



How to make an online archive

Case study of the British Pathe archive digitisation

Date: 1 September 2004

Written by:

Paul McConkey

Cambridge Imaging Systems Ltd
The Grange
44 High Street
Willingham
Cambridge CB4 5ES

Tel: 01954 262000

Fax: 01954 262001

paul.mcconkey@cambridgeimaging.co.uk

Introduction

British Pathe have been creating newsreel films since the earliest days of motion pictures. Many readers of a certain, more distinguished, age will remember these films which were shown before the main feature at cinemas across the land. The newsreel activities gradually came to an end as television news replaced newsreel films as the public's source of moving visual reportage. Almost all of the newsreel films and many of the rushes were collected so British Pathe ended up with an archive of over 3000 hours of newsreel film collected between the start of the twentieth century and around 1970.

Since the news gathering effort was stopped, the film archive has been increasingly used as a historical resource by film and television producers of all genres. The archive is now managed by ITN Archives and the entire collection is available online at <http://www.britishpathe.com>.

The creation of the online archive was a two-year effort that followed almost two years of discussion, planning and fund-raising. A technical solution needed to be found and a team of people gathered together to transfer aging 35mm film prints to a digital archive. British Pathe investigated many possible solutions, but after a software demonstration at the Ministry of Defence, eventually chose the Imagen archive management system from Cambridge Imaging Systems.

Who are Cambridge Imaging Systems?

CIS are a small development company with around 10 staff. We have been developing software for media archive management systems since 1991. We only work on large-scale projects for corporate and government and academic clients. Our software systems have been developed over the last fifteen years to provide a flexible architecture, able to interface with existing software and able to integrate with existing working practices. As we are a small development company, we are able to commit high-level staff to a project which helps to successfully complete projects and to build very good relationships with all of our clients.

Our archive management system, Imagen, contains a number of modules that are designed to work together in a coordinated way to reduce the manual intervention required to manage very large-scale media archives. It is possible to use Imagen as a complete media asset management system, with an associated website and e-commerce system.

We have developed all of the code in our digital media archive management system including the text retrieval DBMS, stream servers, MPEG transcoders and all other utility and service applications. Because of this we are able to customise systems extensively to fit very closely with customer requirements and work practices.

We prefer to use industry standard protocols and interfaces and embrace the use of open-source code to speed development and to keep costs down. We have experience of integrating our software with many different types of administrative, commercial and industrial systems.

The Imagen asset management system includes ingest, storage, database management, web access, e-commerce and output applications. For this encoding proposal we will only be using the ingest and asset management capabilities of Imagen.

Imagen was originally designed to manage a very large photographic in 1991 and has since been developed for use in many large archive management and media monitoring systems. Imagen Server's indexing system allows for very fast searching through metadata.

The core of the system is our database management system and media control centre software. This is a very fast text retrieval system with built in archive management functionality allied to software that can manage a very scalable collection of encoders, transcoders and distribution systems.

How to make an online archive

Film transfer

The original 35mm film needed to be transferred to a medium more suited to automation and easy handling. The earlier film is based on nitrate stock which is extremely flammable and must be treated as a hazardous material.

The solution was to make telecine transfers to Sony DigiBeta tape. Although modern 35mm stock has a much better resolution than Sony DigiBeta tape, the older film didn't suffer from any noticeable losses and became far easier to handle.

When the film transfer was done, the descriptions and other data describing each clip were completely updated. Importantly, the tape time codes for each clip were recorded. This allowed many of the succeeding steps to be completely automated.

Even at this stage, the usability of the archive had been greatly improved. All 90000 clips could be easily and quickly copied from DigiBeta masters to VHS show reels or professional quality video sub-masters. As the descriptions had been completely revised from a modern perspective, searches became much more relevant.

Choosing digital formats

DigiBeta is a digital tape format so, in a sense, the telecine transfer digitised the film. We needed to go further, to create a digital archive stored on an online storage system accessible by computer. The choice of digital video format for this was the subject of many hours of research and discussion. At the time, the only sensible format was MPEG-2. This would have longevity, based upon its use for digital TV broadcasting and DVD, and offered the best compromise between quality and compression.

Archive quality format

Having plumped for MPEG-2, what bit rate should we use? Production quality bit rates and 4:2:2 encoding would have been too expensive because of the cost of storage. In any event, we can always go back to the DigiBeta tapes to encode really high bit rate video, or to the film itself if that isn't good enough.

Sample clips from the early 1900's through to the late 60's were encoded and through trial and error we selected a bit rate of 5Mbits/s. This was the lowest bit rate we could use before seeing noticeable artefacts. We knew that we were going to end up with over 10TB of video files at this bit rate and couldn't afford to go any higher. Nowadays we would probably just go straight for 8Mbit/s and still spend less than half the amount on storage.

The earliest footage gave us a bit of a shock when we saw it! The video image looked like it was vibrating slightly. It turns out that the nitrate stock tends to shrink so the image jumps about in the projector gate. Fortunately MPEG-2's motion compensation algorithms just lapped up this type of footage and we ended up with clips that were as good as the original film.

