CAPTURING AND REPLAYING STREAMING MEDIA IN A WEB ARCHIVE – A BRITISH LIBRARY CASE STUDY

Helen Hockx-Yu **Lewis Crawford** **Roger Coram** **Stephen Johnson**

The British Library
96 Euston Road
London NW1 2DB

ABSTRACT

A prerequisite for digital preservation is to be able to capture and retain the content which is considered worth preserving. This has been a significant challenge for web archiving, especially for websites with embedded streaming media content, which cannot be copied via a simple HTTP request to a URL. This paper describes the approach taken by the British Library in capturing and replaying streaming media in a web archive. A working system is now in place which will lead to the development of more generic tools and workflows, contributing to addressing a common challenge for the web archiving community. The British Library recently archived a large scale public arts project website, http://www.oneandother.co.uk, which contains 2,400 hours of flash videos, streamed over Real Time Messaging Protocol (RTMP). The case study also presents an overview of the non-technical issues relevant to archiving this high-profile website.

1. INTRODUCTION

The web has become an increasingly important information resource for research and learning. However, the web is also ephemeral; websites disappear regularly. If not archived for long-term preservation, valuable web resources could be lost forever.

National libraries and archives around the world have been archiving the web in the 1990s. The Legal Deposit Framework of many countries now also includes the free web, with the national libraries carrying out periodical crawls of the respective national domains to capture and preserve a historical record of the web. A similar legislative framework exists in the UK but is yet to come into effect.

The importance of preserving web resources has been illustrated by the establishment and ongoing activities of the International Internet Preservation Consortium (IIPC), which was initiated in 2003 and currently has 38 member organisations across four continents. IIPC fosters the development and use of common tools, best practices and standards. Being brought together by common challenges, many national libraries and archives are active members of the IIPC, including the British Library.

1.1. Web Archiving at the British Library

With permissions from rights holders, the British Library has been selectively archiving UK websites since 2004. The Library has established an ongoing Web Archiving Programme to collect, make accessible and preserve web resources of scholarly and cultural importance from the UK domain. Archived websites to date are made available through the UK Web Archive, along with additional material archived by the National Library of Wales, the Joint Information Systems Committee, the Wellcome Library. The National Library of Scotland and the National Archives have previously contributed to the Archive.

The UK Web Archive contains regular snapshots of over 8,000 websites and offers rich search functionalities including full-text, title and URL search. The archive in addition can be browsed by Title, by Subject and by Special Collection. The UK Web Archive was formally launched in February 2010, raising awareness of the need for web archiving, which has generated a great level of interest from the press as well as the general public.

Web Curator Tool (WCT), a tool developed by the British Library in collaboration with the National Library of New Zealand, is used to manage our selective archiving processes. WCT embeds the commonly used open source crawler software Heritrix, and has added functionalities to manage workflow. The Open Source Wayback Machine (OSWM) is utilised to render and provide access to archived websites.

In anticipation of the implementation of Legal Deposit for UK online publications, the British Library is also exploring the technical and curatorial challenges of archiving in future a much larger proportion of the UK domain, through periodical domain harvests.

1.2. The One and Other Project

The 4th plinth on Trafalgar Square in London, originally intended for an equestrian statue, has been empty for many years. This is now the location for specially commissioned art works. Between 6th July and 14th October 2009, the famous British artist Antony Gormley undertook a large scale public arts project, during which 2,400 participants occupied the 4th plinth for an hour each, doing whatever they chose to do. The project was intended to create a living portrait of the UK, providing an open space of possibility.

All participants, or *plinthers*, were filmed and the videos were brought together on the project's website: http://www.oneandother.co.uk. The websites received over 7 million visits during the project.

When the project ended in October 2009, the British Library was approached to archive the website. It was a matter of urgency as the project funding would only last to maintain and keep the website live for a limited period of time beyond the project, till end of December 2009 initially, and then extended to March 2010. This time restriction has played a significant role in some of our technical choices.

2. PROGRESSIVE DOWLOAD VERSUS STREAMING MEDIA

Broadly speaking there are two ways to deliver digital media over the Internet between a server and a media player (used locally by end users): **progressive download** and **streaming media**. The former is also referred to as **HTTP download** because media files are typically transferred from the server to a client using the HTTP protocol. In addition, the media files are downloaded physically onto the end users' device, buffered and stored in a temporary folder for the local media player to use for replay. With streaming, data packets are constantly transferred and replayed to the end users, at no time leaving locally a copy of the entire file, as is the case with **progressive download**. There are protocols, such as the Real Time Streaming Protocol (RTSP) and the Real Time Messaging Protocol (RTMP), which are specifically designed to support streaming media.

