From the World Wide Web to digital library stacks

Clément Oury, Sébastien Peyrard

From the World Wide Web to digital library stacks: preserving the French web archives

Clément Oury
Bibliothèque nationale de France
Legal Deposit Department
clement.oury@bnf.fr

Sébastien Peyrard
Bibliothèque nationale de France
Bibliographic and Digital Information Department
sebastien.peyrard@bnf.fr

ABSTRACT
The National Library of France is mandated by French law to collect and preserve the French Internet. It is now a 10-year old project with collections ranging from 1996 to the present. To ensure their long-term preservation, the choice has been made to ingest these web archives into the institution's existing digital preservation repository, SPAR (Scalable Preservation and Archiving Repository). There were numerous implementation challenges, on the modeling as well as the technical sides, which the library met with solutions drawn from international collaboration and widely adopted standards, whenever possible.

- Web archive-specific formats (W/ARC files) lacked validation and characterization tools, which led to the development of a Jhove2 module for the ARC format.
- The heterogeneity of BnF's web archives in terms of formats, production workflows and tools, was managed by aligning all of them on a single model, the current production workflow using NetarchiveSuite.
- The specificities of web archives were matched to the PREMIS data model and dictionary and SPAR's global METS profile.
- Finally, the need to express technical information about ARC files in a concise, manageable fashion led us to define a format-specific metadata scheme for container files, containerMD, which will be released to the preservation community (on BnF's website).

All this development work means new services for digital curators in general and preservation experts in particular. They will be able to know their collection better, to check its comprehensiveness, and, with that deeper understanding, to investigate new preservation strategies. Allowing differentiated service level agreements for specific sets of documents, with richer metadata extraction, better quality insurance and differentiated preservation strategies, will be the logical next step of the web archives long-term preservation project.

Keywords
Web archives; Metadata; Characterization tools; ARC file format.

1. The issue: ingesting the legal deposit of the French Internet in BnF's digital repository

1.1 The legal deposit mandate

As of August 1st, 2006, a copyright law gives the National Library of France the mandate to collect a new kind a resource: the whole set of data that is publicly available on the French Internet. This mandate has been given to the library thanks to an extension of legal deposit, which obliges every publisher to send copies of his output to the library. The Internet having obviously become the favorite place to create and distribute knowledge and information, it was necessary to give French heritage institutions the legal means to ensure its preservation[1].

Although the law was voted in 2006, the project of collecting French websites at BnF dates back to the early years of the last decade. In 2002 was launched the collection of all websites related to the presidential and parliamentary elections; this operation was renewed two years later, for European and then regional elections. These crawls were performed with a small-scale harvesting robot, called HTTrack[2].

The library was still lacking the technical means (hardware and software), skills and experience necessary to realize large-scale crawls of the French web. This is the reason why BnF agreed on a partnership with Internet Archive (IA), a not-for-profit foundation involved in world-wide web archiving since 1996. Thanks to this agreement, five annual broad crawls (from 2004 to 2008) of the .fr domain were performed by IA and enriched BnF's scollections [4]. They were performed by Heritrix[3], a harvesting robot developed by Internet Archive and several Scandinavian libraries in the framework of the International Internet Preservation Consortium (IIPC)[4].

Along with the results of the annual .fr crawls, Internet Archive delivered to BnF large extracts of its own collections, from 1996 to 2005. These so-called historical collections were not directly crawled by Internet Archive, but given to this institution by Alexa Internet [3]. In terms of value, these early collections may be compared to the first printed books.

At the same time, BnF was building an infrastructure (technical as well as organizational) to perform in-house crawls. The library decided to use Heritrix, and started by conducting focused crawls on a continuously growing number of websites and resources (from 130 million URLs in 2007 to 350 million in 2010).

Finally, in 2010, BnF launched its first in-house .fr domain crawl. To achieve this goal, NetarchiveSuite, developed by the Royal Library of Copenhagen and the University Library of Aarhus, was added on top of Heritrix[5]. This tool helps curators manage the harvesting workflow for very broad or very frequent crawls. Successfully tested on the .fr domain crawl in 2010, NetarchiveSuite is now used for all focused and domain crawls.

