

Putting it all together: creating a unified web harvesting workflow at the Bibliothèque nationale de France

Annick Le Follic
Peter Stirling
Bert Wendland

Abstract

This article was written for the IIPC-sponsored workshop *How to fit in? Integrating a web archiving program in your organisation*, held at the Bibliothèque nationale de France, Paris from 26th to 30th November 2012. The workshop was intended to deal with organisational issues such as advocacy, strategy and communication, based mainly on experiences at the BnF, with contributions from other institutions.

The article was co-written by two digital curators and a crawl engineer, representing the fact that web archiving activity at the BnF is divided between the Legal Deposit and IT departments. The aim of this article is to present the complete web harvesting workflow at the BnF, with the different tools that are used at each stage and the different profiles of the staff involved.

| | |
|---|-----------|
| 1. INTRODUCTION | 3 |
| 1.1. PRESERVING THE FRENCH WEB: THE MIXED MODEL OF HARVESTING..... | 3 |
| 1.2. CHOICE OF NETARCHIVESUITE: STANDARDISATION AND FORMALISATION OF PROCEDURES.. | 4 |
| 1.3. THE BNF WEB ARCHIVING WORKFLOW..... | 5 |
| 2. PREPARING THE HARVEST | 6 |
| 2.1. BNF COLLECTE DU WEB: A SELECTION TOOL FOR LIBRARIANS | 6 |
| 2.1.1. <i>Structure and organisation of BCWeb</i> | 7 |
| 2.1.2. <i>Adding URLs to BCWeb</i> | 7 |
| 2.1.3. <i>Modification and management of URLs in BCWeb</i> | 8 |
| 2.1.4. <i>Transfer of URLs to NetarchiveSuite for crawling</i> | 8 |
| 2.2. NAS_PRELOAD: A SPECIFIC TOOL TO PREPARE THE BROAD CRAWL..... | 9 |
| 2.2.1. <i>Aggregation of a large number of sources</i> | 9 |
| 2.2.2. <i>Domain Name System lookup</i> | 10 |
| 2.2.3. <i>Retrieval of redirected domains</i> | 10 |
| 2.2.4. <i>Transfer to NetarchiveSuite</i> | 10 |
| 2.3. NETARCHIVESUITE | 11 |
| 2.3.1. <i>Role and users of the application</i> | 11 |
| 2.3.2. <i>Three main modules: harvesting, archive, access</i> | 11 |
| 2.3.3. <i>Data model: domain, configuration, schedule, harvest definition</i> | 12 |
| 3. HARVESTING..... | 13 |
| 3.1. HERITRIX | 13 |
| 3.2. MONITORING WITH NETARCHIVESUITE | 15 |
| 3.3. VIRTUAL SERVERS FOR A FLEXIBLE INFRASTRUCTURE..... | 15 |
| 3.4. QUALITY ASSURANCE..... | 16 |
| 3.4.1. <i>International quality indicators</i> | 16 |
| 3.4.2. <i>Production of statistics with NAS_qual</i> | 17 |
| 3.4.3. <i>Some metrics used at the BnF</i> | 17 |
| 3.4.4. <i>Quality assurance using NetarchiveSuite</i> | 18 |
| 4. POST-HARVEST PROCESSING | 18 |
| 4.1. THE INDEXING PROCESS | 18 |
| 4.2. ACCESS TO THE COLLECTIONS: “ARCHIVES DE L’INTERNET” | 21 |
| 4.3. SPAR: THE BNF DIGITAL REPOSITORY | 22 |
| 5. CONCLUSION | 24 |

1. Introduction

The Bibliothèque nationale de France (BnF) performed its first in-house domain crawl in April 2010. This marked an important turning point, as all web harvesting activities were now performed in-house. To achieve this the Library had to put in place a hardware and software infrastructure that allowed it to handle the volumes of data involved in a crawl of some 2 million domains, but that would be flexible enough to manage crawls of different kinds, as required by the mixed model of web harvesting implemented by the BnF to fulfil its obligations under French legal deposit legislation. This infrastructure had also to work with the organisational framework that has evolved within the Library, involving many different actors spread across different departments. This article describes the technical choices that allow the BnF to perform both large-scale and selective crawling, the tools that are used and the organisational structure in place, with the aim of presenting the whole production workflow.

1.1. *Preserving the French web: the mixed model of harvesting*

Legal deposit has existed in France since 1537, when François I created a legal obligation for all printed books to be deposited at the Royal Library, later to become the National Library. Over the years legal deposit law has evolved to include other forms of publication to ensure that all forms of cultural heritage are preserved for posterity. The most recent extension of legal deposit, in 2006, covers material published in digital form on the Internet. Under the Code du Patrimoine (Heritage Code) the BnF thus now has a legal mission to collect and preserve the French Internet, with the exception of sites related to television and radio which are collected by the INA (Institut national de l'Audiovisuel). In the tradition of legal deposit, the aim is not to judge the value of what is collected or to preserve only the “best” material, but rather to collect and preserve everything without distinction, or as this is in practical terms impossible in relation to Internet material, to preserve a sample that is as complete and as representative as possible¹.

To fulfil this mission, the BnF has adopted a “mixed model” combining annual broad or domain crawls with focused or selective crawls. These crawls are thus very different in nature:

- **Broad crawls**² are performed once a year, on a very large number of domains (over 2 million in 2012), with a limited budget (usually 10,000 URLs, or files, per domain). The broad crawl consists mainly of domains in .fr as the BnF has a contract with the Association française pour le nommage Internet en coopération (AFNIC), the registry for the .fr top level domain (TLD). In 2012 domains in .nc (for New Caledonia) were added thanks to an agreement with the Office des postes et télécommunications de Nouvelle-Calédonie (OPTNC). In previous years the robots exclusion protocol (the robots.txt file) was respected for the broad crawl, however in 2012 it was decided to no longer respect robots.txt;
- **Selective crawls** are performed at different frequencies (from daily to once a year), on a much smaller number of domains. Seed URLs are chosen by subject librarians at the BnF, and some external partners, such as the network of regional libraries (BDLI) and

¹ For more details on the legal framework and the definition of the scope of French internet legal deposit, see “The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future”, ILLIEN G., SANZ P., SEPETJAN S., STIRLING P. In: *IFLA journal*, 2012, vol. 38, n° 1. [http://www.ifla.org/files/hq/publications/ifla-journal/ifla-journal-38-1_2012.pdf]

² For more details, see “Legal deposit of the French Web: harvesting strategies for a national domain”, LASFARGUES F., OURY C., WENDLAND B. In: *Proceedings of the 8th International Web Archiving Workshop*, Aarhus, Denmark, 2008. [<http://iwaw.net/08/IWAW2008-Lasfargues.pdf>]

other partners such as researchers or associations. Selective crawls can have a bigger budget than broad crawls, going up to several hundred thousand URLs per domain.

