Preserving the Web Archive for Future Generations

Practical Experiments with Emulation and Migration Technologies

Matt Holden

Institute national de l'audiovisuel (Ina), mholden@ina.fr

Abstract

Ina shares the responsibility of the web legal repository in France with the National Library of France (BnF)[1] *and has been archiving regularly since the beginning of February 2009. In this paper we look at how we can use emulation/migration technologies to help preserve an obsolete format and as a result, we hope to be able to shed light on which parts of the web archive may be most at risk and whether this can help shape and prioritize our long-term preservation strategy.*

Author

Matt Holden, MSci, MSc, ARCS studied for a bachelor's degree in Physics at Imperial College, London before taking a master's degree in Telecommunications at University College London (UCL). This led to an Operations IT role working at Nortel Networks UK in a group developing solutions for the mobile phone GSM-HLR network. Late in 2006, Matt became Data Centre Manager at CMC Markets UK plc, which culminated in a multi-million pound construction project to create a new data centre in East London. In early 2009, he joined Ina with the mission of running IT Operations for the web archiving (DLWeb) team.

1. Introduction

The emulation vs migration debate is one which has been discussed at great length in many papers and journals too numerous to catalogue. The intention of this paper is to look at practical experiments that can be conducted during the operation of a working web archive.

As Ina is a member of the International Internet Preservation Consortium (IIPC), we were very interested in building upon the work already undertaken by the National Library of Australia (Long, 2009) as part of a project for the IIPC's Preservation Working Group (PWG). We wanted to revisit the work to see how it could be applied to the operation of our own web archive as well as investigate if the tools had changed.

We were also interested in using the process of Transparent Format Migration (Rosenthal, Lipkis, Robertson, & Morabito, 2005) conducted at the LOCKSS (Lots of Copies Keep Stuff Safe) web archive in which image formats were converted automatically when pages were loaded.

[1] Title IV of French Law n°2006-961 dated 1st August 2006 entitled "Loi DADVSI"

Finally, we were curious to find out whether there really were file formats in the oldest part of our archive that were in danger of becoming obsolete.

1.1 The History of Legal Deposit at Ina

Since its creation in 1974 by law, Ina has been in charge of collecting, preserving and making available French audiovisual collections. Initially organized to meet professional needs for a public broadcast archive facility, it quickly grew into a national repository for the French audiovisual heritage. In 1992 due to its existing obligations, Ina was designated by law as being responsible for the Legal Deposit of radio and television. Today, over twenty years later, the institute is considered to be the world's largest broadcast archive, holding over 4-6 million hours of television and radio recordings, dating back to the dawn of the medium. Annually, we see an increase of a million hours of programs from 100 television channels and 20 radio stations which are digitally recorded around the clock.

Following the fast pace of publishing technologies, French Legal Deposit Law was then amended to include the Web, splitting responsibility between the French national library (BnF) and Ina, thus ensuring coherence and continuity of their respective collections. Ina was thus designated as national repository for audiovisual related Web sites – with a broad remit – as well as on-demand audiovisual media services available from Web platforms.

1.2 The History of Web archiving at Ina

Web archiving duly commenced in February 2009 with 3600 seed websites covered in the initial collection process. In late 2010 we worked in partnership with the Internet Archive to recover websites pertaining to our collection from their archive from 1996-2009 (18Tb of data containing 524 million URLs). As of late July 2012, the whole collection contains 15 billion URLs consisting of nearly 10,000 seed websites represented by over 150Tb of compressed data.

1.3 Ina's Web Archive: technologies and statistics

- No URLs are changed during the consultation of the archive
- All access to the archive is controlled via a proxy which matches urls to archived content. We currently use Firefox v3.5 with a number of plugins to facilitate archive access, such as displaying the date of capture and navigation between different archived versions of the same page.
- Ina uses its own Digital Archive File Format (DAFF)
- All content is de-duplicated – there are no multiple copies of objects in the archive, but the objects themselves may be referenced multiple times[2].
- When we talk about the overall size of the archive we include multiply referenced objects as this indicates the enormous savings due to the de-duplication process (since sites have been crawled with a high frequency). However from a preservation perspective, it is useful to consider the number of unique objects in the archive (
- Figure 1).

[2] E.g. a 2Mb audio file is captured on a web page five consecutive times whilst a text file of 0.1Mb referencing it changes each time. We have five versions of the text but one single audio file referenced five times, so technically our archive size indicates 10.5Mb as opposed to the 2.5Mb of actual data linked to these versions.

