http://conference.ifla.org/ifla78 Date submitted: 31 May 2012

Radio, TV and audiovisual Web content collections: continuity and disjunction

Claude Mussou
Institut national de l'Audiovisuel (INA)
Bry-sur-Marne, France

Session:

148 — *Copyright law and legal deposit for audiovisual materials* — Audiovisual and multimedia with Law Libraries

Abstract:

Since it was created in 1974, 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 French audiovisual heritage. Indeed, Ina is now the world's largest digital audiovisual archive, holding over 4 million hours of television and radio recordings, dating back to the earliest broadcasts, with a further 800,000 hours of programs added each year, within the Legal Deposit framework. At the turn of the XXIst century, when the web provided a new dimension to publishing and broadcasters and new players seized opportunities offered by the digital revolution to distribute audiovisual content online, the legal deposit was extended in France to the Web. Interestingly, French law makers assumed it was necessary to share the responsibility between the BnF and Ina, which was designated by law as national repository for audiovisual media web sites and on demand audiovisual media services.

Ina began collecting over 8000 broadcast related web sites in February 2009. This collection currently complements and ensures continuity for radio and TV programs collections and innovative technical frameworks for harvesting, storage, access have been developed and implemented to handle and tackle specificities of "new media" contents.

Ina: The French heritage institution for sounds images and audiovisual web contents

Since it was created by law in1974, the French Institut national de l'Audiovisuel, 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 French audiovisual heritage when national broadcast public monopoly no longer prevailed.

Effectively, in France, in the second half of the 1980's, broadcasting licenses were granted to the newly born private sector and no regulation ensured that their programming and output would be collected and saved for heritage purposes. The academic community lobbied fiercely and argued strongly that radio and television broadcasts should and would be testimonies and evidence for future generations of scholars. Due to its prior experience and legitimacy in collecting and preserving audiovisual material, Ina was designated, by a law voted on June 20th 1992[1], as responsible for the Legal Deposit of radio and TV. Some twenty years later, the institute is considered the world's largest broadcast archive, holding over 4 million hours of television and radio recordings, dating back to the earliest broadcasts, with a further annual increase of 800 000 hours of programs from 100 television channels and 20 radio stations digitally recorded, around the clock, seven days a week.

Evidently the digital era also opened broad perspectives for Ina's aging collections and a vast digitization plan was launched as early as 1999 to safeguard and transfer in digital formats about 830 000 hours of at risk analogue material. By the year 2015 all of these will have effectively been digitized and several times migrated so as to ensure their long term viability and access. IPR issues were negotiated for part of these collections (30 000 hours) so as to make them available online to the general public, restricted online access is offered for professional usage of any archived broadcast for which Ina holds producer's rights (over 1M hours), all of it can be searched, viewed and analyzed on site for study and research purposes (over 4M hours).

In parallel to the transition from analogue to digital for sounds and images transmission and preservation, the World Wide Web became a paramount tool for publishing and accessing various types of content. Broadcasters and new players from the telecommunication business evidently seized the opportunities offered by digital technologies (compressed digital files, broadband access) to distribute audiovisual content online and launch cross platform strategies, both accompanying and triggering an evolution – if not revolution - in usage and video consumption from multiple devices[2].

---

[1] French Legal deposit Law on broadcast material, voted June 20th 1992 http://legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000723108&categorieLien=id (Accessed on May 6th 2012)

[2] Research in viewing behaviours has shown that online video is currently experiencing unprecedented growth. According the the Internett adverstising Bureau, it took 38 years for radio to reach 50 millions users, 13 years for TV, less than 5 for Internet, and less than 2 for Internet video …In the UK, 27.3 million, of the 38.5 million people who went online on their PC in February 2012, viewed streamed content. (Source: UKOM/Nielsen, Feb 2012)

Following the fast pace and rhythm of publishing technologies, the Legal Deposit regulation in France was then extended to the Web. Interestingly enough, French lawmakers assumed it was necessary to share this responsibility between the French national library, BnF, and Ina, to ensure coherence and continuity of their respective collections[3]. Ina was thus designated as national repository for audiovisual related Web sites – as broadly as this may point to – as well as on-demand audiovisual media services available from Web platforms. An enactment of the law was recently published that precisely defines the joint and shared mission of each institution[4].