Selective crawls had been performed by the Library since the first experiments in 2002, but the first five domain crawls, 2004-2008, were performed for the BnF by Internet Archive. However, as the legal deposit mission is fundamental to the BnF, it was vital to integrate these activities into the organisational structure of the Library, which involved not only management and organisational questions to define the different actors and their roles, but also technical questions of how to implement all aspects of web archiving activity at the BnF.

The 2010 broad crawl was the first to be performed entirely in-house, and marked the moment that the complete harvesting workflow was handled internally at the Library. To allow this the BnF had put in place a technical infrastructure that would allow it to handle simultaneously both the large volumes of data involved in the broad crawl, and the different frequencies and technical settings necessary for selective crawls. Just as importantly, the different applications and tools that have been put in place relate to the organisational structure and the different roles and responsibilities of each actor in the workflow.

1.2. Choice of NetarchiveSuite: standardisation and formalisation of procedures

Implementing a complete harvesting workflow in-house enabled the BnF to better control crawl configurations, monitoring and collection building policy to ensure quality in relation to overall Library policy and budgetary constraints. It also allowed the Library to develop and maintain skills in managing big data and associated workflows. The BnF already had experience of using Heritrix as a stand-alone crawler to perform selective crawls since 2006. However, in order to perform the first broad crawl an additional software layer that could programme and plan large-scale crawls was necessary. In addition, the aim was to automate crawl preparation processes and include selective crawling in the same technical solution, to maintain the entire production process within the same workflow.

The BnF chose NetarchiveSuite (also known as NAS), which was originally developed by netarchive.dk, a joint web archiving project of the two national deposit libraries in Denmark (the Royal Library in Copenhagen and the State and University Library in Aarhus) where it has been used since 2005. NetarchiveSuite was open-sourced in 2007, and the BnF and the Austrian National Library joined the project in 2008³. The harvesting workflow put in place for the Danish web archive was very similar to the model envisaged by the BnF, combining large-scale broad crawls with librarian-curated selective crawling at variable frequencies. NetarchiveSuite was therefore well suited to the needs of the BnF and was chosen as the central part of the production workflow. The BnF had to develop some features to suit its needs; these developments have been integrated into the main NetarchiveSuite software, and the BnF remains active in its development as an open-source project.

Another strength of the NetarchiveSuite software was the possibility to define functional responsibilities depending on the different roles within the web archiving team. At the BnF web archiving activity is organised by the legal deposit department, where a team of five digital curators works in partnership with IT experts on one hand and collection experts on the other hand. As described below, the responsibility for different functions in NetarchiveSuite represents these different roles.

The work of internalising the broad crawl and transferring the selective crawl to NetarchiveSuite, thus creating a single workflow, allowed the BnF to formalise its organisation

³ For more information, see <https://sbforge.org/display/NAS/NetarchiveSuite>.

and create a flexible and adaptable harvesting procedure that can handle the different types of harvest necessary for the BnF to fulfil its mission of Internet legal deposit.

1.3. The BnF web archiving workflow

The following diagram represents schematically the different parts of the BnF web archiving workflow.



The workflow is shared between three areas of activity, performed by staff from different departments of the Library. In the beginning is a **selection** of websites to be archived. This is done by *librarians* from all departments, represented in yellow in the above schema. The selected websites are then **validated** by a *digital curator team* from the legal deposit department (orange in the schema) to ensure that the websites are within the legal scope of the crawl and are technically collectable, followed by a **planning** of how and when to harvest them. The **harvesting** itself is managed in the *IT department* (red in the schema), which is also in charge of **indexing** the harvested data and of ingesting it into the long-term **preservation** system.

Some tasks are performed jointly by the legal deposit and IT departments as these tasks have both collection and technical aspects: **monitoring** the crawl jobs, **quality assurance** after the end of a harvest and the indexing process, **experience** evaluation, and making the data **accessible** for consultation by users. Experience of how to harvest a website in a better way or why a harvest did not achieve the expected result, from the evaluation of monitoring and quality assurance, may influence the further validation of websites. Finally, the librarians may review their selection by consulting the archive.

The activities in the workflow are based on certain applications. **NAS_preload** is used to analyse and insert the data for broad crawls into the database. Selection and validation of data for selective crawls is done by means of **BCWeb**. The central application in the workflow is **NetarchiveSuite**, used for planning the crawls, for creating and launching jobs according to predefined schedules and for monitoring, quality assurance, and experience evaluation. **Heritrix** is the crawl robot. The **Indexing Process** and **NAS_qual** perform post processing after the harvest. The long-term preservation is ensured by the **SPAR** system of the BnF. The **Wayback Machine** gives access to the archived data.

Some applications – NetarchiveSuite, Heritrix, and the Indexing Process – run on a virtualised environment which is based on **VMware** products to obtain a better performance, a better management of physical resources of the machines, and a more flexible administration of functional aspects in the operations.

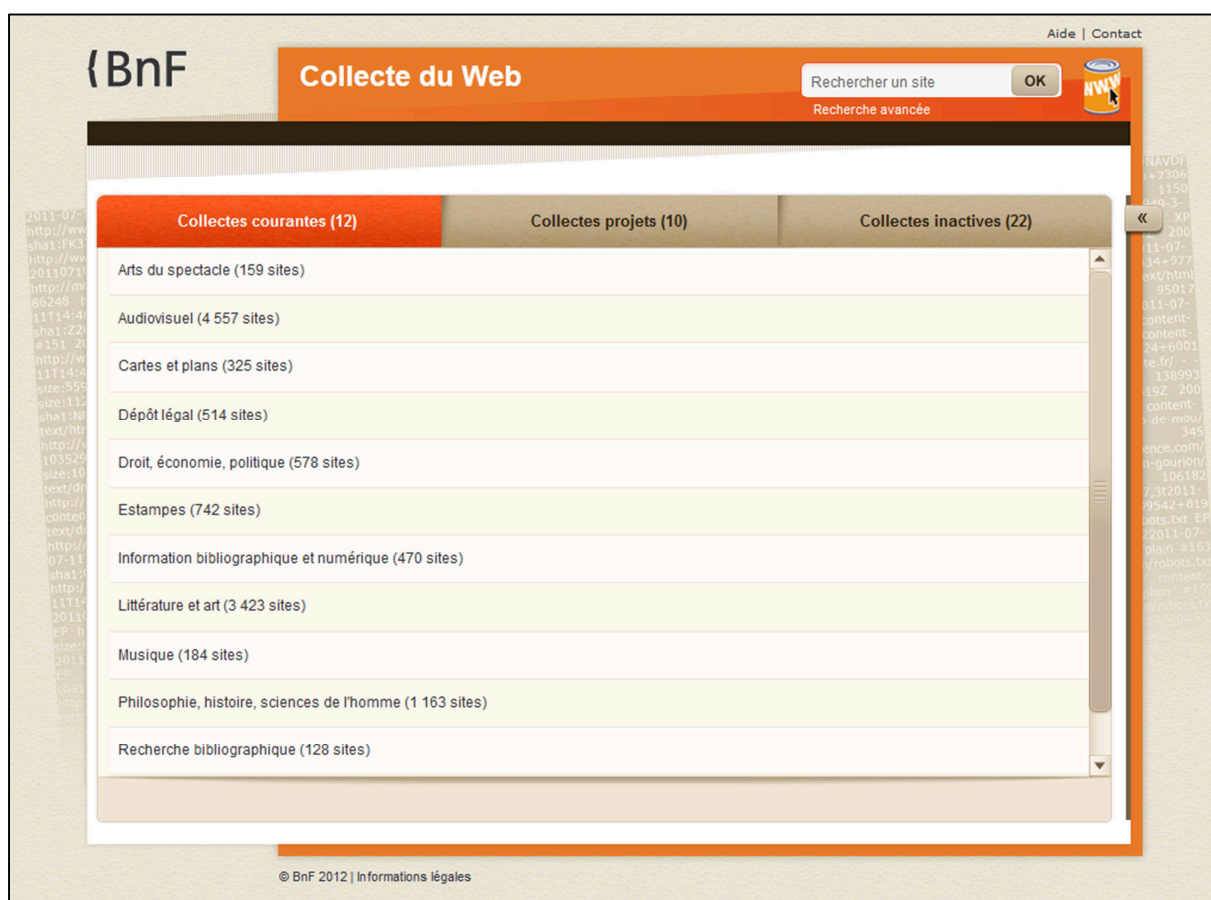
Where possible, the BnF has chosen open-source solutions and aims to contribute to their development: this is the case for Heritrix, the Wayback Machine and NetarchiveSuite. A selection tool (BCWeb) has been developed in-house. Finally, web archives are ingested into SPAR, a system which is being developed in-house for the preservation of all BnF digital collections of the National Library.

