Archiving before Loosing Valuable Data?
Development of Web Archiving in Europe

France Lasfargues
Internet Memory Foundation
45 ter rue de la Révolution
F-93100 Montreuil
E-Mail:
france.lasfargues@internetmemory.org

Chloé Martin
Internet Memory Foundation
45 ter rue de la Révolution
F-93100 Montreuil
E-Mail:
chloe.martin@internetmemory.org

Leïla Medjkoune
Internet Memory Foundation
45 ter rue de la Révolution
F-93100 Montreuil
E-Mail:
leila.medjkoune@internetmemory.org

0 Introduction: Why is it so important to archive websites?

By nature, Web content is ephemeral: sites disappear continuously and are frequently updated, involving the disappearance of what is often very valuable online information. It has been observed that 80% of website pages are updated or disappear after 1 year (Ntoulas, Cho & Olston, 2004). This media is pervasive in our society and certainly today one of its most important representation. Even printed publications suffer from the effect of web data transience, for instance when pointing to online resources that become unavailable (Spinellis, 2003). From presidential elections to music festivals, every event is an opportunity to create, update or close a website, a Twitter, a Flickr or a Facebook profile; to communicate about it, to feed blogs, forums, social websites with opinions, point of view, issues, etc.

Like any other media, the Web deserves a memory and it is essential to preserve what has a cultural, heritage and historical value.

But challenges come from the feature and richness of the Web: dynamic content, volatility, variety of formats and Internet users contributions, all attributes increasingly used. Archiving the Web requires a special attention in order to retain its value and ensure its greater fidelity to the original.

First we will present an overview of Web archiving activity, with a highlight on Europe, then continue with the State of the Art in this field and challenges Web archiving community will have to face.

1 Inventory of Web archiving

1.1 International interest

Our aim here is not to do an exhaustive description of every such program but to summarize and highlight facts and issues regarding Web Archiving.

Since the beginning of the Internet Archive (http://www.archive.org/), the first notable Web archiving initiative and the most ambitious one in 1996, a number of international, national and institutional Web archiving programs have been launched. From major national libraries or archives to consortia, foundations and university departments, we can find a wide diversity of Web archiving projects in developed countries.

The International Web archiving Workshop (IWAW – http://www.iawaw.net/) began in 2001 and yearly presents updated work about this topic.

From 12 members in 2003, the International Internet Preservation Consortium – IIPC (http://netpreserve.org/) counts now 36 members all over the world (4 in Asia, 24 in
Europe, 8 in North America, 2 in Oceania and 1 in North Africa). Goals of this group are:
- To enable the collection of a rich body of Internet content from around the world to be preserved in a way that it can be archived, secured and accessed over time,
- To foster the development and use of common tools, techniques and standards that enable the creation of international archives,
- To encourage and support heritage institutions to address Internet archiving and preservation.

IIPC has its main achievements in the development of common open source tools and standards.

In Europe, since 2004, the Internet Memory Foundation (formerly European Archive Foundation) actively supports the preservation of Internet as a new media. To fulfill its mission, the Foundation developed a wide range of collaborations all around the world, and especially in Europe, with half-dozen cultural institutions, about thirty research teams and the spin-off Internet Memory Research. Thanks to these partnerships, the Foundation is archiving dozens of Terabytes of data per months, via a shared Web archiving platform, ArchivetheNet (http://archivethe.net), in order to help institutions to easily and quickly start collecting websites including dynamic content and rich media, enabling navigation in past versions of a website, as if it was still live.

The growing interest for Web archiving is also noticeable through presentations and workshops in international professional and research conferences. Almost all of them now include a focus on Web archiving: European Conference on Digital Archiving in 2010, Association of European Research Libraries, LIBER 2010, International Federation of Television Archives, FIAT/IFTA 2010 and 2011, Museums and the Web, 2011 etc (See Presentations in References).