Because of the potential risk of piracy related to progressive download, many content owners choose to publish high-value multimedia data using streaming based solutions.

The collective term *rich media* is used in this paper to refer to **progressive download** as well as **streaming media.**

For the purpose of web archiving, web crawlers are commonly used to capture snapshots of websites. It generally starts from a list of URLs (*seeds*), visiting and downloading them, before identifying all the hyperlinks within the visited pages and recursively visiting and downloading these too.

Capturing multimedia content can be just a matter of determining URLs. If the content can be served by requesting it, as web pages, then the crawler will be able to download a copy of the file via a simple HTTP request, by going to the right URL. However, parsing arbitrary URLs is not always a simple task as the URL syntax can be stretched to address almost any type of network resource and URLs can be generated dynamically. Overly complex URL structures include numerous variables, marked by ampersands, equals signs, session or user IDs as well as referral tracking codes. In some cases, multimedia files are served or initiated by embedded web applications which retrieve data from the server in the background, without explicitly locating the files in the HTML.

When streaming is not via HTTP, but proprietary protocols such as RTMP developed by Adobe Systems, it is even more difficult to capture and replay the multimedia content as this requires an understanding of the implementation of the particular protocol.

3. ARCHIVING RICH MEDIA

A prerequisite for digital preservation is to be able to capture and retain the content which is considered worth preserving. This has been a significant challenge for web archiving, especially for websites with embedded streaming media content, which often cannot be copied via a simple HTTP request to a URL.

The tools currently used by the British Library, and many other national libraries and archives, do not yet have the capability of capturing and playing back streaming media content embedded in archived websites. Heritrix, the crawler software, can only capture data delivered over HTTP and/or FTP. In addition, the OSWM does not have any streaming capability.

Many organisations engaged with web archiving have long recognised the need for a solution to dealing with rich media. The Internet Archive, the Institut National de l'Audiovisuel (INA) and the European Archive for example have been actively carrying out research and developing projects to address the problem. The effort focuses on adding capabilities to crawlers for them to be able to interpret complex code and extract URLs for media files so that these can be captured by the crawlers through HTTP requests. A problem with this is that the exercise of URL parsing needs to be frequently repeated as sites such as YouTube constantly change the way of publishing videos to prevent direct downloads. In addition, the replay aspects of the captured media have pretty much been left to the capability of the browsers, or occasionally solutions developed specifically for individual media player applications.

Adding capability of capturing and replaying rich media in web archives is a key area of work for the

IIPC.

4. CAPTURING AND REPLYING PLINTHER VIDEOS

The One and Other website contains 2,400 hours of video in .flv format, approximately 1TB, streamed directly over RTMP. Initial test crawls of the sites using the Web Curator Tool (essentially Heritrix) only brought back static HTML pages without the videos, which the artist and curator considered as significant and essential components of the project and the website.

As previously mentioned, the One and Other website had a planned take-down date of end December 2009, which only allowed us a couple of months to find a solution to capture the website (the take-down date of the website was later extended to end March 2010). There was additional pressure to develop an access solution too, as the plan was to invite the artist Antony Gormley to speak at the formal launch of the UK Web Archive three months later to maximise the impact of the event. The tight timescale meant that our goal was to find a working solution for an immediate problem, rather than setting out to develop a generic technical solution for the long term within that phase of the project.

4.1. Capture

Essentially a combination of a browser and a streaming media recorder was used to initiate and capture the video streams from the One and Other website. The choice of software was largely determined by its functionality being adaptable to the project at hand. Apart from test captures to check reliability and quality, the main criteria used to select a streaming media recorder included the ability to capture media steamed over RTMP, to schedule captures and the ability to import a schedule so that a degree of automation was possible. It was equally important that the chosen sofware's method of naming the captured files should allow easy identification of the video along with the web page it was captured from.

Based on the above criteria, we chose Jaksta as our media recorder. Jaksta can detect videos and music streamed over RTMP, using port 1935, and capture the TCP/IP packets as they are sent to the embedded flash player in the browser. Although not allowing imports, Jaksta uses a sqlLite database which gave us the opportunity to automate some parts of the scheduling.

Prior to the actual captures, a Unix shell script was used to identify pages containing video streams, which output a list of URLs of pages containing videos. Four virtual machine instances, all configured with Jaksta for capturing video and SqlLite2009 for scheduling, each based on the schedule launched Internet Explorer instances at three different URLs at a time, to initiate the video streams. It was then was a matter of letting Jasksta do the job of capturing the videos. The scheduling, also inserted using a Unix shell script, was set at 90 minutes intervals. We knew in advance that each video was approximately an hour long, so this was the metric used as a static variable to create scheduling.