These applications are described in detail in the following sections of this paper.

2. Preparing the harvest

2.1. *BnF Collecte du Web: a selection tool for librarians*

The selection of sites for selective crawls is performed by a network of around 80 content librarians in the BnF, with in addition external partners for some projects. To give them the possibility to propose and manage URLs, the BnF has developed an application called *BnF Collecte du Web* (BCWeb), which can be translated as “Building Collections on the Web”. BCWeb is a web application that is accessible via a standard web browser within the BnF and via a secure connection for external users. The aim of this tool is to facilitate the work of content librarians, and also to provide more visibility on what URLs are collected.



2.1.1. Structure and organisation of BCWeb

BCWeb is organised into “collections”, which correspond to the organisation of selective crawling at the BnF: each department has its own collection, while there are also collections for project crawls. These collections are created and defined by the BCWeb administrators (the digital legal deposit team) to allow certain technical settings depending on the type of collection.

Each content librarian has a user account for BCWeb. For those at the BnF, this is linked to the Library’s LDAP user directory, permitting a single login identical to other BnF applications. External users log in using a secure connection, with login details created by the digital legal deposit team. Each user account is associated with one or more collections, so that users have the right only to enter and modify URLs in the collections on which they work. However all the collections can be consulted by all BnF staff, which allows managers to have an overview of the work of their teams and gives other library staff an idea of the content of the selective crawls.

URLs are chosen according to the collection policy of each department or project, and may correspond to a whole website or a part of a site. The URLs in each collection are stored in the form of “records”, which contain three kinds of information:

- *management information*: mostly administrative information such as the date of creation and last modification, the name of the person who created the record;
- *technical settings*: these include the seed URLs and the settings that determine how they will be collected;
- *description*: information which allows records to be classified and searched.

As of November 2012, BCWeb contains over 30,000 records in total, with the size of collections ranging from around 100 records to over 4,000.

2.1.2. Adding URLs to BCWeb

To enter a URL into BCWeb the librarian creates a record using a two-stage process.

- The first stage performs two kinds of test on the URL: An online verification process confirms whether or not the URL is valid: if the server returns an HTTP response in the range of 400 or 500 or if there is no DNS entry, the URL is considered invalid and cannot be selected. In addition, redirections (HTTP response in the range of 300) are indicated to allow the selector to choose the URL which will give the best quality crawl.
- A second verification within the BCWeb database shows whether the URL has already been entered, or whether a similar URL has been entered (within the same domain or host). This should prevent URLs being entered multiple times, however, if necessary the same URL can be entered in different collections.

The second stage is to fill in the different kinds of information outlined above.

Management information

The date of creation and the person who creates the record are filled in automatically. For any update, the date of modification and the name of the user are also added. Another field in this section contains the name of the content librarian who is “responsible” for the record; in large collections this means users can quickly find the URLs on which they work.

Technical settings

The three main technical settings – budget, frequency, depth – define the manner in which the URL will be crawled by the robot, and therefore define the target. They also give the elements for the digital legal deposit team to organise the crawls, by grouping together URLs by size and frequency:

- Budget: this setting defines the number of URLs, or files, collected from the seed URL. The options are usually “small” (less than 50,000 URLs), “medium” (10,000 – 100,000 URLs), or “large” (over 100,000 URLs).
- Frequency: this setting defines a period adapted to the rate of change of a website. The usual settings are weekly, monthly, twice-yearly and annual.
- Depth: this setting tells the robot how far to go from the seed URL; depth ranges from one “mouse click” (for example, to collect the headline articles of an online news site) to the whole domain.

These three settings are combined to define the crawl configuration. Certain combinations are impossible to use in practice: for example, a large site cannot be collected in its entirety every day. BCWeb prevents users from selecting incompatible values, and indicates visually which values may be combined.

The technical settings also include the possibility of adding supplementary URLs to improve the quality of the collect, for example by including the sitemap. Finally, a record can be made “inactive” which means it will not be transferred to the robot for crawling; this is usually in the case of URLs which no longer exist online, or which are no longer pertinent for the collection.

Description

The description section contains additional information to organise and follow up the observation of chosen websites. The “theme” field is mandatory and classifies records in each collection. The “keyword” field is optional and contains terms describing the site. Finally, two notes fields (content and technical notes) allow librarians to add any other information.

2.1.3. Modification and management of URLs in BCWeb

Records can be modified by all users who work on the same collection. They can also be transferred from one collection to another.

BCWeb also allows users to perform searches, for example to find only the records for which they are responsible, or a subset of records which they want to check. Searches which a librarian often performs can be saved to their account, to be rapidly accessible. From the list of search results it is possible to modify a set of records simultaneously, for example to transfer the responsibility to a new librarian. Users can also export results as a CSV file.

2.1.4. Transfer of URLs to NetarchiveSuite for crawling

BCWeb communicates directly with NetarchiveSuite, transferring the selected URLs directly to the workflow for subsequent crawls. Administrators, i.e. the digital legal deposit team, use a separate section of BCWeb to define the settings for the transfer of URLs into NetarchiveSuite, by establishing a mapping between the technical settings in BCWeb (budget, frequency, depth) and the equivalent settings in NetarchiveSuite. Each “collection” in BCWeb is also mapped to a Harvest Definition in NetarchiveSuite.

The transfer can be performed as needed, for instance the records with a “monthly” frequency will be transferred just before the crawl is launched each month, whereas “annual” records will be transferred just before the annual crawl.