Coherence and continuity for the collections were indeed the backbone concepts behind the legal framework assigning Ina this new mission, yet its technical and practical implementation departed largely from the methods, tools or practices in use for archiving broadcast material. This essay will attempt to provide a general overview of the issues at stake in selecting, acquiring, organizing, accessing, storing and preserving Web archives that link, refer to or complement traditional broadcast media.

Selecting

In France as in most countries a Government agency, the CSA, is charged with regulating the transmission of most national broadcast communications. This mainly refers to radio and television. Its responsibilities include the allocation and regulation of various frequencies for broadcast transmissions. Those are thus subject to authorization and limited in numbers. Before a new channel goes on air – which does not happen on a daily basis as it does for Web sites -, it is made public ahead of time and if it falls within Ina's scope for the archive, the process generally is anticipated and goes smoothly. It operates very differently for Web sites. Those are born and disappear without clear notice. The French domain registration office (AFNIC) can regularly provide a list of the .fr domain hosts, yet it represents only 30% of the French Web and is not classified according to business activities or theme of the Web sites.

Selecting and assessing which Web sites should be collected, at which frequency and depth in accordance to their refreshing update and size, is thus the first step in the archive workflow. Because the Web is boundless, fleeting and ephemeral, defining the scope of the collection and frontiers of the domains crawled, is essential, but a rather time consuming, ongoing process, involving above all human judgment. Based on objective criteria stated in the enactment of the law, documentalists at Ina take turn in keeping track of the relevant Websites, and nominating them for crawls.

Following the terms of the law enactment, the selection is both based on the activity of the publisher behind the domain (broadcasting activity), on the fact that its chief topic relates to radio and television (many blogs and fan Web sites in that category), or that it delivers on demand

[3] Law on IPR in the information society, extension of the Legal Deposit to the web, voted August 1st 2006 http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000266350 (Accessed on May 6th 2012)

[4] Enacment of the law, regarding the Legal deposit legislation published on December 19th 2011 http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000025002022 (Accessed on May 6th 2012)

video or provides non linear on line access to broadcast programs (replay, catch-up and original Web only video content).

The process has been in production for over two years, it started out in February 2009 with 3600 Web sites and has so far expanded to 10 000.
It is noteworthy that the prevailing ambition of the Legal Deposit to gather comprehensive collections needs to be reconsidered as far as Web content is concerned, the guarantee that all relevant Web sites will be identified and selected cannot be achieved and the commitment rather aims at a « best effort » approach with active monitoring of the live web by dedicated professionals. Semi automatic semantic crawl methods have been tested to support human judgment that have not proved relevant regarding the scope Ina is considering.
We will demonstrate later on that this « best effort » approach also prevails for the acquisition or crawls aspects.

Collecting

In the years of the analogue age, Ina had been collecting broadcast material on physical carriers, following traditional methods of library acquisition techniques. With the widespread use of digital technologies, as early as 2001, direct around the clock recording was organized on a large scale for over one hundred channels, and Ina tapping directly from the broadcasters flows into its storage system.

In a sense the techniques and frameworks implemented to harvest Web content follow that trend and « invasive approach » to collecting information, even though unlike radio and television, the Web is a non linear media which does not “stream in” but requires specific and dedicated techniques.

Collecting content directly from Web servers is indeed the general approach, while the objective is to simulate all possible human interactions within the Web pages in order to generate as many responses and download as many contents from the remote website. These techniques commonly known as “harvesting” or “crawling” are used by all search engines to gather and process Web data. The tools used are commonly named “crawlers” or “spiders”.
The word “spider” might sound quaint and obsolete, but it is probably the most accurate to define the method, as it hints to the multiple paths and crossroads the automated tool will discover, follow or ignore in order to collect part of the world-wide “Web”.
This is not mere harvesting but rather a very precise and systematic technique starting from a seed list and following links to discover and download web content according to specific crawling parameters. All this would be simple enough if the Web was not continuously changing, with new links unpredictably appearing and others disappearing, with new or updated items to be discovered in each node. It is in fact possible to somehow predict or receive notice of these changes when they happen, RSS feeds acting as alerts, just like vibrations would for a spider on its web, meaning that something to be swallowed came up somewhere on the Web. But this is more the exception than the rule, and the spider usually has to continuously forage over the Web to discover new alluring food.