---

[1] About the legal aspects of Web archiving in France, see [2].

[2] HTTrack Website Copier website: http://www.httrack.com.

[3] Heritrix home page: http://crawler.archive.org.

[4] IIPC website: http://www.netpreserve.org.

[5] NetarchiveSuite website: http://netarchive.dk/suite.

Figure 1 : BnF web archives collections as of July 2010

Collection	Historical collections	First election crawls	IA crawls	In-house crawls without NetarchiveSuite	In-house crawls with Netarchive-Suite
Date	1996-2005	2002 and 2004	2004-2008	2006-2010	2010-…
Size	70 Tb	0.5 Tb	45 Tb	22 Tb	23 Tb
Operator	Alexa Internet	BnF	IA	BnF	BnF
Robot	-	HTTrack	Heritrix	Heritrix	NetarchiveSuite and Heritrix

In short, the harvesting process has been now fully internalized. Access to these web archives has been opened in BnF reading rooms. Ensuring their long-term preservation was a further step in order to achieve a complete library lifecycle, but two main issues arise in tackling this challenge:

- The size and variety of these collections make them invaluable and harder to preserve at the same time.
- BnF's existing digital repository, SPAR, was a natural choice for preserving our web archives, but some adjustments were necessary on both sides.

1.2 SPAR

Ingesting BnF's web archives in SPAR [1] is indeed an opportunity and a constraint at the same time.

The opportunity is great: the core preservation functions of the system have already been defined, developed and are up and running, which lowers implementation risks. Using the same preservation system for all the digital collections at BnF also has the benefit of being cost-efficient.

Apart from project and cost management issues, this is also clearly an opportunity from a librarian point of view. From its early stages, SPAR has been designed to manage heterogeneous digital documents with different preservation policies to be applied. It would be something of a waste not to use these features.

Finally, using a single repository solution for all kinds of digital documents in a single system seems more manageable over the long term: the distinction made between web archives and, say, born-digital acquisitions, can shift over time. Being able to manage them in a single system can make things easier later.

However, integrating the web archives with SPAR also has its constraints: there is a pre-existing data model [1], which could be adapted and enhanced, but not replaced by a new one; in addition, BnF's web archives are poorly described as there is currently no cataloguing of these collections, whereas the first ingested collections, of digitized books and audiovisual documents, are far better-known and described.

2. Implementation: from local issues to international cooperation

2.1 Knowing our collections: the Jhove2 modules

Before ingesting BnF's web archives, BnF digital curators should be able to know the technical characteristics of their collections and, thanks to this, what they can do with them in terms of preservation. The huge amount and variety of the harvested files, impossible to encompass directly, led us to concentrate for the moment on the container file levels, in particular the ARC file format[6], used for all the collections. It was vital to have tools that were able to validate and extract information about these files, *and* that allowed, at least, content files to be identified – and thus, multi-level file-format analysis features.

In order to achieve this goal, we decided to use the framework proposed by Jhove2[7]. However, this tool lacked an ARC-specific module; so it was necessary to develop one, along with a GZIP module to handle ARC GZ files, to have these features.

Apart from listing the properties to extract from the ARC files, the challenges that appeared at the design stages were mainly linked to the interpretation of the often ambiguous ARC specification.

First, we defined a unique, unambiguous terminology for the constituent parts of an ARC file:

– *Version-block*: the header and structure declaration of the file; comprises a *filedesc* (metadata about the ARC file) and a *URL-record-definition* (structure of the ARC records).

– *ARC record*: a specific entry of an ARC file, comprising a *URL record* (header for the ARC record) and a *network doc* (whatever the protocol returned to the crawler). This *network doc* is itself divided into a *protocol response* and an *object* (the harvested file).

We typically encountered difficulties in finding a way to manage the peculiarities of the ARC 1.1 format, an Internet Archive ARC profile with an XML metadata file inserted after the *filedesc*. Even though not referenced in the ARC specification, it was assumed to be compliant with it, since its author, IA, produces all its ARC files produced on this model since 2005. However, aligning this to our terminology was not simple: should this XML file be considered as a part of the *version block*, or as a particular *ARC record*? We combined the two, considering the *version-block* as a header built on the structure of an ARC record with an XML file as a possible content *object*, as can be seen on figure 2.