Once completed the captured video was saved onto local disk. Jaksta uses the URL query as a naming convention when possible, which suited us and allowed easy identification of the link between the video file and the web page which it was embedded in and captured from.

The method described above was used to capture the video content from the One and Other website. File size was an immediate attribute used to monitor the capturing process because all the videos are of similar length, and significant variance in file size was an indication of error. File size in itself, however, cannot determine definitively if the full was captured. A shortcoming of Jaksta was that that it did not recognise or report when the full video was not captured. The videos were also spot-checked by viewing them, validated using the FLVCheck tool (by Adobe), and where required and possible, repaired using FFmpeg.

A second attempt was made to re-capture a portion of the videos which appeared shorter in length but this made no difference, which made us suspect that the error may be inherent to the video files themselves. This was confirmed when SkyArts, who sponsored the One and Other Project, later provided us with the original video files on a disk which unfortunately contained the same errors. The errors were believed to be caused by the videos in question not being recorded as one file, resulting in a mismatch between the metadata layer and the content layer. As a result, these videos have been curtailed in the web archive and cannot replay to the full length. SkyArts is currently looking to fix these videos.

4.2. Replay

Capturing the videos only completes half of the job. In order to provide access to the archive version of the One and Other website, we also needed a solution to play back the videos, as part of the end use interface of the UK Web Archive.

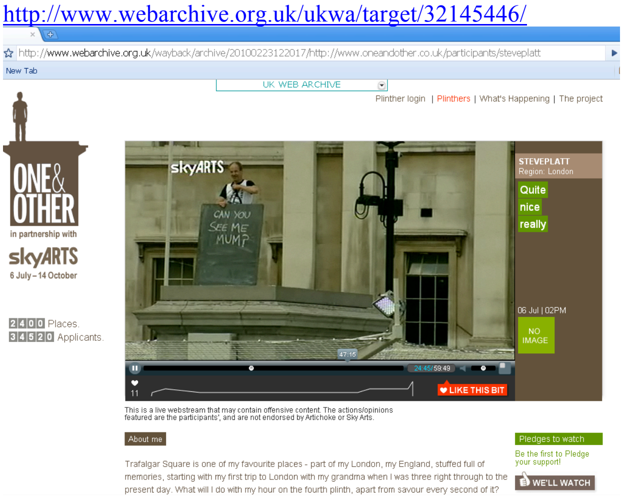
When granting a licence to the British Library, SkyArts explicitly required that the video content may only be streamed to the archive users, having in place the copy protection equivalent to that applied to the original website. This requirement eliminated the possibility of implementing any solution based on progressive download.

Two open-source software tools have been used to stream and replay the videos. Red5, a Java media server, was chosen as our streaming mechanism. In addition to the base streaming server, Red5 requires an application to access and serve the media. Several demo. applications can be installed by default and the 'oflaDemo' application, designed simply to serve from a flat file system, was adequate to serve this purpose. For the client side, Flowplayer has been selected as the video player, used to play back the streamed Flash videos.

In order to replace the original flash objects and to reference the local videos, a modification has been made to our Wayback timeline implementation, which is a banner inside rendered HTML pages inserted by the OSWM, allowing users to navigate between individual archived versions of the current page. A few lines of JavaScript has been added to firstly, if not already defined, reference the flowplayer() function by calling the flowplayer-*.min.js file. The window.onload function has then been amended to load a Javascript file with the same name as the original domain from the Flowplayer location (i.e. http://...wayback/*/http://www.oneandother.co.uk/ will load www.oneandother.co.uk.js). This contains a single function - streamVideo() - which does two things:

1. Replace any existing Flash elements with an object of equal dimensions.
2. Call the now-defined flowplayer() function, passing in (among other things) the name of the video file, derived from the plinther's name, and the name of the above, new object.

The One and Other website is no longer live on the web since 31 March 2010. The domain name oneandother.co.uk now redirects directly to the archival version in the UK Web Archive: http://www.webarchive.org.uk/ukwa/target/32145446/



This is a live webstream that may contain offensive content. The actions/opinions featured are the participants', and are not endorsed by Artichoke or Sky Arts.

Trafalgar Square is one of my favourite places - part of my London, my England, stuffed full of memories, starting with my first trip to London with my grandma when I was three right through to the present day. What will I do with my hour on the fourth plinth, apart from savour every second of it?

Figure 1: A screenshot of the archived One and Other page

5. NOT JUST TECHNICAL CHALLENGES

One and Other was the most talked-about arts project in the UK in 2009 and it caught the media's attention from the very beginning. Archiving such a high profile website involving multiple stakeholders meant there were also legal, curatorial and communication challenges which required the team's extensive attention.