In addition to this, many surveys and studies are published: in 2008, the IIPC published results of web archiving project of its members (IIPC) and in 2011, the futures of Web archives (Meyer, Thomas & Schroeder, 2011); The Foundation for National Scientific Computing created a Wikipedia page named “List of Web Archiving Initiatives” to enable collaborative updating (Wikipedia) and also published a survey on web archiving initiatives in 2011 (Gomes, Miranda & Costa, 2011) focused on an international point of view (42 initiatives worldwide). This latter gives among others interesting features:
- On the evolution of the number of Web archiving initiatives: since 2003, 31 initiatives have been seen light; one possible explanation exposed is the concern raised by the United Nations Educational, Scientific and Cultural Organization (UNESCO) regarding the preservation of the digital heritage (UNESCO).
- On the size of Web archives worldwide, which preserved since 1996 a total of 181 978 million contents (6.6 PB); the Internet Archive itself holding 150 000 million of web content (5.5 PB).

1.2 European overview

In the framework of a research project, Living Web Archives (LiWA), funded by the European Commission from 2008 to 2011 (http://lliwa-project.eu/), the Internet Memory Foundation carried out a survey on Web archiving in December 2010, sent to European and International institutions. We will summarize here its main findings.

Which institutions do Web archiving?

This survey has been sent to more than 360 institutions (most of them are European heritage institutions):
- National and Regional Libraries (19 %)
- University or specialized Libraries (22 %)
- National and Regional Archives (11 %)
- Audiovisual and/or Broadcasting Archives (20 %)
- Institutional Archives, as Parliament Archive, European Institutions and International Institutions (18 %)
- Documentation or Archive Department of Museum (10 %)

Most of the respondents (73 European institutions fulfilled the survey) were National Libraries (25 %) and Audiovisual Archives (25 %), followed by National and Regional Archives (15 %). It is also interesting to notice that 42 % of the institutions that do not currently preserve Web material, consider that it is or should be in the mandate of the National Library.

Maturity of Web archiving programs

This survey shows that more institutions than expected launched a Web archiving program: 41 active institutions (fully operational, experimenting, and starting) have answered and many of them are not part of the IIPC (26) and that 17 institutions answered they do not have a current Web archiving program but they plan to create one in the next 3 years.

What do they archive?

Web archiving can target only one’s own website, a selection of external sites, or all sites of a given domain (like all German speaking sites, or all sites under the top level domain .se).

A Web archiving selection policy may have to be formulated in the context of an existing organisation-wide selection policy, or analogous selection policies for their type of resources. It is indeed quite likely that the National Library of a country where a legal deposit on printed resources is in place would articulate its Web archiving policy along the same lines.

One approach is to take the decision not to select individual sites, but try to archive all those that are being automatically discovered, as the Internet Archive did (http://www.archive.org/web/web.php). This holistic approach is also possible within a more limited context, such as national domains, as it is the case in France or Sweden for example.

According to the results of our survey, only national libraries (33 % of them) operate these holistic crawls on their national Top-Level Domain, but the large majority (93 %) focus on selective and thematic crawls.

Audiovisual archives (75 % of them) mostly focus on their own websites (only 38 % collect according to Thematic and Selective crawls).
Regarding selective crawls, institutions can focus on specific domains (such as cern.ch), on events (elections) or subject or categories of websites (such as all governmental websites).

Overall, content collected depends on the kind of institution but also on legal limitations.

**Legal context**

50% of institutions are operating their Web archiving programs under a specific legal framework. The others are expecting one or considering that the existing one is sufficient albeit not specific.

**Management**

To fit with their Web archiving project, these institutions develop their own strategies and skills. 64% operate their Web archive in house versus 36%, which outsource this service.

Most institutions use the open source crawler Heritrix (80%) and manage their collection with a curator tool together with an in house solution for 35% of them. 35% use open source solution such as NAS, WCT or Pandora…

**Which access to collections do they provide?**

Depending on their legal framework, all institutions do not provide the same level of access to their collections: 41% provide an open access, 28% an online access with restrictions and 21.3% do not provide access (content is in a completely dark archive).

When providing access, users can browse Web archive by URL (68%), search by keyword (70%) and/or through thematic collection (65%).

**2 State of the art and challenges**

More than a decade of research and practice has established a relatively solid consensus on the tools and methods to be used in Web archiving. Nonetheless, as the Web is still changing at a very rapid pace, permanent adaptation is required in this domain. In this section, we will present the main stages of a typical Web archiving workflow, with for each, a state of the art as well as some of the challenges that current practitioners face.