The other major problem was handling the gzip compression of an ARC: whereas a gzip compression is typically applied after the file has been created, Heritrix directly interlaces the two formats when creating arc.gz files: the version block and first ARC record are respectively the first gzip and second gzip members at the same time. Jhove2 therefore had to be able to process an arc.gz file simultaneously with the gzip *and* ARC modules. We modelled this as an ARC structure with a gzip encoding:

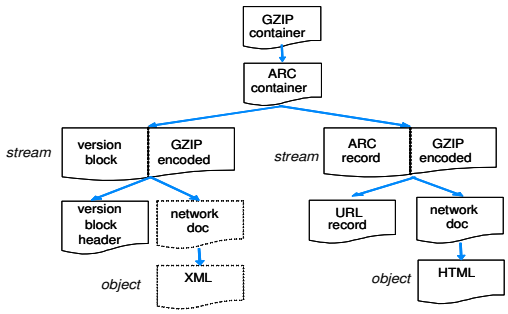

Figure 2. Structure of an arc.gz file according to Jhove2

[6] An ARC file is a container file aggregating each file harvested on the Web in a dedicated ARC record. For technical reasons, the size of an ARC file is generally limited to 100 Mb. Cf. http://www.archive.org/web/researcher/ArcFileFormat.php.

[7] Jhove2 website: http://www.jhove2.org.

2.2 Aligning the heterogeneous web archive collections: the NetarchiveSuite target

As explained in 1.1, BnF currently uses NetarchiveSuite, or NAS, for all its crawls, and the data harvested thanks to NAS represent the only growing part of BnF web collections. These are the two reasons why NetarchiveSuite data structure and data model have been chosen as a target to organize all our other collections of web archives in SPAR. For example, all the metadata describing Heritrix crawling process (configuration, log and report files from the crawler, or CRL) are packaged by NAS into ARC "metadata files", where each CRL file represents an ARC record within the ARC container file. We applied this rule to the CRL files coming from other harvesting channels whenever possible; we do not for example have any of them for the "historical" collections.

Another critical choice was related to the collections data model. The data coming from NAS are organized in three layers of granularity:

- At the base there is the ARC container file.
- Each ARC file is part of a specific "harvest instance", or "job". In our domain-specific terminology, a job is "a particular harvest process, realized by a crawling machine and monitored by a human engineer, which produces a set of data and metadata ARC files, and that has a beginning and an end". Each job is launched on a list of seeds (a seed is a URL from where the robot will start its crawling process), with defined parameters that will order the robot to conform to certain rules[8].
- On the top, the "harvest definition" is globally equivalent to a collection. A harvest definition groups all the jobs that have been launched in order to achieve the same purpose. For example, the aim of "performing a .fr domain crawl" is achieved thanks to hundred of jobs, each of them grouping thousands of domains. The harvest definition "news websites" launches every day a crawl of around 80 seeds – i.e. there are 365 jobs a year to achieve this harvest definition.

These three layers match three kinds of information packages in SPAR: ARC "data" files are ingested as "web data" information packages; ARC metadata files (that contain information at the job level) are ingested as "harvest metadata", and harvest definitions are ingested as OAIS Archival Information Collections.

In order to homogenize our collections to a certain extent, we decided to use this three layered data model everywhere, which can sometimes be artificial. For example historical collections only have two layers: the ARC files and the harvest definition. There is no layer for the job, as the data given to BnF has not been crawled, but extracted from a larger web archives collection. However, in order to maintain homogeneity, we virtually created a third layer. We declared that all the harvest definitions of the historical collections had been produced by a single harvest instance, or job.

2.3 From web archive concepts to PREMIS

Even if PREMIS is conceived as core preservation metadata and web archive-specific concepts are clearly out of its scope, it is a cornerstone of the SPAR data model [1] so we had to know where our web archive concepts fit in the PREMIS data model. Here again we encountered some difficulties.

The job. We have defined in 2.2 what a job is. However, if we try to fit this "job" concept in PREMIS, it can match three different entities depending on what you consider. It is a typical Event since it has a beginning and end date, produces Objects (ARC files) and has Agents (software, administrators…) involved in it. But it is also an Object, if we use this term to designate the result of a crawl. In addition, the job is also a set of parameters, given to a crawler at a certain time and impacting the crawling event and produced objects. From this standpoint, a job can be viewed as an Agent that activates a crawl.