•

Object Type	Unique objects (millons)	Unique objects size (Tb)	Multiply referenced objects (millons)	Multiply referenced objects size (Tb)
Image	145	3.1	7454	114
Text	1759	82	3390	99
Audio	2.6	25	82	649
Video	1.4	22	29	527
Other	288	4.8	1785	65
Total	**2196**	**137**	**12740**	**1364**

Figure 1. Ina web archive statistics (June 2012)

Given the importance of images and sound at Ina, we have concentrated on rich multimedia websites and this is very much reflected in our archived content, with nearly 50Tb of compressed audiovisual material counting for approximately a third of the size of the archive.

1.4 Long-term bit preservation (using short-term migrations)

The long-term bit preservation of the archived files falls into line with the typical working processes of many other institutions, using short to medium term migrations (around 20 year strategies). Two copies of the archive are stored on differing generations of disk storage and 2 offline backup copies will be stored on tape. This is based on migrating the archived files onto a new disk storage media every 3-5 years and a new tape based media every 4-5 years.

1.4.1 Disk Storage

Unlike tape storage, disk storage typically has much higher associated costs regarding power and cooling. We also must take into account the reliability and failure rates of disk technology, which limits the typical use of these systems to between 3 to 5 years. To facilitate migration and cost savings we look to double to storage capacity on each migration (e.g. from 1.5 to 3Tb disks)

1.4.2 LTO Tape Storage (Linear Tape-Open)

LTO is a magnetic tape data storage technology, developed in the late 1990s under an open standards initiative as opposed to the proprietary magnetic tape formats that were available at the time. The capacity will approximately double with each new generation (from LTO-1 to LTO-8) but retain the same physical size, thus facilitating migration strategies and allowing standardization in physical storage depots. However, the LTO specification only mandates the that each version of LTO is read compatible with the 2 preceding versions (eg, LTO-4 drives can read LTO-2 tapes) Therefore, to reliably migrate data we

must transfer the data to the new format when it becomes widely available and is economically viable, which is approximately every 4 to 5 years.

1.4.3 Checksums

After the creation of the DAFF archive file, its content is verified using various tools. If the content of the file is found to be valid, we take a checksum (SHA) of the file and store this in a separate database. Checksums are also stored with their associated files when they are stored on magnetic tape. In this way we can periodically verify that the contents of the file are valid and if necessary replace a corrupted copy from a backup.

2. Method

The methodology was as follows:

1. Extract all audio, video and image files (identified by their MIME type) from Ina's archive during the period 1996/7.
2. Use file identification software to verify the content.
3. Assess whether there are any unknown or obsolescent file types.
4. Attempt to read these file types through emulation or migration.
5. Attempt to play these file types by modifying the existing archive interface.

2.1 Small Sample Testing

In order to find the best candidates for emulation/migration we decided to look at the older part of the web archive – content collected in 1996/7 which contained approximately 4-5Gb of data. Although this is a very small part of the archive's overall content, we thought it would produce a more manageable dataset. Again, given Ina's interest in audiovisual material we focused on image, audio and video content.

2.2 Hardware used

Desktop: Dell Optiplex 760, 2Gb RAM, Intel Core 2 Duo E8400 3GHz - Windows 7

Data Processing: Dell R710, 16Gb RAM, 2 x Intel Xeon Quad-core X5560 2.8GHz - CentOS 5.5 64-bit + chrooted CentOS 5.5, 32-bit installation

Storage array: Transtec 2 x Intel Xeon Quad-core X5650 2.67GHz, 48Gb RAM, Adaptec 5805Z RAID controller , 2 x 28Tb RAID 6 virtual disks - CentOS 5.5 64-bit

2.3 Format Identifying Software

There are a number of format identifiers available, some well known to the preservation community such as PRONOM and JHOVE. However we decided against using JHOVE in this instance as there is limited support for audio and video file formats.

2.3.1 Format Identification for Digital Objects (FIDO) v1.1.0 (DROID signature file v60)

https://github.com/openplanets/fido

Description:
Developed as part of the Open Planets Foundation, FIDO is a Python command-line tool to identify the file formats of digital objects using PRONOM's Digital Record Object Identification's (DROID) signature files.

Comments:
Performed the tests very quickly and was capable of exporting the information to csv format for easy conversion into excel. Backed by the powerful PRONOM library which is capable of using signatures to positively match filetypes (not just by filename extensions). Simple to use and easy to integrate into existing workflows.
Positive Identification Criteria:
We considered a positive match to be where a file format signature was used to identify the object.