This analogy now reaches its limits, as spiders obviously spin and control their own webs, whereas crawlers involved in web archiving are invasive, like parasites spreading across a

gigantic web that millions of spiders could have spun. Moreover, when least expected “Web traps” (set intentionally or not), allow automated tools to fall, as ants would in a sand pit trap. So we’ll now cut short with the insect metaphor and eventually resume to using the word “Crawler”.

Because Ina’s scope for the web legal deposit is rather focused and will never amount to the number of websites included in large domain crawls, an in-house scalable crawling system has been developed to better fit the diversity of the web sites (regarding update frequency, depth and interactive features).

The system is based on a two-tier architecture, with a main scheduler sending orders to a multitude of crawlers, each one handling one Web site at a given time, with 500 to 1000 of these crawlers typically running on a single machine.

The scheduler (the top part of this two-part architecture) evidently handles the scheduling aspects and the configuration of each Website. Using a multi-sampling strategy, it monitors update frequency and rhythms:
Websites are categorized (in a semi-automated way) depending on frequency features (constant, hourly, daily, weekly updates etc.) and crawling is scheduled accordingly.
As surface Web pages (defined by a limited amount of interactions or « clicks » to access them) are more likely to be updated than “deeper” pages, the surface pages for each Websites are crawled more often than the deeper ones,

Some Websites have syndication feeds than link to new contents. These feeds are used to initiate specific crawls for the single page of an article or content as soon as it is published within a given website, automated revisits follow up for updates and comments

This approach deals separately, on the one hand with scheduling strategies (frequency and depth crawl parameters matching each website specific requirement), and on the other hand, with crawling issues (Web traps avoidance, crawling rules and politeness enforcement, storage). It allows the use of different crawlers at once. Because the web is a jungle haunted by a multitude of heterogeneous technical systems and formats, crawlers are prone to errors and misses while striving to trigger interaction.

This dedicated multi crawler approach aims at collecting various types of contents following the previously mentioned « best effort » strategy, enhancing at once the quality of the archive.
Up to three different in-house crawlers can be connected to the same scheduler, each one dedicated to a specific task:

PhagoSite, is a general-purpose crawler, which can handle very large Websites and does not require much computer resources.

Fantomas, is a more specific crawler based on the *Phantom-JS* Web kit, which uses the same core functions as Google Chrome or Apple Safari browsers. This crawler is able to crawl most “2.0” Websites with fancy JavaScript interactions, without being too stringent with computer resources.

Crocket, is based on the Firefox browser and can crawl rich and complex Websites (rich-media and rich-interactions), it is however very computer-intensive.

In addition, and as video content is a large part (both in number and in size) of the archive, dedicated crawling tools for downloading UGC videos from YouTube or Dailymotion, or collecting live streaming contents have also been developed.

This dedicated, customized crawling system has been running continuously since February 2009, at a current pace of 6 billion requests a year. Tools are evidently continuously updated and upgraded to adapt to the ever-changing and still maturing world of the Internet.

Description, Metadata and Access

Film and broadcast archives just as digital or web archives imply a technological bias to access their content that did not exist for any previously published document. Content from a book can be immediately available, provided you can read, whereas physical carrier such as film, tape, disk, or digital files store content information that need to be reconfigured, calculated, interpreted to be humanly intelligible.

At Ina the tasks of the documentalists have gradually been reoriented considering the vast amount of data involved both for broadcast flows and Internet content, and metadata extraction and management is gradually becoming the rule.

Yet, in keeping with a traditional approach to organizing collections, Web sites are documented and catalogued according to dedicated taxonomies that attempt to bridge the broadcast and web collections.

Archived Websites contents are also automatically indexed to enable random-access from an URL and date combination.
Storage on disk, of files and metadata effectively builds up the archive. Due to the characteristics of a digital archive storing discrete information, and to the nature of the web as a publication tool, access to the « archived » web pages (a recent estimate considers that on average, over 50 individuals files now make up a single web page) implies a reconstruction of the publishing process by accessing the files that have been crawled at a given time then allowing their display within a reconstructed page.