Web archiving projects follow a generic production scheme, which includes several phases from selection to hosting and quality review of archived content.

Once all steps are completed (quality review), some phases such as capture, might start again depending on the overall quality of captured content and in regard of the institution archiving policy.

These practical phases of the workflow are presented in Figure 1.

Web archiving requires tools to collect, preserve and access archived Web content. The International Internet Preservation Consortium (IIPC) has developed tools and methods to help in these processes.

Some institutions, such as the Internet Memory Foundation, but also companies or universities also develop their own tools.

---

Figure 1: Web Archiving production scheme

**2.1 Crawls strategy**

Collecting the Web is usually done using crawlers, or “robots” working in a similar way to search engines crawlers. They must adapt to the heterogeneity of architectures, (dynamic, social Web, etc.), languages and formats (video, text, audio files, etc.) and must also interpret links. Links are the centre of crawling problem as they are the key to content discovery.

Several types of links can be found:

- Explicit links: source code is available and full path is explicitly stated
- Variable link: source code is available but use variables to encode the path
- Opaque links: source code is not available.

For extracting links, two main strategies and tools are developed:

- Parsing
  
  This method consists in creating crawlers that parse content, extract link and archive the discovered content on the fly.

  This approach deals with explicit and some of the variable links type. However, many of them, as well as opaque links remain out of reach, which entails quality loss.

- Execution
  
  Another method consists in mimicking a browser, by executing content to capture it. Hanzo Archives, a private company, developed such an execution-based crawler (http://www.hanzoarchives.com/) and currently use it to archive a wide range of sites that would be impossible to parse. Analyses were made as part of the LiWA (Living Web archives–http://liwa-project.eu/) project and demonstrated significant im-
provements in the quality of crawls, especially when targeting dynamic content.

The issue with this approach is the time it takes to detect link on any given page, which poses serious scaling issues.

A combination of both approaches, as done by the Internet Memory Foundation, makes it possible to keep the best of both worlds.

- Large scale crawl

The ever-growing size of the Internet raises a serious challenge to the design and implementation of web crawlers. Among issues that must be solved, SPAM and trap detection, politeness compliance, dynamic HTML and hidden Web analysis, to name a few, the frontier management (e.g., the ability to detect that a page has already been visited), is probably the most prominent from a pure scalability point of view. Current tools don’t really scale to the size of national domains, not to mention the entire Web.

A distributed approach on several servers must be taken to tackle the task of collecting billions of resources during weeks of harvesting campaigns. The Internet Memory Foundation has taken this approach to develop of a new fully distributed and scalable crawler. Its design divides the crawling task in individual units, each of which can be assigned to a server. At the global level, these servers cooperate to exchange the core information involved in the crawl processing, in particular newly discovered URLs. The frontier itself is partitioned, each piece being maintained locally without need to exchange. This makes possible very large crawls whose scope can be broaden by adding new servers capacity. A single crawler node (server) achieves a throughput of approximately 100 collected resources per second, about 8.5 million per day, and 180 million over a 3-weeks period. By assigning dozens of such nodes running in parallel, we can envisage the capture of billions of pages.

This implementation effort deals with the software issues inherent to large distributed system: risk of failure, load balancing, networking issues, etc.

To improve crawl processes, additional modules and plugins are developed by institutions, consortiums or companies.

In the context of the LiWA Project, new technologies were developed to extend the current state of the art for capture, analysis, and enrichment services altogether, in order to improve fidelity, coherence, and interpretability of Web archives. We present some of these technologies below.

**Spam**

Web spam (automatically generated Web content to cheat search engines), has become a real industry estimated to represent up to 20% of the Web. In a production process, Spam can be seen as a waste of time and resources. In the context of Web archiving, we can easily understand that archiving spam is a big challenge in terms of costs (saving time during crawl process and hosting) and content compliance.

To detect spam, a module has been developed during the Project LiWA. Using machine-learning techniques to detect spam based on several features, this module enables assessment of the crawled Web hosts with a very good accuracy (over 90%).