2.3.2 Apache Tika v1.1

http://tika.apache.org/

Description:
The Apache Tika™ toolkit detects and extracts metadata and structured text content from various documents using existing parser libraries.

Comments:
Tika is also easy to setup and comes with an extensive set of libraries for integration into the user's environment. However for these tests we simply used the included jar file to launch the application.

Positive Identification Criteria:
We considered a positive match where tika was able to extract significant metadata (i.e. other than the name, size, or content type) from the file.

3. Results

3.1 Extraction Process

The Data processing server ran 4 parallel processes to scan through 270Gb of metadata to extract objects matching the MIME type video/*, audio/* or image/* and the date of capture. From the resulting extraction, we used the content-id checksum in the metadata to link through to the corresponding DAFF data file which could then be downloaded from the archive servers and stored locally on the Data Processing Server.

The extraction took approximately 4 hours and produced the following fileset (
Figure 2)

	1996	1997
Files	4607	12206
Size (Gb)	1.4	2.7

Figure 2. 1996-7 Archive Sample

3.2 File type assessment

After the extraction we ran both file format identification tools on the extracted files (
Figure 3). We were surprised at the high success rate of FIDO, which identified the vast majority of formats, whereas Tika struggled with the less common audio and video formats.

Positive identification	1996	1997
FIDO	97.3%	98.3%
Tika	85.1%	87.4%

Figure 3. 1996-7 Positive identification

After using FIDO and Tika we were left to categorise the remaining files (
Figure 4) without signature based identification (or extensive metadata) by hand. Looking at the list, it became clear that the majority of these files formats are well known, well supported and still in wide use today - with the notable exception of the Vivo Active Video file format.

File Types	1996	1997
GIF 1987a/1989a	1340	3914
Audio Video Interleave (avi)	566	949
JPEG	1447	5169
WAV	580	599
TIFF	0	2
Quicktime (mov)	595	1030
AIFF	31	0
MPEG	0	1

AU	12	3
MIDI	4	15
Real Audio (ra/ram)	32	490
VivoActive (viv)	***0***	***34***
Total	**4607**	**12206**

Figure 4. 1996-7 Identified Filetypes

3.3 File Format – at risk: Vivo

Out of all the formats identified during this process, only one could be considered as being 'at risk' - the VivoActive VIVO video format. The format was one of the first used for video streaming in the 1990s before being rendered obsolete by other formats supported by Real Player, Quick Time and Windows Media Player. RealPlayer (Real Networks) acquired VivoActive and consequently the VIVO format in 1997.

3.4 Second Archive Sample (1996-2008)

Now that we had identified a particular format of interest, we decided to take a look as to whether other such files existed in the archive. Returning to the metadata we re-scanned files pertaining to a larger subset of the Archive between 1996-2008, this time slightly widening the search to include Vivo's MIME type (video/vnd) as well as any files matching the filename extension ".viv" (
Figure 5).

Year	**Vivo files**	**Year**	**Vivo files**
1996	0	2003	31
1997	209	2004	2
1998	159	2005	1
1999	0	2006	8
2000	17	2007	0
2001	31	2008	0
2002	38	**Total**	**496**

Figure 5. Archive 1996 - 2008 Vivo format files

We identified 496 files corresponding to these parameters of which 211 were unique (multiple references within the archive). As expected, the usage of this format declined sharply shortly after VivoActive was taken over in 1997.

3.5 Reading the Vivo format

Finding a player for the vivo format proved problematic – there are still links (RealNetworks, Inc) to the original software on the Realplayer site, but it appears that any development stopped since the acquisition of the vivo format in 1997 – the software is designed for Windows 95 and browsers (Netscape 4 / IE 3 or 4) of that period.

The next port of call was RealPlayer itself (current version 15.0.5.109), however the Vivo format files were no longer supported (Formats - RealNetworks, Inc).

In the end, we turned to opensource alternatives such as MPlayer (MPlayer) which supports the VIVO format versions 1.0, 2.0, I263 and other H.263(+) variants (using x86 DLLs).
MPlayer is also bundled with Mencoder which allows all readable formats to be converted to modern variants such as MPEG-4 H.264 and Flash Video.

3.6 Problems with supporting Vivo libraries

MPlayer includes support for a very large range of video formats by allowing the inclusion of numerous codecs in the form of libraries. With some of the older formats, including for Vivo, this means using the existing libraries as they were developed i.e. for windows in the 1990s in the form of 32-bit DLLs.