2.2. NAS_preload: a specific tool to prepare the broad crawl

At the BnF, the broad crawl is launched once a year with a large number of URLs from different sources. The challenge of the broad crawl is to collect a representative sample of the national domain and to illustrate the French production on the web at the time of the harvest. Crawlers harvest content without any distinction between its academic, institutional, commercial, or any other background, as defined in the legal deposit mission of the BnF. But it is equally important to find a way to record and to freeze a moving space as a snapshot of the French web.

This type of crawl is a less expensive approach than focused crawling in terms of the costs of the harvest (machines and humans) in relation to the retrieved amount of data. This also concerns the process of preloading the sources. The different sources involved made it necessary to develop a specific tool to prepare the seed list to be imported into NetarchiveSuite. This tool, called “NAS_preload”, is a command line Java application which is launched manually by the crawl engineers.

The in-house development of NAS_preload started in 2010 for the first in-house broad crawl at the BnF. The goal at that point was to transfer the domain list (1.6 million domains in .fr and .re) directly into the NetarchiveSuite database. In 2011, it was adapted with the addition of other sources (selections of URLs from the BnF librarians, lists of URLs from other library workflows) and with the need for specific treatments on these sources: de-duplication, DNS checks, analysis of HTTP response code to identify redirections, and statistics. Finally, in 2012, the processes were optimised.

2.2.1. Aggregation of a large number of sources

The first step of these operations is to ingest different sources. In 2012 these sources were as follows:

- 2.4 million domains in .fr and .re, provided by AFNIC (Association française pour le nommage Internet en coopération – the French domain name allocation authority);
- 3,000 domains in .nc, provided by OPT-NC (Office des postes et télécommunications de Nouvelle-Calédonie – the office of telecommunications of New Caledonia);
- 2.6 million domains already present in NetarchiveSuite database;
- 13,000 domains from the selection of URLs by BnF librarians (in BCWeb);
- 6,000 domains from other workflows of the Library that contain URLs as part of the metadata: publishers’ declarations for books and periodicals, the BnF catalogue, identification of new periodicals by librarians, print periodicals that move to online publishing, and others.

The script used to ingest these sources also gives the possibility to mark each domain or each URL to produce statistics on their origin. That means that we can measure the importance of the librarians’ selection in identifying domains not present in other sources, or the benefits of including the list of publishers’ declarations.

These statistics show how many new domains were discovered in 2012 from these different sources, respectively:

- 650,000 for AFNIC;
- 3,000 for OPT-NC;
- 600 for librarians’ selection;
- 1,800 for other BnF workflows.

This model has proved to be well suited for BnF purposes and will be reused in 2013.

At the end of this step, these sources are de-duplicated to generate a list of unique domains: the total is around 3.3 million unique domains.

2.2.2. Domain Name System lookup

The second step of NAS_preload is to define which domains are inactive with the aim of excluding them from the harvest. Excluding inactive domains can avoid slowing down the crawl, which is important as creating a snapshot of the French Internet requires “temporal coherence”, i.e. ensuring that the content is collected at the same time, or at least over as short a time as possible.

The DNS (Domain Name System) is a service used to establish a correspondence between the IP address of a computer and a domain name. A negative response usually indicates that the name has been reserved with a registrar but no service has yet been created. A script of NAS_preload goes through the list of domains to check the DNS responses, with as many concurrent outbound connections as possible for the BnF bandwidth and for the DNS servers.

Measuring the activity of a site is a recurrent issue among web archiving institutions in different countries. Of all the sites listed at the beginning of the snapshot, how many are there that really have content? Can a site with a single page be considered as active? Should we include mirror sites? Should we exclude parking sites? At the moment, the BnF excludes only the domains that have a failed DNS lookup, because it is certain that these sites do not exist at the moment of the crawl. However, it has been decided to maintain sites with only one page, mirror sites and parking sites because they represent the nature of the web.

The results for 2012 show that of the 3.3 million unique domains, approximately 1 million have a failed response:

- 250,000 domains had been registered at AFNIC but did not exist on the web. NAS_preload did not transfer these domains to NetarchiveSuite.
- 770,000 domains were already present in NetarchiveSuite database and NAS_preload modified their configuration in NetarchiveSuite to exclude them from the harvest.

2.2.3. Retrieval of redirected domains

NAS_preload also includes a script to identify new domains using HTTP response codes for redirections. For each domain, it creates two seeds, with and without *www*, to which are added any specific URLs present in the different sources. If the HTTP request gives a response code indicating a redirection (type 301, 302, 303, 304), the name of the new domain is added to the list in NAS_preload if it is not already present. This operation is useful as a way of discovering French websites on TLDs other than .fr, mostly on .com, .net, .eu and .de.

2.2.4. Transfer to NetarchiveSuite

The last step is to load the final lists of domains and of seed URLs from the NAS_preload database into NAS. New domain names, that is those that are not already present in the NAS database, are created with their list of seed URLs and their default configuration, while for existing domains the previous seed lists and configurations are updated. In 2012, NAS_preload transferred 2.3 million active domains along with 4.6 million associated seed URLs. The tool also changed the configuration of the 777,000 inactive domains (DNS failed) already present in the NetarchiveSuite database so that nothing will be harvested for these domains.

2.3. NetarchiveSuite

The screenshot shows the 'Edit domain' interface for the domain 'bnf.fr'. The interface is divided into several sections:

- Menu:** A sidebar on the left with a 'Menu' header and a list of navigation options including 'Definitions', 'Selective Harvests', 'Snapshot Harvests', 'Schedules', 'Find Domain(s)', 'Create Domain', 'Domain Statistics', 'Alias Summary', 'Edit Harvest Templates', 'Global Crawler Traps', 'Extended Fields', 'Harvest status', 'Quality Assurance', and 'Systemstate'.
- Domain information:** Fields for 'Domain name: bnf.fr' and 'Comments:'.
- Configurations:** A table listing various configurations with their names, descriptions, and 'Default' status.
- Seed lists:** A table listing seed URLs for various configurations.
- Crawler traps:** A section at the bottom for managing crawler traps.