Even before any technical solution was attempted or experimented, the media had already reported that the British Library would archive the One and Other website and preserve it in perpetuity. Publicity, when well managed, can however help the cause of web archiving. Antony Gormley was invited to the formal launch of the UK Web Archive at the end of February 2010, who spoke positively about working with the British Library. This has helped generate positive publicity and illustrate the importance of web archiving.

The One and Other project had many stakeholders, including the artist, the sponsor, the producer, the technology provider and the 2,400 participants. Intensive interaction with the stakeholders took place to coordinate and communicate the archiving process. It is not always possible to balance the interests and expectations of all the stakeholders. There is generally an expectation for an archived website to behave exactly the same as the live website. For some plinthers, it is difficult to appreciate the concept of an archival website and understand why message boards and discussion forums no longer work.

The British Library has a standard licence which website owners sign to grant us permissions to harvest, provide public access to and preserve their websites in the web archive. A customised licence had to be developed specifically for the One and Other web site, introducing additional terms and conditions and specifying in detail the involved parties' obligations.

There had also been a couple of occasions in which a plinther or a third party had requested that certain pages of the website to be taken down. Delay in taking actions or non-compliance could have potentially resulted in legal proceedings. These occurred when the live website still existed and were dealt with by the sponsor and the artist directly. Although all that the Library was required to do was to recapture a "cleaned" version of the website, such situations, however, do raise interesting and significant legal and curatorial questions. When the live website no longer exists, the Library would be seen as republishing the website by providing public access to its archival version. This in itself will transfer certain legal risks from the original publisher to the Library.

6. NEXT STEPS AND CONCLUSION

The British Library has invested considerable resources in archiving the One and Other website and successfully implemented a solution within the required, extremely tight timescale. The addition of the One and Other website to the UK Web Archive has helped raised the profile and awareness of web archiving. Our approach to capturing and replaying streaming media seems to be the only way at the moment to capture the video streams as they are not available via standard (HTTP) protocols. It is the first practical streaming media implementation within the international web archiving community and provided us with valuable hands-on experience which will lead to more generic solutions.

A main issue with the solution described in the case study is that it sat outside the operational workflow. The video files for example were not stored as ARC files with the rest of the web archive but streaming from a separate video server in the native format. This introduces data management complexity and potential digital preservation risks.

It is desirable to build streaming media capability into the current web archiving tools commonly used by the national libraries and web archives. Alternatively we could extend the open source tools we used for them to interpret archived websites. We are pleased to report that in the subsequent months following the project, we carried out further development work on Red5 which can now stream from non-compressed WARC files.

The LiWA project, funded by the European Commission, has recently released a rich media capture plug-in for Heritrix which aims to enhance its capturing capabilities to include HTTP downloads as well as streaming media. It is still an experimental version of the software but nevertheless shows potential of adding rich media capturing capability to Heritrix.

The advent of HTML5 in addition seems to offer the most effective solution to replaying HTTP media in a web archive. The introduction of the <video> tag explicitly marks up the content which means video can be streamed over HTTP and replayed directly by the browser without the necessity of additional applications.

The recent technological developments are encouraging and it is not unrealistic to expect in the foreseeable future a solution to capturing and replying rich media in web archives. In parallel, the web archiving community perhaps should also consider approaching major rich media publishers (e.g. YouTube) to achieve collaborative arrangement so that more focused solutions can be developed at the API level.

7. REFERENCES

[1] The British Library Web Archiving Programme: http://www.bl.uk/aboutus/stratpolprog/digi/webarch/

[2] European Archive : http://www.europarchive.org/

[3] FFmpeg: http://www.ffmpeg.org/

[4] FlowPlayer: http://flowplayer.org/

[5] FlVCheck Tool: http://www.adobe.com/livedocs/flashmediaserver/3.0/docs/help.html?content=06_admintasks_11.html

[6] Heritrix: http://crawler.archive.org/

[7] International Intenert Preservation Consortium (IIPC): http://netpreserve.org/about/index.php

[8] Internet Archive: http://www.archive.org/

[9] Institut National de l'Audiovisuel (INA) : http://www.ina.fr/

[10] Jaksta: http://www.jaksta.com/

[11] LiWA project: http://www.liwa-project.eu/

[12] LiWA rich media capture module for Heritrix: http://code.google.com/p/liwa-technologies/wiki/RichMediaCapture

[13] Open Source Wayback Machine: http://archive-access.sourceforge.net/projects/wayback/

[14] Red5: http://osflash.org/red5

[15] The UK Web Archive: http://www.webarchive.org.uk/.

[16] The Web Curator Tool: http://webcurator.sourceforge.net/