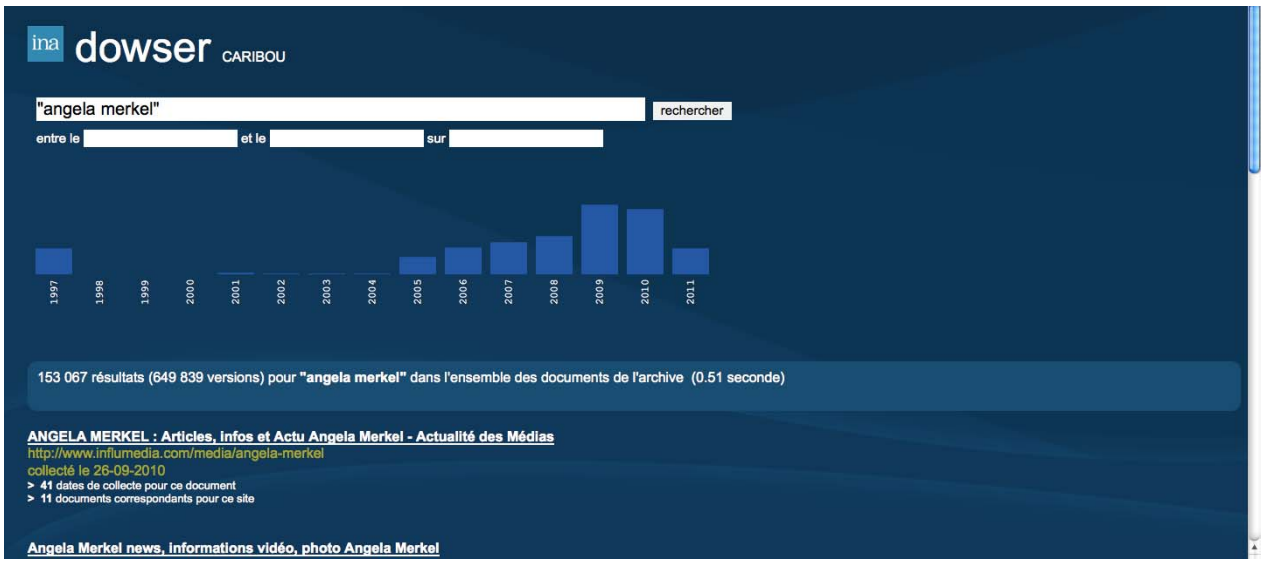
The archive browsing tool is a customized Firefox browser. From one given capture, it is possible to go forward or backward in time and eventually display all available versions. Most interaction remain active (link navigation and basic interactions still represent 95% of the user experience), some don’t for technical reasons (interactive Flash contents, or complex JavaScript interactions, including some video players). Evidently the aim is to capture and display the « look and feel » of the content and pages as they were originally published yet this attempt is not fully successful, most of the time due to technical impediments.

For instance, some of these lost interactions need to be somehow recreated to enable a thorough browsing experience, i.e.:

Videos that will not play from their native embedded player will be played with an external player; additional features can be provided, such as seeking inside the video.

As search engines such as Google or Bing cannot be archived (a simulation of all possible user interaction is indeed impossible), a specific search engine was developed for searching Ina's web archive, it was customized to handle the time dimension in the archive and cluster the duplicates or near duplicates results.

Dowser, is our "magical" search engine that allows the user to access any of the archived content from text queries in a « Google like » way. Histograms are used to help the researcher navigate through dates and occurrences for any given search query.



Authenticity supports the legitimacy of an archive, yet the notion has been greatly questioned in the digital environment. The sheer notion of « original » document will soon have disappeared and digital data is bound to be copied, migrated or manipulated. Access to the archived web content offers consecutive displays and views of Websites, but does not provide access to a canonic version of the Website which does not exist. The option at Ina is to inform the user of any potential modification of the original online content or context (temporal discontinuity between a linked page and its referrer, video played in an external player instead of the original one, missing contents due to crawling limits, etc.), stressing the fact that the archived web document is actually not an artifact but an incomplete reconstruction of an original media. Again the "best effort" approach is we think the right way to go!

Storage and preservation

Content published on the web is for the most part born digital and has no physical existence anywhere else. Provided it will hold many of the historical records of our times, and be testimony of the trends in our societies, long term preservation aspects are central, as well as the need for

large storage solution that enable migration of data through the ever developing technologies of magnetic tape and hard disk storage.
As everyone in the web archiving community recognizes the over-arching need to develop a better framework, the long-term preservation strategies for web archiving are still very much under development and most institutions rather focused on giving as wide an access as possible to their collections to keep the collection process alive. We will point hereafter at the main issues that are being debated by the web archivists' community and which actually meet those of all digital archives.