| Configurations | Default |
|---|---------------------------------------|
| BnF_collecte_courante_annuelle_moyen_hote (note, [BnF_collecte_courante_annuelle_moyen_hote]) | Edit <input type="radio"/> |
| BnF_collecte_courante_annuelle_petit_chemin (chemin, [BnF_collecte_courante_annuelle_petit_chemin]) | Edit <input type="radio"/> |
| BnF_collecte_courante_annuelle_petit_hote (note, [BnF_collecte_courante_annuelle_petit_hote]) | Edit <input type="radio"/> |
| BnF_collecte_courante_annuelle_petit_page+1 (page+1, [BnF_collecte_courante_annuelle_petit_page+1]) | Edit <input type="radio"/> |
| BnF_collecte_courante_mensuelle_petit_page+2 (page+2, [BnF_collecte_courante_mensuelle_petit_page+2]) | Edit <input type="radio"/> |
| BnF_collecte_courante_semestrielle_petit_hote (note, [BnF_collecte_courante_semestrielle_petit_hote]) | Edit <input type="radio"/> |
| BnF_publications_officielles_annuelle_moyen_hote (note, [BnF_publications_officielles_annuelle_moyen_hote]) | Edit <input type="radio"/> |
| BnF_urgence_unique_illimite (page+2, [BnF_urgence_unique_illimite]) | Edit <input type="radio"/> |
| defaultconfig (default, [defaultseeds]) | Edit <input checked="" type="radio"/> |

| Seed lists | Edit |
|---|------|
| BnF_collecte_courante_annuelle_moyen_hote (http://expositions.bnf.fr) | Edit |
| BnF_collecte_courante_annuelle_petit_chemin (http://www.bnf.fr/lettre_gallica/) | Edit |
| BnF_collecte_courante_annuelle_petit_hote (http://enfants.bnf.fr) | Edit |
| BnF_collecte_courante_annuelle_petit_page+1 (http://www.bnf.fr/fr/professionnels/actualites_de_la_conservation/i.index_actus_conservation.html) | Edit |
| BnF_collecte_courante_mensuelle_petit_page+2 (http://guides.bnf.fr/francophonie; http://www.bnf.fr/fr/professionnels/actualites_de_la_conservat...) | Edit |
| BnF_collecte_courante_semestrielle_petit_hote (http://blog.bnf.fr; http://lajoieparieslivres.bnf.fr) | Edit |
| BnF_publications_officielles_annuelle_moyen_hote (http://www.bnf.fr; http://www.bnf.fr/la_bnf/a.chroniques.html) | Edit |
| BnF_urgence_unique_illimite (http://www.bnf.fr/fr/outils/a.facebook.html) | Edit |
| defaultseeds (http://blog.bnf.fr; http://bnf.fr; http://enfants.bnf.fr; http://expositions.bnf.fr; http://lajoie...) | Edit |

2.3.1. Role and users of the application

At the BnF, NetarchiveSuite can be considered as the central piece of the web archiving workflow in the Library. It takes the seed URLs for both broad and selective crawls, organises them by domain and by crawl settings, thus creating “packets” in the form of individual jobs which are passed to the crawlers. It is a production tool which is used both by digital curators (to organise the harvest from a content point of view) as well as by crawl engineers (to organise the harvest from a technical point of view).

The profiles of NetarchiveSuite users are therefore in correlation with that of the digital curators, who have a general understanding of web languages, analysis of page structure and URL syntax and knowledge of harvesting parameters, but also with that of the two IT crawl engineers dedicated to web archiving at the BnF, who have technical knowledge of the web archiving infrastructure underlying NetarchiveSuite. The interactions of the users with the different functions of NetarchiveSuite described below reflect these different profiles.

2.3.2. Three main modules: harvesting, archive, access

NetarchiveSuite makes it possible to manage, store and retrieve large-scale harvesting, due to its distributed Java-based system. It is built around Heritrix 1.14 which performs the individual web crawls.

The suite is split into four modules, all of them using the programming languages JDK 1.5.0 and Java JMS on Linux 2.6 operating system. The first module concerns general administration: it includes several applications which manage all the system components using a unique configuration file “settings.xml” which gives a central access to manage

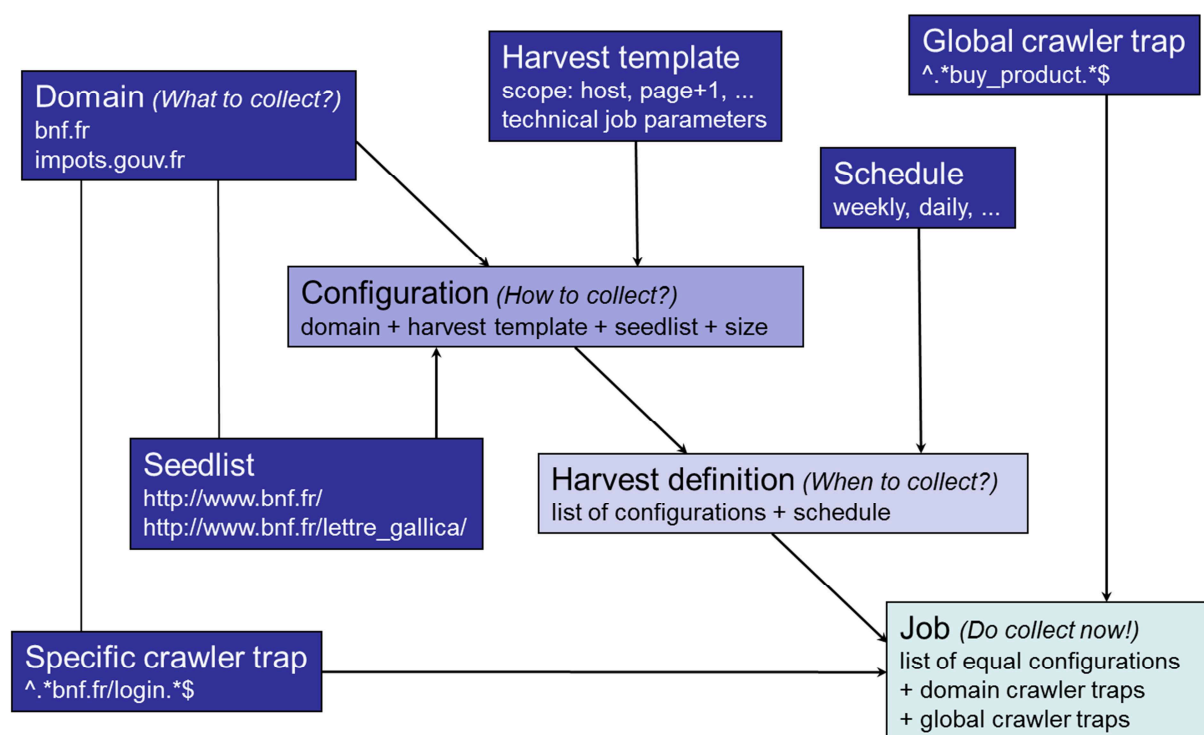
NetarchiveSuite. The remaining three modules deal with the major operations of web archiving: harvesting, preservation and access.

The harvester module is a web-based interface for the application “HarvestController” which is used to define and configure the crawls, to distribute jobs to Heritrix instances on the crawl servers and to package metadata about the harvest together with the harvested data. As described above, the NetarchiveSuite database can be automatically updated using the selection tool “BCWeb” for selective crawls and “NAS_preload” for broad crawls. Thanks to this module, the process of managing the crawl itself can also be automated by handling the different elements of the NetarchiveSuite data model, which is outlined in the next section.

The archive module provides bit preservation functions to check the integrity of crawled data with the help of two applications, “Bitarchive” and “ArcRepository”. This module is not active at the BnF because the web archives are preserved in the separate digital repository SPAR, described below.