These ambiguities forced us to adopt a clearer, PREMIS-compliant terminology:

- The Event is a **harvest** eventType.
- The Objects produced by this harvest event are **harvest instances**. They typically consist of at least one ARC metadata file and many ARC data files.
- The Agent activating a harvest Event launching a crawler is a **job**. Since it is currently impossible to track accurately – as the parameters can be changed by an administrator during a crawl – it was considered out of the scope of our digital repository, so we merely kept track of it as a linking Agent of the harvest Event. However we kept track of two key parameters: the user agent and the robots policy.

The user agent is an identity under which the crawler declares itself, typically a particular web browser, e.g. "Mozilla 5.0". We modelled this as a distinct Agent involved in the Event, because the same crawler could declare itself under different identities.

The robots policy is the behaviour of the crawler towards the robots.txt protocol (comply with it or ignore it). We considered it as a particular outcome of the harvest; this debatable choice was made in want of a current PREMIS field for "input" information.

The reports on the produced ARC files and crawled hosts were typical outcomes of the harvest Event, documented in Extensions.

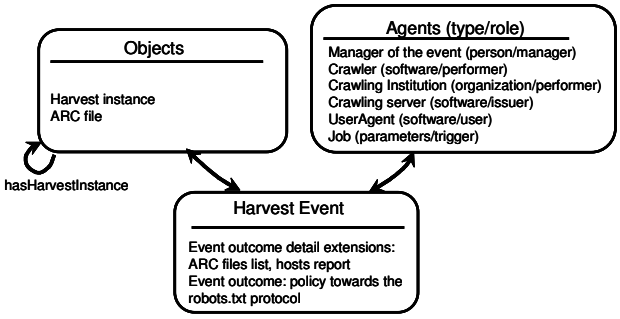

Figure 3. Aligning web archiving concepts with PREMIS

2.4 From core to domain-specific preservation metadata: the containerMD solution

Having validation and extraction tools for ARC and arc.gz files was not enough: we also had to define how to record this format specific information in our AIPs. While we had the XML output of Jhove2, it was considered far too verbose to be manageable[9].

[8] For example: do only crawl URLs in a given list of domain names, do not follow too many clicks from a seed URL…

[9] Even with mere identification of the content files, the Jhove2 XML output for an ARC file could sometimes exceed the ARC file size itself. For the typical 100 Mb ARC file, this was considered too heavy to handle (processing, rendering, etc.).

Core information about a file and its content files could be modelled as premis:file containing premis:files[10]. However, some ARC-specific features needed dedicated fields to be recorded; and, above all, PREMIS has been designed on an all-or-none principle: it is necessary to choose between describing just the container file and describing all the contained files individually.

No characterization schema for container files being available to the community yet, we felt there was a gap to be bridged and we therefore designed containerMD[11]. Its key features are a description of the container file itself in a <container> section and two verbosity levels:

– A "non verbose mode": aggregated information about the entries in an <entriesInformation> section;

– A "verbose mode": dedicated description for each entry in an <entry> section, with the ability to include other characterization schemes if needed.

The <container>, <entriesInformation> and <entry> all have an extension section for format-specific fields, with only the ARC format for the moment. Each content *object* is thus referenced as an entry, with additional information about the ARC record being embedded in the ARC-specific extension section.

