upfront investment. Items ingested to the digital archive are assigned to different storage solutions depending on their type, for example CD quality audio might be assigned to disk whilst high quality video at $80MBs^{-1}$ might be assigned to tape. The relative cost effectiveness of different solutions changes over time, for example allowing transition from tape-based to disk-based solutions. Our model maps ingested items to digital media types and has a migration matrix between media and hardware solutions based on estimated life expectancy. The combined profile of ingest and migration determines the storage needed and cost over time. Combined with the transfer plan this produces a projection of cost and loss on a yearly basis which can be optimized using a spreadsheet to investigate tradeoffs. The result is a projection of long term transfer and storage needs as shown in Figure 4

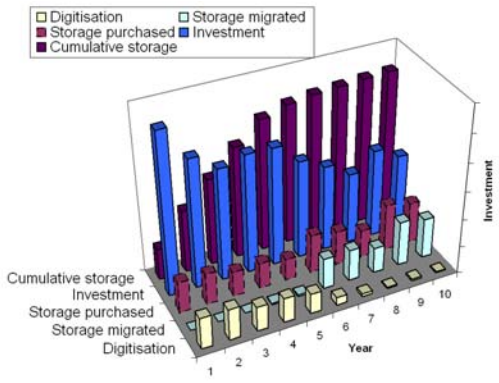

Figure 4 Transfer and storage projection

3. Future Work

The next stage of our work is to calibrate the models using data from archives on their ongoing preservation projects (expected by the end of 2005). This includes media degradation rates, transfer costs, actual costs for storage systems, and inflationary factors. Over the next two years, attention will turn to the needs of small archives.

4. Conclusions

Broadcast archives face many decisions when planning the long term preservation of their assets. These problems are not unique to broadcast archives and apply equally to any large archive held as items on shelves (analogue or digital) which doesn't have a program of ongoing migration. Cost modeling can assist with identifying the best strategy and allows the year-on-year costs and losses to be calculated and trade-offs investigated. Ongoing migration coupled with an 'on demand' mass storage purchasing strategy provides a cost effective model for long-term digital archiving of large audiovisual collections.

5. Acknowledgements

This work is supported by the European Commission (FP5 IST PrestoSpace project). The authors would also like to thank archive staff at the BBC, INA and ORF for kindly sharing their knowledge and expertise.

References

[1] EC FP5 IST PrestoSpace project. Project website: http://www.prestospace.org
[2] "Final report on users requirements". Deliverable D2.1 of the PrestoSpace project. Available from http://www.prestospace.org/project/public.en.html
[3] IAT/IFTA International Appeal for the Preservation of the World Audiovisual Heritage. http://www.fiatifta.org/aboutfiat/policy/petition/index.php
[4] Survey of Endangered Audiovisual Carriers 2003. George Boston. Technical Committee of the International Association of Sound and Audiovisual Archives with assistance from the International Council of Archives on behalf of UNESCO's Information Society Division United Nations Educational, Scientific and Cultural Organization Paris, 2003. http://portal.unesco.org/ci/en/ev.php-URL_ID=13437&URL_DO=DO_TOPIC&URL_SECTION=201.html
[5] "Magnetic Tape Storage and Handling A Guide for Libraries and Archives" Dr. John W.C. Van Bogart. National Media Laboratory. June 1995. http://www.clir.org/pubs/reports/pub54/
[6] Annual Report on Preservation Issues for European Audiovisual Collections (D22.4)" Deliverable D22.4 of the PrestoSpace project. Available from http://www.prestospace.org/project/public.en.html