The access module has two applications. The “IndexServer” creates indexes for the harvested data. It allows fast access to the harvested materials within NetarchiveSuite and makes it possible to generate a specific index for each harvest before starting the jobs. This last index is used to de-duplicate URLs that have been already crawled, by allowing Heritrix to indicate the location of the file that has already been collected rather than collecting it again. The “ViewerProxy” gives access to harvested material using the index generated by the IndexServer, through a proxy solution. It can be useful to look at the domains of a single job and check the quality of the crawls. It allows the curators to identify unharvested URLs while browsing and to include them in a subsequent harvest.

2.3.3. Data model: domain, configuration, schedule, harvest definition



In practice, the digital curators at the BnF use only the harvester module, to define configuration and crawl settings and to monitor crawls. The data model is based on the

domain unit, which in turn can have one or multiple **seedlists**, lists of URLs from which the crawl starts, to collect an entire website or a part of it. Once it has been created, a domain cannot be manually deleted from NetarchiveSuite database. At the end of 2012, this database contained 3.8 million unique domains.

For each domain, one or several **configurations** can be created. The configuration describes the ways in which the domain will be harvested: it combines a seedlist with a **harvest template**, the order.xml file that specifies the depth of the crawl and other technical parameters. For selective crawls, the configuration is defined by the migration of data from BCWeb, including the budget and the depth chosen by the curators. For broad crawls, a default configuration is used, using parameters defined in advance by the digital legal deposit team.

Finally, the harvest definition gathers different domains with their dedicated configurations and the same **schedule** (the frequency chosen by the curators). This system of planning harvests allows the creation of an annual overview of web archiving production and thus makes it possible to manage the monitoring and quality assurance activities of the team. The **harvest definition** is the fundamental element that gives structure to the crawling activity, and the manner in which these definitions are set up can vary depending on the needs of the institution, within the limits of the data model. At the BnF, the organisation of the harvest definitions reflects both collection policy and the needs of harvesting: for ongoing selective crawls, the URLs from different department collections are divided among generic harvest definitions defined not only by schedule but also budget, while project collections each have their own specific harvest definition.

When the time arrives to start a harvest, as defined in the schedule, one or more **jobs** are created by taking sub lists of identical configurations from the harvest definition. Global and specific **crawler traps** are added to prevent certain URLs being collected. The jobs are then sent to the harvesters.

It is however important that the digital curator team takes care to define the structure and standardise the names of the different elements before starting, because they cannot be deleted from the NetarchiveSuite database. Before beginning these technical operations, it is necessary to have defined a collection policy on web archiving in the institution, which will be reflected in the organisation of data in NetarchiveSuite.

This data model is sufficiently flexible to manage broad crawls as well as selective crawls and to run them at the same time with different settings. Also important from this point of view is the possibility that NetarchiveSuite provides to define “high” or “low” priority for crawlers; in this way selective crawls will not be blocked by the large number of jobs generated by a broad crawl.


3. Harvesting

3.1. *Heritrix*

Heritrix is a web crawler, developed by Internet Archive in cooperation with the different Scandinavian National Libraries, which was specially designed for web archiving⁴. Its name is an archaic English word for heiress (woman who inherits); since the crawler seeks to collect and preserve the digital artefacts of human culture for the benefit of future generations, this name seemed apt to the developers. It is open-source and written in Java. The main interface

⁴ For more information, see <https://webarchive.jira.com/wiki/display/Heritrix/Heritrix>

is accessible using a web browser, and there is a command-line tool that can optionally be used to initiate and control crawls.

| | | | |
|---|---|--|---|
| HERITRIX | | Status as of Nov. 22, 2012 10:45:10 GMT | Alerts: 7 (7 new) |
| Admin Console | | CRAWLING JOBS | RUNNING job: 5304-28 |
| | | 0 jobs pending, 0 completed | 160190 URIs in 1h39m13s (14.25/sec) |
| Console | Jobs | Profiles | Logs Reports Setup Help |
| Crawler Status: CRAWLING JOBS Hold | | | |
| Jobs | | Memory | |
| Running: 5304-28 | | 856440 KB used | |
| 0 pending, 0 completed | | 1351104 KB current heap | |
| Alerts: 7 (7 new) | | 2796224 KB max heap | |
| Job Status: RUNNING Pause Checkpoint Terminate | | | |
| Rates | | Load | |
| 14.25 URIs/sec (26.93 avg) | | 56 active of 200 threads | |
| 274 KB/sec (634 avg) | | 1 congestion ratio | |
| Time | | 5177 deepest queue | |
| 1h39m13s elapsed | | 391 average depth | |
| 42m11s remaining (estimated) | | | |
| Totals | | | |
| downloaded 160190 |  | 68054 queued | |
| 228300 total downloaded and queued | | | |
| 3.6 GB crawled (3.6 GB novel) | | | |

Heritrix has been designed with a great flexibility, modularity, and extensibility. A job can be composed of a wide variety of modules. There many different filters, extractors, and processors available for different tasks in the treatment of the URLs. Further extensions can be included, either very simple by adding BeanShell scripts or more in-depth by writing processors in Java.

Heritrix can run as a standalone application or as an embedded module, which is the case within the NetarchiveSuite installation. The communication between Heritrix and NetarchiveSuite is performed via JMX commands.

A crawl is configured as a job in Heritrix, which consists mainly of:

- a list of URLs to start from (the seeds),
- a scope (collect all URLs in the domain of a seed, stay on the same host, only a particular web page, etc.),
- a set of filters to exclude unwanted URLs from the crawl,
- a list of extractors (to extract URLs from HTML, CSS, JavaScript),
- many other technical parameters, for instance to define the “politeness” of a crawl or whether or not obey a website’s robots.txt file.

The politeness is established as a delay between two requests to the same website to prevent the site from being attacked by too many requests in a short time. There is a static amount in the delay (always wait a minimum and never wait longer than a maximum of seconds) as well as a dynamic one which is computed by taking the duration of fetch of the previous URL multiplied with a delay-factor.

The question of whether or not to obey robots.txt has both technical and political aspects. Many websites prevent robots from crawling directories that contain images, CSS files, or administrative resources. But this content is often essential to reconstruct web pages from the archive. The non-respect of the rules given in the robots.txt file may, however, provoke complaints by webmasters who do not understand the need of collecting this kind of content

and who therefore consider Heritrix an unethical crawler. French legal deposit law allows the BnF to ignore robots.txt to fulfil its legal mission of collecting the French Internet, and selective crawls have always been configured to ignore robots.txt while broad crawls previously respected this protocol. Since 2012, broad crawls too ignore robots.txt to improve the quality of the capture.

All these parameters can be stored in a profile, with the different profiles being managed by NetarchiveSuite in its database. The profile for a given job is thus created according to the configuration defined in the data model in NetarchiveSuite (see above). NetarchiveSuite also creates, configures, launches, and monitors the jobs in Heritrix.

The seeds and all the extracted URLs are organised in queues, defined by domain, by host, or by other parameters. A certain number of queues can be treated in parallel – at the BnF, this number is currently 250.

Even though the crawl jobs are completely controlled by NetarchiveSuite, there is still the possibility to make use of Heritrix’s web interface to perform certain operation on a running job such as pausing and terminating the job, change technical parameters of the job, or add filters to block unwanted URLs from being collected. Those operations can alternatively be performed via JMX commands which can be called in a shell script or directly on the command line of the operating system.