To use these DLL codecs it is necessary to install or emulate a 32-bit operating system, either through the use of virtual machines or via the use of 32-bit libraries . Given that our server environment is now completely 64-bit linux based, we decided to use a chrooted 32-bit installation of linux on an existing 64-bit machine.

The installation method is too complex to describe in detail in this paper, but essentially this allowed us to preserve all the 32-bit software and libraries necessary to execute Mplayer and Mencoder in this environment.

3.7 Ina's Media Player tool

In the development of the consultation interface for the archive (based on Firefox) we decided to adopt the use of an external media player via a toolbar button (Figure 6) to help play audio and video files in the archive. The media player works by identifying links within the page as well as being able to access metadata which allows it to link media files associated with the page and passes them to an external player (we are currently using the JWPlayer (v5.4) for its enhanced Flash and HTML5 capabilities).

In one example, embedded flash videos on a page are captured separately by data mining the links on the archived page and downloading the video content in a separate procedure. In Figure 7 we can see how the archived web page is missing its embedded video. In Figure 8 we can see how the video content has been 're-associated' with the page by using the media player plugin.

3.8 Integrating Mplayer/Mencoder with Media Player

Using the media player plugin and combining it with Mencoder we are able to convert 'on the fly' associated vivo files to a more compatible format. In this case we have chosen the Flash file (flv) format.

The vivo video format is identified through its MIME type (video/vnd.vivo) and converted automatically by Mencoder. The converted files are then stored in their new format and can be played directly (without the need for conversion) the next time the video is requested. In essence, we are performing 'on-demand' migration.

In Figure 9 we have a webpage with links to vivo files. Using the modified Media Player we can access the automatically converted content (Figure 10).

3.9 Conversion Compatibility Problems

It should be noted that we had some difficulty in choosing options with which to convert the files. There are a wide range of options available in terms of bit rates, size, quality synchronization, etc which changed the length and quality of the resulting video considerably. Further investigation is required on how to better normalize the process.

4. Conclusions

In the end we were able to use transparent format migration, underpinned by some emulation, to help us access an obsolete file format. This work was very much a test of the technology, able to be built on to the existing structure of the archive (Media Player) and we were able to see how well the process could be integrated. This has resulted in a practical tool which now enables users of the archive to access videos which had been effectively unreadable.

That said, the 'obsolescence' of the Vivo file format could be considered relatively insignificant given the very small percentage of files as compared to the size of the archive. However, given that a large percentage of these files were found on old archived pages from Ina's corporate website, they take on a much greater importance!

4.1 Further Work

This work opens up some interesting avenues to pursue with future projects. With the increasing use of data mining on our own archive, we can imagine conducting file format identification on the whole archive to begin to analyse trends in file format obsolescence. The tools seemed to be robust and useable on a much wider scale.

Even though MPlayer/Mencoder is a modern up to date piece of software, we are still relying on emulation in the form of codecs and a 32-bit environment upon which to play the videos. We need to look at ways of cataloguing and preserving such information so it can be used to help the emulation of software or to support renderers in the future.

All of the work presented in this paper will be fed back into the Preservation Working Group of the IIPC, where a number of projects are underway to help the web archiving community track format obsolescence as well as the tools to support them.

4.2 Looking Forward

Preserving a web page in the archive, as it was, will become increasing complex task as software develops to keep up with the myriad of formats and standards. Although we believe we will be able to preserve a lot of the text, images and videos we think there will inevitably be a loss with regards the style,

feel and interactivity of web pages over time. Already with the use of the Media Player we are, in essence, breaking up the flow of a web page to be able to extract the content.

We think it is best that we should concentrate on using current tools to extensively data mine as to enrich the archive with as much metadata as possible.
In this way we should look to leave as many tools as possible for future generations to re-interpret these digital artifacts.

References

Brown, R. H., & Davis-Brown, B. (1998). The making of memory: the politics of archives, libraries and museums in the construction of national consciousness. *History of Human Sciences , 11* (4), 17-32.

Carr, E. H. (1961). *What Is History.* New York: Random House.

Consultative Committee for Space Data Systems. (2012). *Reference Model for an Open Archival Information System (OAIS).* 650.0-M-2.

Cox, R. J. (1994). The Documentation Strategy and Archival Appraisal Principles: A Different Perspective. *Archivaria* , 11-36.

Diamond, E. (1994). The Archivist as Forensic Scientist -- Seeing Ouselves in a Different Way. *Archivaria* , 139-154.