Assuming the integrity of the data and bit level is maintained, will the stored data be reinterpreted correctly in the future? Just as the possibility to read film or videotapes must be maintained, the question of migrating file formats as well as maintenance of browsers or plugins is essential.

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 digital archives, using short to medium term migrations (around 20 year strategies).
At Ina, 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.

Disk Storage
The issue of storage requirements is all the more fundamental that from pages of formatted text (HTML) in KiloBytes, the shift to hosting a growing amount of audio and video files requires many MegaBytes.
Unlike tape storage, disk storage typically has much higher associated costs regarding power and cooling. The reliability and failure rates of disk technology should be taken into account, which limits the typical use of these systems to between 3 to 5 years. To facilitate migration and cost savings attention is given to double the storage capacity on each migration (e.g. from 1.5 to 3 TB disks)

LTO (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. Starting in 2000, the LTO standard defined a standard sized cartridge with increasing capacities over 8 generations - approximately a 20-year period (2000 to 2017). The capacity will approximately double with each new generation, facilitating migration strategies and allowing standardization in physical storage depots.
Even though the magnetic tapes can be theoretically stored up to 25 years, there is no guarantee that the LTO vendors will produce drives in the future capable of reading old tapes.
In fact the LTO specification only mandates the compatibility standard that each version of LTO is read compatible with the 2 preceding versions (e.g., LTO-4 drives can read LTO-2 tapes)
Therefore, to reliably migrate data, it needs be transferred to the new format when it becomes widely available and is economically viable. For example, despite LTO-5 drives becoming available in Q2 2010, only in Q2 2012 is the price per cartridge becoming favorable in comparison to LTO-4.

Similarly, economies in pricing will be considered when migration is decided, which, to maximize efficiency and space will be between 2 LTO generations, quadrupling the capacity i.e. migrating from 4 LTO-4 cartridges to 1 LTO-6 cartridge.

Checksums

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

Long-term archive preservation: emulation or migration?

It is necessary to consider the web archive from 2 different viewpoints:

- The « look and feel » of web pages – the « aesthetic » form which communicates the style, imagination and flair of its creator
- The component data parts; the written word conveying the ideas of the author, images, sounds and videos

The first form seems to be more open to emulation type strategies.

Currently a large number of organizations and institutions are running projects to categorize and emulate old browsers and plugins.

A way to verify that emulation is working correctly could be envisioned by taking either a video or snapshot of web content.

The second form lends itself much better to a migration type strategy; where, at the point of capture (or even within several years) a standard file format can be identified for medium-term (say 20 years) migration of the data.

In the above example PDF or ASCII might be chosen for all forms of text, JPEG2000 for all images, AIFF for sound and MPEG4 H264 for all video formats. These are currently the standardized formats used at Ina for the long-term migration of audio and video archives.

This type of approach could also lend itself to extracted metadata, where a number of software tools can be run to help identify files and formats at the time (DROID, JHOVE or Apache Tika for example) and storing of the results in standardized formats is organized.

Ultimately, the aim is to seize as much information as possible about a web page during the point of capture. Even with the best migration and/or emulation strategies, there is no guarantee that any given page will be able to be recovered in the way it was originally stored. Visual aspects of the page will possibly be restored but interaction will almost certainly not be achieved in the same way.

However mixing the above strategies will enable the use of various tools by a future user.

For example, he or she may be able to see a page created by an emulator, which may not be able to read an associated document in the reconstructed page. However, this emulation could be compared to an image snapshot of the page so as to give an idea of its visual form, and then regard the (migrated) extracted metadata to perhaps recover the content of the original document.

Conclusion

We have no means to really estimate the long term impact of the digital storage of all knowledge, but archives and national or academic institutions are trying their best to ensure that it is kept alive in the long run. Web archiving is a step forward which notably involves archiving content that has a short life expectancy on the live web, keeping up with numerous and ephemeral formats, maintaining interactivity in the archive or enabling « user experience » as it was originally intended. It is now a shared activity and even if approaches and options may differ from one institution to another, the cooperation among them within the IIPC fosters dialogue, exchange in practices, involving users – academic or professional - from many countries so that the history of World Wide Web and its content can be written by future generations.