**Temporal coherence**

Website may take hours to be crawled during which changes can be made or pages can become unavailable. The LiWA module includes methods for proper dating of Web pages and provides coherent crawls with complete and correct temporal metadata, including time-aware reference sources such as the Wikipedia history.

**Rich Media Capture**

Internet is a perfect media to offer a good visibility to videos but capturing Web video is quite complicated because technology used to publish or stream them aims at avoiding direct access to the files by the users. Two main categories of issues are highlighted below:

**HTTP protocol**

Websites that use standard HTTP protocol to deliver video content can be difficult to archive. The issue lies in the various techniques used to obfuscate links to video files (2 or 3 hops and redirects). YouTube is a representative example of this technical problem. The current version of Heritrix is mainly based on the HTTP/HTTPS protocols and it cannot treat other content transfer protocols widely used for the multimedia content, such as streaming.

**Other protocols**

Some sites use transport protocols other than HTTP, such as RTMP streaming protocol (e.g. http://www.swr.de). RTMP stream life protocol is used to deliver big files and delivers content by package to facilitate access. Videos are more or less embedded in a player (Flash, Silverlight, Real media,) to protect both content and access. Players use methods to hide files. To collect a video, the crawler has to discover hidden links. One of the possible methods consists in using an external tool that downloads the dump of the content and saves it into a file.

The Rich Media Capture module (RMC), developed in LiWA and used and extended by the Internet Memory Foundation, is designed to enhance the capturing capabilities of a crawler, with regards to different multimedia content types.

**Social Web**

Our society is characterized by radical changes in the way we create and communicate information. People are also more and more involved in social networks (there are now more social network items created than Google searches created). Social medias are becoming more and more pervasive in all areas of life and therefore a challenge for preservation institutions.

On one hand, this material is both ephemeral and highly contextualized, making it increasingly difficult for archivists to decide what to preserve. On the other hand, social Web is based on dynamic websites and uses some specific technologies to publish content.

Web archiving institutions and consortia address these issues by publishing and circulating guidelines on crawls parameters. (See for instance guidelines published by the Internet Archive to enhance the quality of social media
content: https://webarchive.jira.com/wiki/display/ARIH/Archiving+Social+Networking+Sites+with+Archive-It).

The European Commission also funds research projects to address this issue such as the ARCOMEM project (Collect-All ARChives to COmmunity MEMories–http://www.arcomem.eu/). Collecting social Web requires new tools as well as new strategies. In contrast to classical Web harvesting crawlers, ARCOMEM will provide facilities for extracting complex objects from the Web, by changing the behaviour of a crawl depending on the kind of Web application targeted, and by interacting with content analysis modules. An “Application aware crawler” will be developed for this purpose. This module aims at making a crawler aware of the particular kind of Web application it is crawling, in terms of its general classification (wiki, social network, blog, Web forum, etc.), its technical implementation (MediaWiki, WordPress, etc.), and even its specific instance (Twitter, CNN, etc.), in order to adapt the crawling strategy to the task at hand. The crawler will rely on a hierarchical knowledge base of Web applications, which specifies how to recognize an application and how to crawl it.

2.2 Quality Assurance (QA)

As explained above, capturing Web content is full of challenges and harvesting tools used have clear technical limitations.

Because of the limitations and the now obvious incompleteness of Web archives, most European institutions are looking into developing Quality Assurance methodologies and tools that could be applied to Web archives.

The most used Quality Assurance methodology is the visual method, followed by a capture of the identified missing resources. It applies mainly to selective harvesting and consists in visually checking pairs of Web Pages: live version versus archived version. It implies the use of the live version as a reference when checking quality of captured content and therefore means that it should always be done as fast as possible after the crawl, due to the ephemerality of Web content. Visual check methodologies are subject to institutions policies as well as to the capturing tools used. Indeed, technical limitations of tools used have direct consequences on the quality of crawls and a good knowledge of crawling issues and state of the art is mandatory to practice an efficient QA.

The “visual” methodology described above, if it provides good results on selective harvesting, obviously requires trained human resources and is time and money consuming.

To overcome these difficulties and apply quality review on larger sets of Web content, institutions and companies are now looking into implementing automated or semi-automated QA.