Figure 4. Aggregation methods in containerMD

	verbose mode: <entry>	**non verbose mode: <entriesInformation>**	**Reduction method**
existence of an entry	One <entry> per entry	number attribute	Count
creation dates	creationDateTime attribute	firstDateTime and lastDateTime attributes	Only the min and max values are kept
entry size	<fixity>: size attribute	minimumSize and maximumSize attributes	Only the min and max values are kept
format of a content file	<format>: name and version attributes	<formats> container element For each format name and (if any) version: one <format> element with name and version attributes; number attribute counting the corresponding entries; globalSize attribute for the size of all the corresponding entries.	Aggregation and count Sum
encodings at entry-level	<encoding>: type (encryption or compression) and method attributes	<encodings> container element For each encoding type and method, one <encoding> element with type and method attributes	Aggregation
ARC record host	<host>	<hosts> container element For each <host>, number attribute; globalSize attribute for the corresponding entries	Aggregation and count Sum
ARC record declared MIME type	<declaredMimeType>	<declaredMimeTypes> container element For each <declaredMimeType>, number attribute; globalSize attribute for the corresponding entries.	Aggregation and count Sum
ARC record protocol information	<response>: protocolName and protocolVersion attributes	<responses> container element For each <response> with a particular protocolName and (if any) protocolVersion: number attribute; globalSize attribute for the corresponding entries.	Aggregation and count Sum

3. Conclusion: future usages and evolutions

In the end, the ingest of the BnF web archives in SPAR will allow us to build **new curation services for the web harvesting team**:

- **Getting better file formats statistics** on the type of files (text, image, video…) harvested: currently we still use the MIME type sent by the server, which is often unreliable. Using Jhove2 and storing the results in the containerMD <formats> section to be queried will improve this knowledge.
- **Knowing the content of older collections**. The distribution of the data per host is also some key information for web archives. This information is compiled in files called hosts-reports for current harvest instances, but not for historical collections. Jhove2 will be able to calculate a host-report per ARC file, which may later be aggregated at upper levels.
- **Checking collection comprehensiveness**. Each ARC metadata AIP contains a list of all ARC files produced by the harvest instance, as the outcome of a harvest event. Automatically comparing such lists with the ARC data files actually ingested in SPAR may prove very useful with old collections, for which there is a risk that we have lost data.

All this generated AIP metadata will also help us define indicators and **investigate preservation strategies**. Some metadata elements can be used to define risk assessment criteria, e.g. the format of the container file and its content objects, the rendering tool intended for the harvested files (given by the user agent), or the status of a given harvest (finished, terminated or crashed). This will help us define subsets of our collection on which focused preservation actions could be performed: format migration for the container files or not; emulation vs migration of the harvested files, and so on.

Finally, one of the great strengths of SPAR being its ability to manage different collections with different service level agreements, one may imagine **applying differentiated SLAs** to collections that, even though produced by the same harvesting infrastructure, do not share the same preservation policies. For example, if we negotiate with a publisher to harvest PDF online editions provided that they comply with a specific PDF profile, we will be able to ask SPAR for stronger validation procedures to check that these files conform to the negotiated format. In the end, defining differentiated preservation actions and services on our web archives seems to be the next great challenge to take up.

4. REFERENCES

[1] Fauduet, L., Peyrard, S., A data-first preservation strategy: data management in SPAR, *Proceedings of the 7th International Conference on Preservation of Digital Objects (iPRES)*, 2010. Online: http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/fauduet-13.pdf (accessed September 2011, 30th).

[2] Illien, G., Sanz, P., Sepetjan, S., Stirling, P., The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future, *Proceedings of the 77th IFLA General Conference and Assembly*, 2011. Online: http://conference.ifla.org/sites/default/files/files/papers/ifla77/193-stirling-en.pdf (accessed September 2011, 30th).

[3] Kimpton, M., Braggs, M. and Ubois, J. 2006. Year by Year: From an Archive of the Internet to an Archive on the Internet. In *Web Archiving*, J. Masanès, Ed. Springer, Berlin, Heidelberg, New York.

[4] Lasfargues, F., Oury, C., Wendland, B., Legal deposit of the French Web: harvesting strategies for a national domain, *Proceedings of the 8th* International Web Archiving Workshop (IWAW), 2008. Online: http://iwaw.net/08/IWAW2008-Lasfargues.pdf (Accessed September 2011, 30th).

[5] Oury, C., « Large-scale collections under the magnifying glass: format identification for web archives », *Proceedings of the 7th International Conference on Preservation of Digital Objects (iPRES)*, 2010.

[10] Cf. *Data dictionary for Preservation Metadata: PREMIS version 2.1*, p. 45. Online: http://www.loc.gov/standards/premis/v2/premis-dd-2-1.pdf.

[11] For more information on containerMD, see http://bibnum.bnf.fr/containerMD.