Internet formats

For browse quality video files we needed a format that could be quickly downloaded over dial-up Internet connections, provided very good quality and could be universally used. Most of the candidate formats met two out of three requirements!

In the end, it was a choice between MPEG-1, QuickTime 5, or Windows Media 7. MPEG-1 had the advantage that just about every computer had a suitable player, but the quality at low bit rates was not acceptable. QuickTime 5 was outdated – we had been waiting for QuickTime 6

for months but still had no guarantee that it would be available in time. Windows Media 7 gave us better results than the other two at our target bit rates of 128Kbit/s and 512Kbit/s and would at least be available on every Windows PC.

Thumbnails

Of course we chose JPEG for our stills format – any computer that couldn't display JPEG would have problems browsing the Internet!

Over the years, we have found that users like thumbnails for browsing video. It is usually much quicker to scan a screen full of thumbnails to get an impression of the content of a clip than to play the whole clip through. Thumbnails can be captured at scene changes but it is more useful to use a fixed interval. When you use scene change detection you still have to allow for very fast scene changes so you have to set a minimum interval between thumbnails and you have to allow for very long scenes where you may have to allow a maximum interval between thumbnails. If you have a simple fixed interval, then the users can easily judge the length of the clips or scenes.

Constructing the archive

There was only going to be one operator so the system had to be efficient and as automated as possible. All of the server software is controlled across the network, so the operator can manage the system from a single workstation. The video players need to be accessible, but all of the remaining equipment can be kept out of the way.

We designed a system with two video players, which meant that two tapes could be encoded simultaneously. The operator selected tapes from the library and using the Imagen archive management system, she created jobs that were queued up by the video encoder systems.

After the correct tapes were loaded into the players the encoding queue was started and the system could be left to create and distribute all of the digital video assets.

Automated workflow

When designing a large-scale archive ingest system, it is very important to reduce the amount of keying and repetitive jobs to an absolute minimum. If you already have video tape time codes in a computer readable form, it is much better to import those than to re-enter the time codes all over again.

The time codes for the British Pathe tapes were entered into the database when the descriptions were updated so no keying was necessary to start an encoding run. After the operator started an encoding run the MPEG-2 video were created and stored. The transcoders automatically picked up the new files and created the browse video assets and the stills.

The transcoders were able to convert MPEG-2 video into Windows Media video at about 3 times real time. In other words, it would take 3 minutes to convert 1-minute's worth of MPEG-2 video. With current processors, this conversion is now faster than real time. As we were creating two different versions of browse video from each MPEG-2 asset we ended up with three transcoders to service the output from the two encoder systems.

The encoders were only in use during office hours but the transcoders were able to run 24 hours a day, so they would catch up with the encoders overnight. In this way, a single operator produced an average of 12 hours per day of completed footage. The system can be scaled up or down to suit the type of video assets that are required.

Automated distribution

Once a job was complete the Imagen archive system had all of the metadata and digital assets for a tape stored online. The database could be searched and video accessed immediately. The next step was to update the public archive on the Internet.

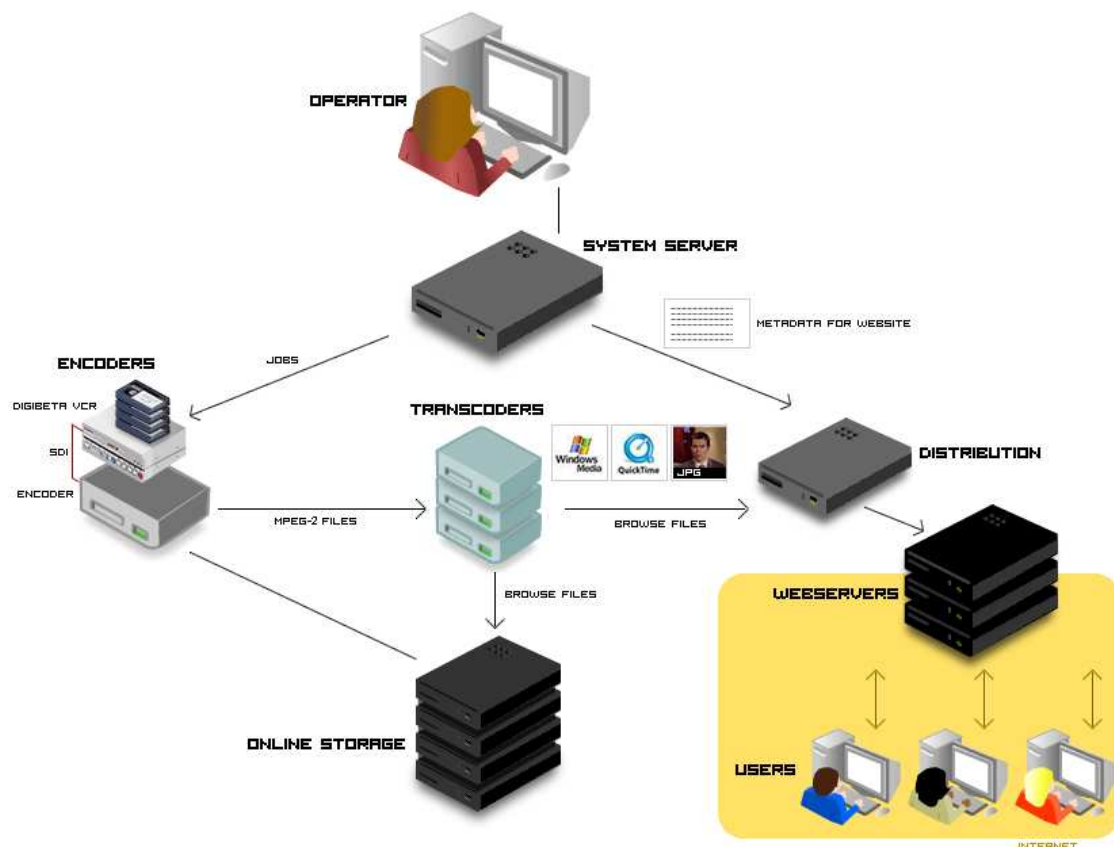
All of the ingest, management and storage servers were based at British Pathe's office in Camden. There was a fast leased line connection to the Internet, but we anticipated that it would be saturated if we tried to host the website in Camden as well.