A first option is to use metrics related to crawls as references and comparison tools. This method, that can be automated to detect problematic crawls, is useful when crawling the same list of domains at different frequencies. For instance, one could decide to store metrics from one reference crawl, which would have gone through visual checking and compare them to all future crawls during a set period of time.

Other options consist in developing specific QA tools based on proxy, execution (Selenium–http://seleniumhq.org/) or image comparison tools (see for example the project SCAPE–http://www.scape-project.eu/), to mimic “manual” QA and either detect missing elements to fetch and index them automatically, either evaluate quality of crawls at a high level to minimize human intervention and reduce costs.

There is currently no packaged QA tool available as this activity relies half on team’s knowledge and available technologies, half on institutions needs. Some Web harvesting “service providers” such as the Internet Memory propose “packaged” offers to facilitate and promote Web archiving inside cultural institutions of all size (http://archivethe.net/en/). Some “curator tools” also try to integrate QA facilities, such as metrics or browsing options (http://webcurator.sourceforge.net/ or http://netarkivet.dk-suite/).

2.3 Access

A Web archive is a copy of website recorded by a crawler at a specific date and time. Each institution involved in Web archiving develops its own strategy. The Internet Archive and the Internet Memory Foundation make the main large public web archives available. Some archives can only provide restricted access. This is the case for the French National Library, where archives are only available on site for accredited researchers, or in Austria for example. To browse the archive, users can usually use URL or full text search.

Most institutions use the open source Wayback Machine developed by Internet Archive with the support of IIPC to provide online access. Through a form, the user submits an URL. The tool provides a table with all archived dates for the requested URL. By clicking on a date, the user can browse the archive. The URL archived is rewriting on the fly by the Wayback machine to ensure that navigation is really done within the archive and not on the live Web. The latest version of the Wayback machine is using server-side rewriting.

The Internet Memory Foundation has been the first to implement server-side link rewriting four years ago and has been constantly improving its access since then. It enables for instance IMF to implement customized rewriting rules for difficult sites or to embed customized video players and other advanced on the fly presentation improvements.

This includes:

- A front-end web server that fetches the closest in time archived content for a specific URL and reference date. Resources of certain types get rewritten on the fly, for instance HTML documents have their links modified so that they point to the archive, or a banner can be inserted.
- A distributed index that allows fast access to a resource given an URL and a date.

This access tool is already supporting millions of requests per month.

Full-text Search

Currently, most implementations of the full text search are based on Lucene (http://lucene.apache.org/) a keyword-based search facility.
As Web archiving becomes more popular, user requirements have been taken into account in order to develop smart access. Thus, several user interfaces have been developed. The UI framework, developed during the LiWA project provides a new timetable browser and an organizer. Consortiums such as UKWAC (UK Web Archive) provide access to the archive by collection and using a time frame.

**Automatic redirection**

The Internet Memory Foundation implemented a redirection service that enables a user to be automatically redirected to the archive when the content is not anymore available online in collaboration with its partner The UK National Archives. This service is one of the first initiatives to embed Web archives as part of the normal web access user experience.

**Improving Web archives use**

More and more institutions are involved in Research projects and activities to develop tools and to get analytics from their web archiving collection. In this context, the UK National Archives have invested in an ‘intelligent discovery tool’ to improve searches of archived UK Government websites. The contract for the development of the Government Web Archive Semantic Knowledge Base was granted to a consortium of semantic technology professional lead by Ontotext AD Company and the Gate team at the University of Sheffield. The goal of this new project is to improve access to information contained in the web archive by providing far richer and more sophisticated information about what it contains by providing semantic search (Spencer & Storrar, IIPC 2011).

This issue is also addressed by the European commission through an European project (LAWA: Longitudinal Analytics of Web Archive data, http://www.lawa-project.eu/, 2010-2013). IMF currently develops an infrastructure apt at scaling to Petabytes of storage for Web archives. This infrastructure will provide extended analysis and extraction mechanisms for institutions interested by the derivation of statistical descriptors from very large datasets. These capabilities are of interest for archivist as a mean to provide an aggregated level of information that can help users to navigate, consult and interpret large and unstructured collections. Among the analysis and extraction methods currently under investigation, it is worth mentioning: entity recognition (e.g., a country’s name), language detection and links between documents.