Duranti, L., & Endicott-Popovsky, B. (2010). Digital Records Forensics: A New Science and Academic Program for Forensic Readiness. *Journal of Digital Forensics, Security and Law , 5* (2), 45-62.

Formats - RealNetworks, Inc . (n.d.). *Supported Media Formats*. Retrieved from http://cache-download.real.com/free/windows/mrkt/help/RealPlayer-15/en/Content/Media_Types.htm

Gardner, J. (2003). The Mark of Responsibility. *Oxford Journal of Legal Studies , 23* (2), 157-171.

Halbwachs, M. (1992). *On Collective Memory (originally published in 1941, edited and translated by Lewis A. Coser).* Chicago: University of Chicago Press.

Hedstrom, M. (2001). Exploring the Concept of Temporal Interoperability as a Framework for Digital Preservation. *Third DELOS Network of Excellence Workshop on Interoperatability and Mediation in Heterogeneous Digital Libraries.*

Josias, A. (2011). Toward an understanding of archives as a feature of collective memory. *Archival Science , 11* (1-2), 95-112.

Lee, C. A., & Tibbo, H. R. (2011). Where's the Archivist in Digital Curation? Exploring the Possibilities through a Matrix of Knowledge and Skills. *Archivaria , 72*, 123-168.

Liu, W. W. (2010). Identifying and Addressing Rogue Servers in Countering Internet Email Misuse. *IEEE/SADFE-2010: Proceedings of the 5th International Workshop on Systematic Approaches to Digital Forensic Engineering* (pp. 13-24). Oakland, CA: IEEE Computer Society.

Liu, W. W. (2011, July). Trust Management and Accountability for Internet Security. *PhD Thesis* . Tallahassee, FL: Florida State University.

Liu, W., Aggarwal, S., & Duan, Z. (2009). Incorporating Accountability into Internet Email. *SAC'09: Proceedings of the 24th ACM Symposium on Applied Computing.* Horolulu, HI: ACM.

Long, A. S. (2009, 12 10). *Long-Term Preservation of Web Archives – Experimenting with Emulation and Migration Methodologies.* Retrieved 04 01, 2012, from netpreserve.org: http://netpreserve.org/publications/NLA_2009_IIPC_Report.pdf

Misztal, B. (2003). *Theories of Social Remembering.* Maindenhead, UK: Open University Press.
MPlayer. (n.d.). Retrieved from http://www.mplayerhq.hu/

Nora, P. (1989). Between Memory and History: Les Lieux de Memoire. *Representations , 26* (Special Issue), 7-24.

Pruzan, P. (1998). From Control to Values-Based Management and Accountability. *Journal of Business Ethics , 17* (13), 1379-1394.

Real Networks. (n.d.). *RealNetworks, Inc. - Acquisition History*. Retrieved 07 05, 2012, from http://investor.realnetworks.com: http://investor.realnetworks.com/faq.cfm?faqid=2

RealNetworks, Inc. (n.d.). *Download the VivoActive Player*. Retrieved from http://egg.real.com/vivo-player/vivodl.html/

Rosenthal, D. S., Lipkis, T., Robertson, T. S., & Morabito, S. (2005, Jan 1). *Transparent Format Migration of Preserved Web Content.* Retrieved 07 05, 2012, from www.dlib.org: http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html

Sachs, A. (2006). Archives, truth and reconciliation. *Archivaria , 62*, 1-14.

Schwartz, J. M., & Cook, T. (2002). Archives, records, and power: the making of modern memory. *Archival Science , 2* (1), 1-19.

Sinclair, A. (1995). The Chameleon of Accountability: Forms and Discourses. *Accounting, Organizations and Society , 20* (2/3), 219-237.

Wallace, D. A. (2011). Introduction: memory ethics--or the presence of the past in the present. *Archival Science , 11*, 1-12.

Wallace, D. A., & Stuchell, L. (2011). Understanding the 9/11 Commission Archive: Control, Access, and the Politics of Manipulation. *Archival Science , 11* (1-2), 125-169.

Yakel, E. (2007). Digital Curation. *OCLC Systems & Services , 23* (4), 335-340.

Yakel, E., Conway, P., Hedstrom, M., & Wallace, D. (2011). Digital Curation for Digital Natives. *Journal of Education for Library and Information Science , 55* (1), 23-31.