Instead, we set up the web servers and database servers at an Internet data centre. As the tapes were encoded the Imagen system automatically uploaded metadata and assets to the website.

The website

When the British Pathe website was publicly launched in late November 2002, about half of the 90000 video clips had been encoded and uploaded to the site. We had quietly launched the site to professional researchers a couple of months earlier to see how well it worked and how well it was received.

Some changes were necessary to the website interface, but we were very encouraged by the enthusiasm of these early users. We began to worry that our estimates of the number of users may be well short of the mark! To provide for this unanticipated load and also to give the system some redundancy we doubled up all of the servers.



This diagram shows the flow of data and video through the ingest system to the website

Fingers crossed, but feeling comfortable with the system's capacity the site was launched in a blaze of TV, radio and press coverage. The first few hours saw the system reach its capacity and for many users the website slowed to a crawl. However, we served up 1.2 million pages to over 60 thousand individual users per day for that first week.

Gradually, the torrent of users died away to a more manageable level, but every so often there would be another peak. Most of the time this could be attributed to publicity somewhere in the world – a double page spread in the New York Times for example!

After a couple of months it settled down into a pattern. There are about 50 thousand visits to the website every month. About 15% of those are repeat visitors and the traffic is spread fairly evenly through the day, suggesting that much of it is international.

Version 2

We have made a series of upgrades to the website since the launch. Once the encoding was completed in April 2003 we then created an automated job that went back through the entire archive and captured new still images at one second intervals. These were full-frame, full-resolution images as well as new thumbnails. The website was modified so that users can browse the archive using thumbnails at any interval they like. We also launched the stills archive which, with over 12 million stills, must be the largest online stills archive in the world!

Version 3

By this time, QuickTime 6 was available, so we wrote a transcoder for QuickTime 6 and set up another automated task to generate new QuickTime 6 assets from the original MPEG-2. This more up-to-date format allows us to embed metadata into the video stream and it has been created at a higher bit rate for a playback quality that exceeds VHS. After one or two false starts, this task was completed and the assets added to the website earlier this year.

There are now over 90 thousand clips available in Windows Media 7 or QuickTime 6 formats with over 12 million thumbnails and still frames.

Version 4?

A crucial decision, taken early on, was to store all of the high resolution MPEG-2 assets online using custom built network attached storage (NAS) servers. This means that batch re-processing of the entire archive is possible with almost no operator involvement. We created the stills and QuickTime assets in this way and apart from an occasional check on progress the re-processing carried on unattended, 24/7.

We are considering updating the Windows Media 7 files to Windows Media 9, which should improve the quality of the video as well allowing us to embed metadata into the video files.

Another option would be to give paying users the ability to edit together sections of different clips and then request them in a video format and bit rate of their own choosing. The orders can be processed automatically by the system – including the e-commerce aspect – and the results emailed anywhere in the world!

Conclusions

Creating a digital version of the British Pathe archive was an exciting project involving a large team of people. We have been involved in larger projects in terms of the amounts of video or racks of equipment, but this has been the largest Internet based system we have developed. There have been well over 3 million unique visitors to the website and nearly half a million downloads from the site!

The key to successfully creating the digital assets was to make the whole system as automated as possible. Network based software and re-using metadata wherever possible helped here, but you still need someone to load tapes into the video players!

The flexibility of having the MPEG-2 assets online at all times helped with re-processing and gives us the option to add more and more features that give the end-user the results that they want. Having over 10 terabytes of data easily accessible over a Gigabit network has made our lives easier more times than I can remember. In fact, 10 terabytes doesn't seem that much nowadays and I'm sure it won't be long before we are dealing with hundreds of terabytes!

Links:

<http://www.itnarchive.com>

<http://www.britishpathe.com>

<http://www.cambridgeimaging.co.uk>