### 2.4 Preservation and Metadata

The issue of the preservation is linked to the issue of format and metadata. The IIPC has achieved to create a dedicated ISO standard (ISO 28500:2009) for Web archives storage of content.

![Figure 2: Archived dates of an URL. Provided by Internet Memory Foundation on behalf of The UK National Archives](image-url)
and metadata. This format specifies a method for combining multiple digital resources into an aggregate archival file with related information. Resources are dated, identified by URIs, and preceded by simple text headers, using MIME entities. By convention, files of this format are named with the extension “.warc” and have the MIME type application/warc. The WARC file format is a revision and generalisation of the ARC originally developed by the Internet Archive to store information blocks harvested by Web crawlers.

In terms of preservation, a Web collection consists of collected documents and their metadata and form together a network of content and description of content.

Technical dependency is key for Web material as well as for digital object in general.

Several levels of metadata have to be collected and preserve to ensure the reliability of the archive. Metadata from selection process give information about the context of the crawl, which policy content was chosen, its original place, and its evaluation... Metadata about documents (size, format, checksum…) can be leverage for preservation policy (migration, emulation and bit-level preservation). In addition to this, process metadata gather information about the communication with the server during the archiving process (giving important context to the documents preserved).

Challenges come from the richness and variety of Internet content and characterisation of content within Web archives.

3 Conclusion

After more than a decade of development, Web archiving has become a mature field and is becoming part of many institutions’ mission. A consensus on best practices and methods exists, and institutions can choose between several tools and services to implement a Web archiving policy.

Since the Web is born, it has evolved in many ways and the continuous development of the Web will require Web archiving actors to think and adapt at fast pace and to develop scaling technologies.

Research areas are mainly the scalability and automation of process and the development of tools to use this Web archives in a relevant way (Meyer, Thomas & Schroeder, 2011). Indeed the final purpose is not only to be able to capture Web content, but also to preserve it, share it back and make it useful to users.

In the future, the use of the massive amount of information that Web archives contain will be diverse, and one can easily foresee that it will not only be about human directly reading but also about applying analytical tools. One can already see how new algorithms can help make sense of large amount of data, specifically when it comes to understanding the global shape and dynamic of information dissemination, influence networks, cultural differences in communication patterns etc. Web archives will be the key resources for supporting open research in this emerging field in the future. It’s the right time to get involved in creating this new generation of collections!

4 References

Books and studies
IIPC, International Internet Preservation Consortium (http://netpreserve.org/publications/reports.php) 2008 IIPC Member Profile Survey Results, by the IIPC Communications Officer
Meyer Eric T., Thomas Arthur, Schroeder Ralph: Web Archives: The future(s), Oxford Internet Institute and University of Oxford, June 2011 on behalf of the IIPC.
Wikipedia http://en.wikipedia.org/wiki/List_of/Web_Archiving_Initiatives

Presentations
European Conference on Digital Archiving, ECA 2010: Archivage du Web: perspective archivistique” by University of Montreal (Aïda Chebbi, 2010), “Catching the Web – Mirroring information society through website archiving” by Austrian Parliamentary Administration (Günter Schefbeck, 2010) and “Intervening early to ensure links persistence: How the National Archives’ Web Continuity project has helped to redefine What to keep?” (Amanda Spencer, 2010)
Projects

LivingKnowledge, http://livingknowledge-project.eu/ (funded by the European Commission under Project No. 231126)

Longitudinal Analytics of Web Archive data – LAWA, http://www.lawa-project.eu/ (funded by the European Commission under Project No. 258105)

Collect-All ARCHives to COMMunity MEMories – ARCOMEM, http://www.arcomem.eu/ (funded by the European Commission under Project No. 270239)


Living Web Archives – LiWA, http://www.liwa-project.eu/ (funded by the European Commission under Project No. 216267)

Large-scale, Cross-lingual Trend Mining and Summarisation of Real-time Media Streams – TrendMiner, http://www.trendminer-project.eu/ (funded by the European Commission under Project No. 287863)