Illustrations

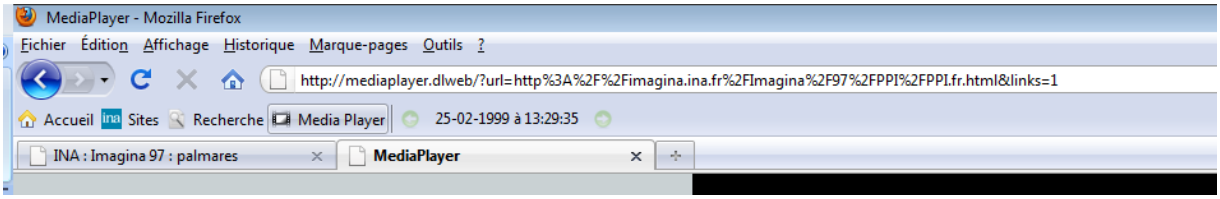

Figure 6. Firefox Media Player Toolbar Button

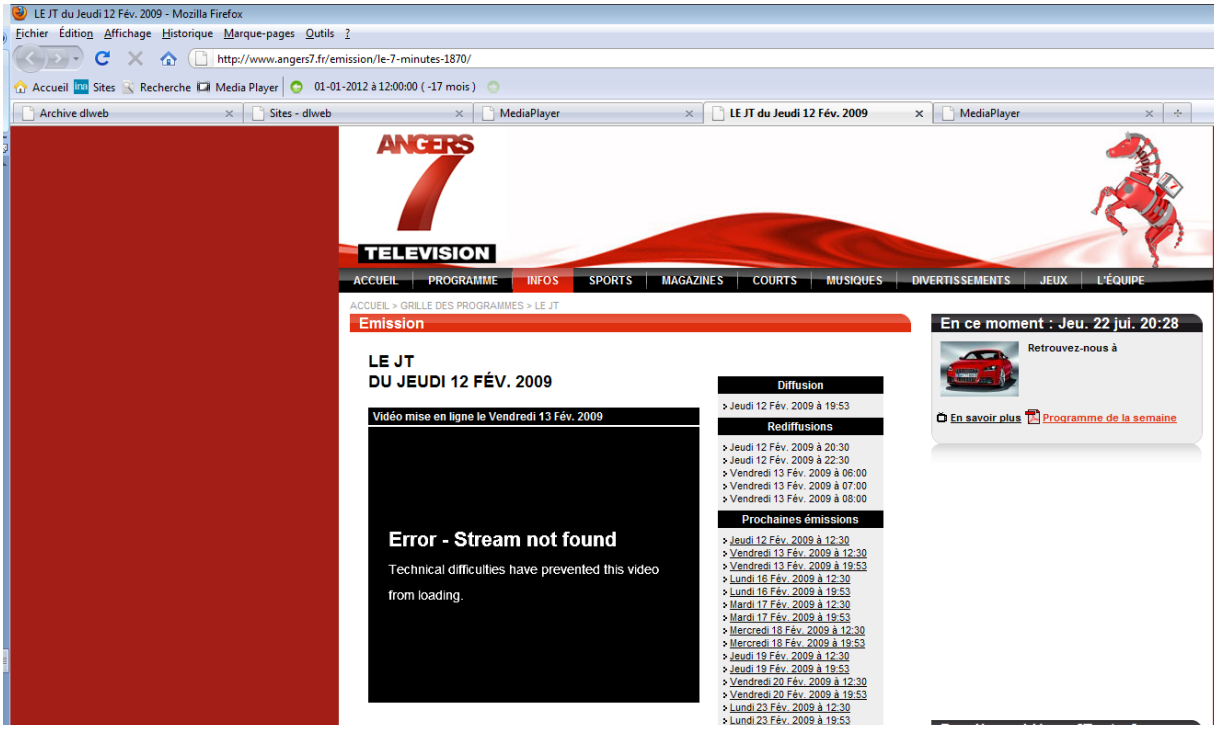

Figure 7. Archived Angers 7 page missing embedded video content

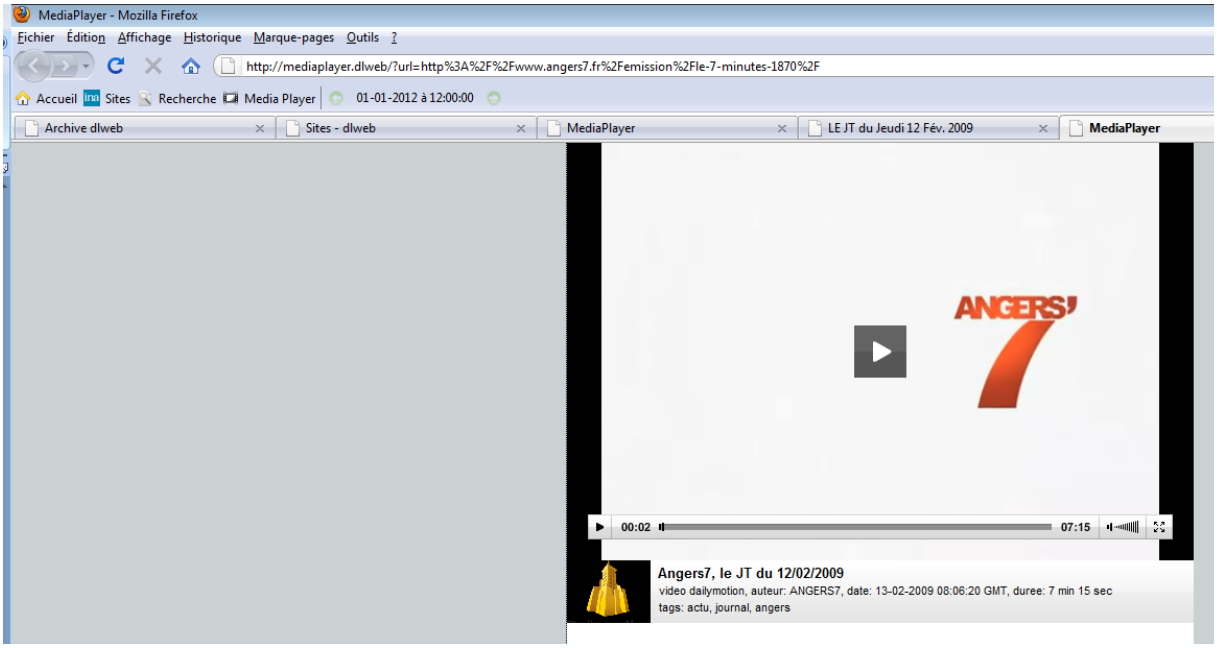

Figure 8. Archived Angers 7 page recombined video content accessible through Media Player