3.2. Monitoring with NetarchiveSuite

When a crawl is started by NetarchiveSuite, one or multiple jobs are automatically created following a load balancing algorithm which has the task to optimise the repartition of domains into as less jobs as possible. The algorithm takes in consideration the number of domains, the size of the harvested domains from last crawls, the similarity of configurations.

The jobs are then submitted via a JMS broker to a queue to be picked up by one of the harvesters. A Heritrix instance is started and a communication through a Java JMX technology is established between Heritrix and NetarchiveSuite. The digital curators can follow the running jobs and consult a panel with the progression rate, the number of active or exhausted queues, the harvest speed performances. If necessary, they can then look at the biggest queues and decide to add specific filters to correct the crawl. These filters are written as regular expressions directly in the Heritrix console and are added to the NAS database for future harvests to improve their quality.

When a job is completed and is stopped by Heritrix, the HarvestController application of NetarchiveSuite generates ARC files containing metadata which include logs and reports produced by Heritrix. These elements are sent back to NetarchiveSuite in the “ArcRepository” to update the database.

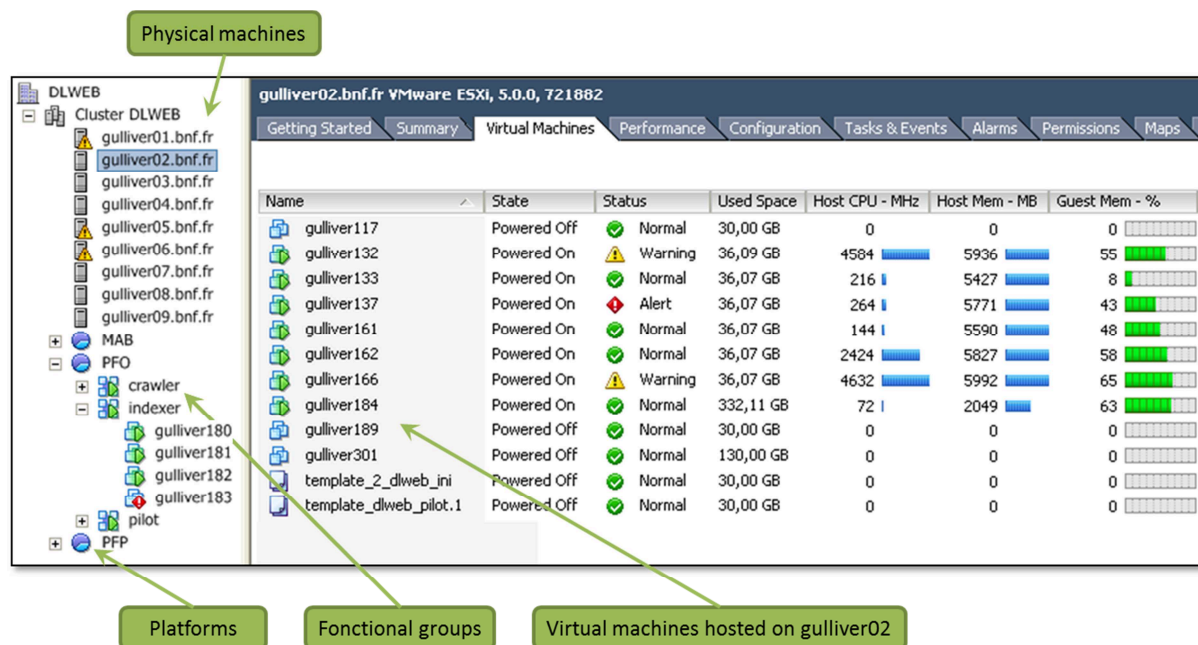
3.3. Virtual servers for a flexible infrastructure

The BnF’s crawling architecture consists of several types of machines: Pilot (NetarchiveSuite), de-duplication indexer (NetarchiveSuite), crawlers (Heritrix), and indexers (Indexing Process).

Three types of platform are defined for different purposes:

- The PFO (operational platform) is used for daily production.
- The MAB (trial run platform, or “marche à blanc”) has an identical setup to the operational platform and is used for test harvests under real conditions by the digital curators team, for example to test new job configurations or to examine new sites.

- The PFP (pre-production platform) is a technical test platform to test architectural designs or deployments by the technical staff.



To maintain this park of a variable and scalable number of machines, a virtualised environment based on vSphere by VMware has been installed. The physical base is an IBM BladeCenter in which 9 blades are currently used for web archiving, each one of them equipped with 2 quad-core Intel Xeon CPUs and 72 GB of RAM. A virtual machine is hosted on a single physical server at a given time. If the load on the servers becomes too heavy, some of the virtual machines are moved onto another host dynamically and without interruption. If one of the hosts fails, all the virtual machines hosted on this server are moved to other hosts and are rebooted.

Virtualisation allows another advantage: an active copy of the virtual machine can be run on another server. If the server where the master virtual machine is hosted fails, the ghost virtual machine instantly takes control without interruption. A new copy is then created on a third server. This function can however use up a lot of resources, and especially network consumption. For this reason at the BnF this kind of protection is used only for the pilot machine.

This gives a flexibility regarding the number of machines allocated to a platform, sharing and optimisation of the use of hardware resources as well as improved robustness and reliability. There is one single instance of Heritrix active on every virtual machine, which can act as either high or low priority harvester to serve selective or broad crawls, respectively. At the BnF, the current configuration allows a number of up to 70 crawling machines on the Operational Platform. The number of crawlers active at any given time will be decided depending on the number of crawls defined in NetarchiveSuite: more crawlers can be allocated to the Operational Platform during busy periods (for instance, during a broad crawl), while during less busy periods more resources can be allocated to other platforms for tests.

3.4. Quality assurance

3.4.1. International quality indicators

For national libraries, web archives now have to be included in the library general performance statistics, and considered as heritage and research library materials. To that end, an ISO

working group has examined how to define the different indicators that may be considered as standards for web archiving. This will aid international measures and comparisons and, at the same time, allow a better evaluation of practices within institutions⁵.

Since the beginning of its web archiving activity, the BnF has produced a large number of indicators, and the digital legal deposit team has been involved in ISO discussions and reports as a main contributor. Since 2010, the ISO recommendations have been integrated into the web harvesting workflow.

3.4.2. Production of statistics with NAS_qual

NetarchiveSuite keeps records of all broad crawls and selective crawls operated by the BnF. This means that, for each job, it is possible to store and analyse Heritrix reports and NetarchiveSuite operations. The tool developed by BnF to perform these analyses is a command-line Java application, called “NAS_qual”. Launched automatically each day, it exports this information in .txt files that digital curators can access via a web browser: general metrics, MIME types, HTTP response codes, TLD repartition, top domains, etc.

Using the statistics produced by NAS_qual comparisons can be made for the total annual production between the current year and the year before, between selective and broad crawls, and between two crawls of the same harvest definition. Finally, these statistics are useful for BnF managers as well as subject librarians to illustrate web archiving activity.