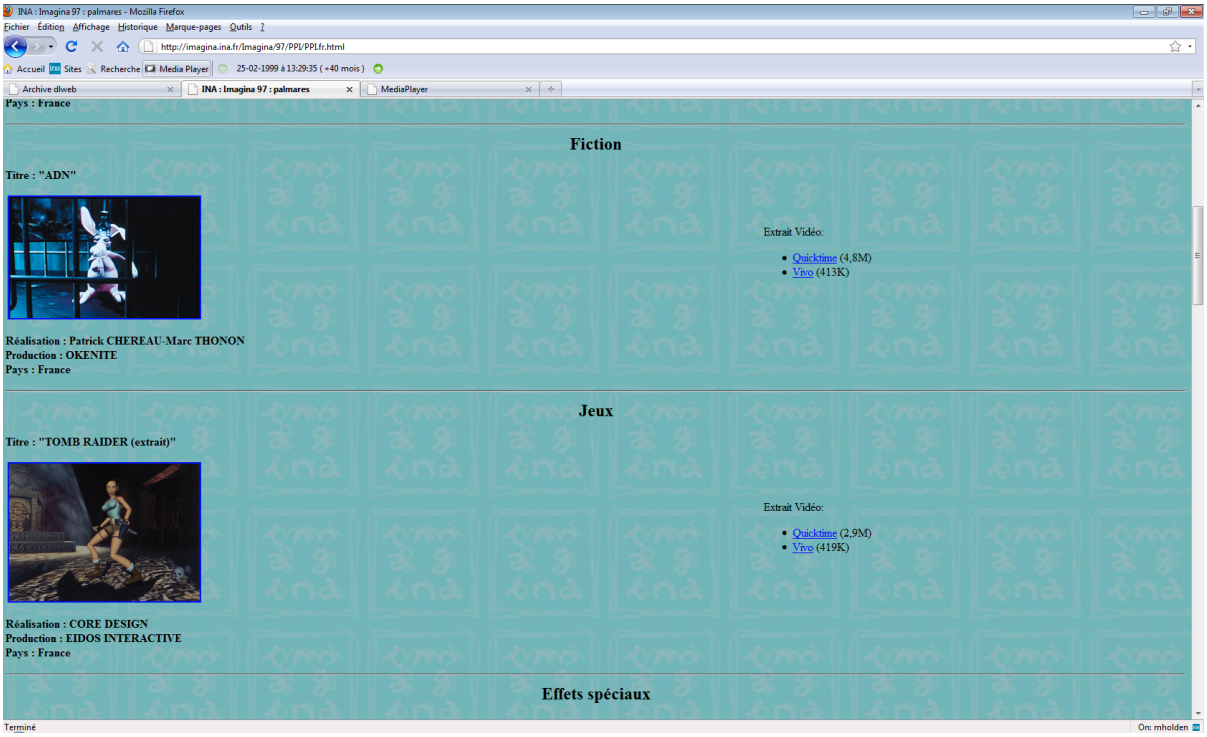

Figure 9. Web page with vivo video links

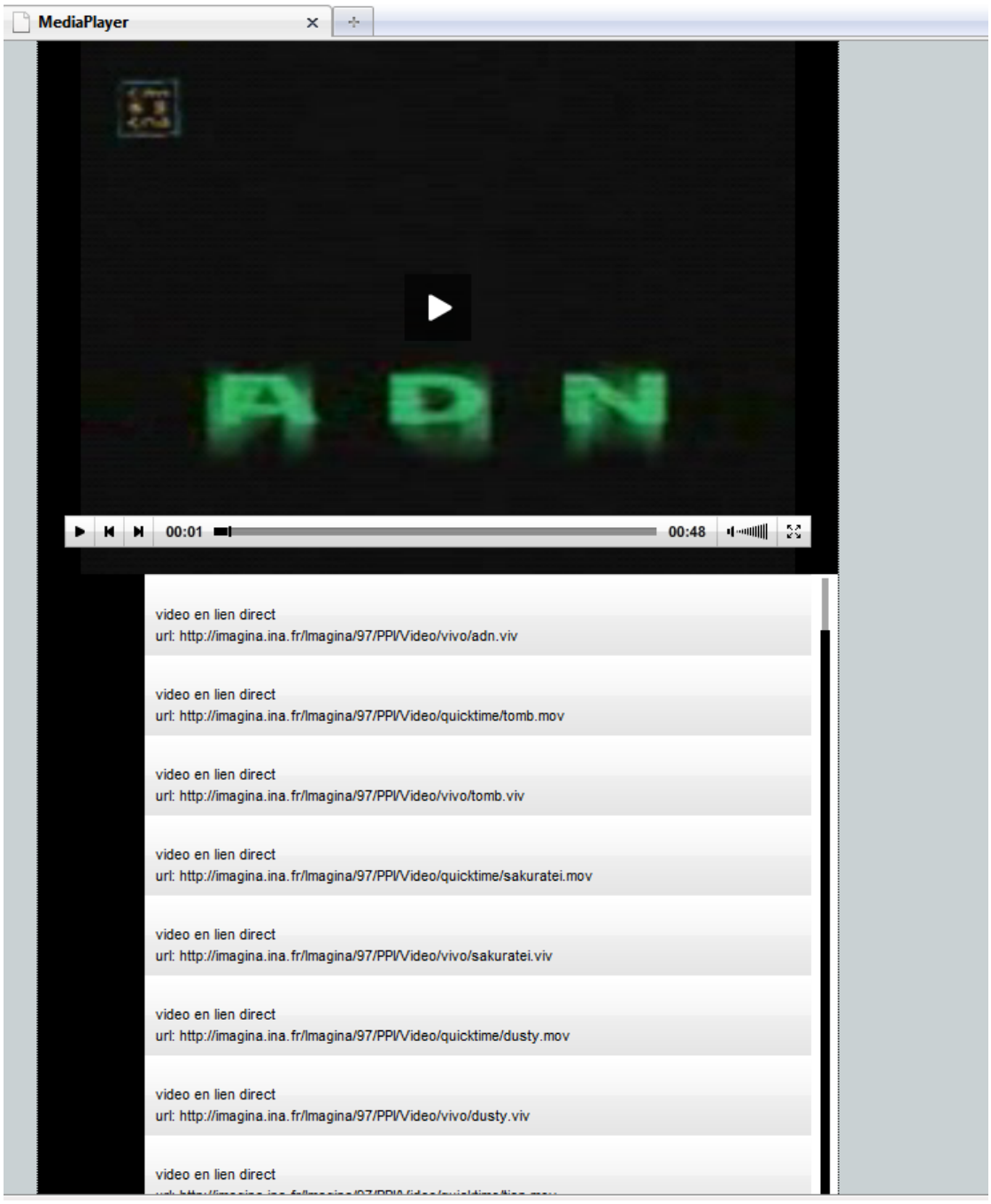

Figure 10. MediaPlayer playing videos converted automatically from Vivo to Flash video format