3.4.3. Some metrics used at the BnF

The four following examples are useful in characterizing a web archive collection, as well as providing a starting point for an interpretation of the web media itself.

The top level domains help to characterise a collection in terms of geographic distribution (e.g. France). For example, in 2011, starting with a seedlist of 2.8 million domains in .fr, the results of all the harvests gave: 62% of collected URLs in .fr, 27% in .com, 2% in .net, and 9% for other TLDs. This shows that the French scope also includes a large part of .com domains.

The digital curators also measure the size of domains in order to verify if the target of a crawl is reached and to improve the knowledge on active web sites. In 2011, almost 57% of collected domains were smaller than 10 URLs: this means that half of the domains were almost empty; many of them were domain names reserved but without any content on the web, redirections, and so on. A further 42% of collected domains had between 10 and 10,000 URLs, which is the limit for BnF French broad crawl. Only 1% of collected domains contained more than 10,000 URLs and were thus not collected in their entirety.

The analysis of MIME format types is intended to give a distribution by type of content comparable with other documents in the Library and also to help preservation tasks. In 2011, the harvests were composed of 59.7% text and 36.2% image formats; this will be similar to the proportion in the “live” web before archiving. The other categories are smaller but, at the same time, will need more special attention as regards their preservation: 3.9% application, 0.1% audio and 0.1% video formats.

The last example concerns the source of the seed URLs, to highlight the portion of the archive that is the result of human selection. In 2011, 7% of seeds in NetarchiveSuite were chosen by

⁵ This working group is part of the Technical Committee 46, Information and Documentation, Sub-Committee 8, Quality – Statistics and Performance Evaluation. Its report will be released in 2013 as the ISO/TR 14873, Statistics and Quality Indicators for Web Archiving. For more information, see http://www.netpreserve.org/sites/default/files/resources/IIPCGA_ISO_Workshop.pdf.

subject librarians and 93% came from the AFNIC list, registered in NAS_preload. These two numbers have to be compared with the proportion of URLs collected during the year: 43% for focused crawls (selection of librarians) and 57% for broad crawls (automated harvesting). The value of librarians is seen in these numbers, as although the number of seed URLs for selective crawling is much smaller than that for broad crawls, the numbers of collected URLs for each type of crawl are almost the same. This is due to the use of specific frequencies, depths and other parameters to achieve a better harvesting quality for the selective crawls. In this way, the BnF maintains a balance between its mission of legal deposit (to preserve a snapshot of all what can be seen on the web media) and its mission of heritage protection (to highlight specific practices and behaviours on the web).

3.4.4. Quality assurance using NetarchiveSuite

To manage quality, the BnF team uses crawl data stored by NetarchiveSuite to refer to previous crawls. Associated with each domain is historical data, including the number of harvested objects, whether the harvest was completed or reached the limit imposed, what configuration was used...

If necessary, the observations made by the digital curators on the crawls can also be indicated in the selection tool BCWeb to give feedback to content curators. And the circle of harvesting is closed: when the quality of the previous crawl was not good enough, the parameters can be changed in BCWeb and sent back to NetarchiveSuite for another crawl.

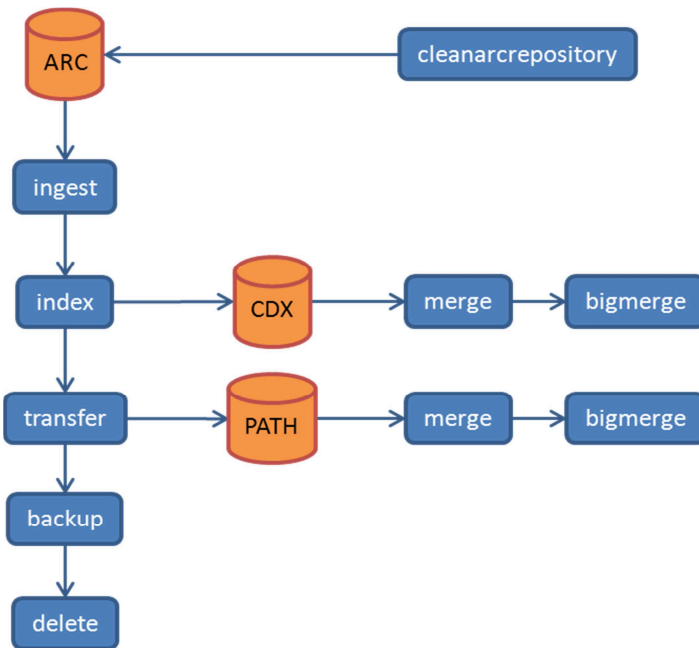
4. Post-harvest processing

4.1. *The indexing process*

To make the collected data accessible in the Wayback Machine, the ARC files produced by Heritrix have to be indexed. Two types of index are created: CDX files that show the location of a given URL within an ARC, and PATH files that show the physical location of the ARC. By means of the information stored in CDX and PATH files, the Wayback Machine is able to retrieve a given URL from the archive.

The indexing process is performed by shell scripts that execute several tasks within the indexing process. The scripts run on dedicated machines and are started automatically after a job is finished and the ARC files have been moved into the *ArcRepository* of NetarchiveSuite:

The scripts form a chain of independent modules, which can be interrupted and restarted at any time. Every module has its own working directory. When started, the module looks into its working directory and treats every file that it finds. After successful treatment, the file is moved into the working directory of the next module in the chain.



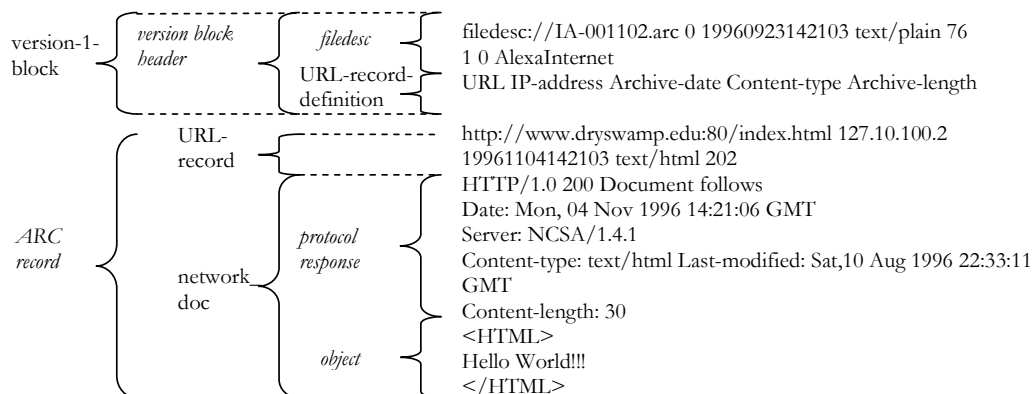
ingest

The first module acts as a kind of master in the indexing process. It maintains a manifest of indexed ARC files and a list of indexing machines. The module searches for ARC files in the ArcRepository of NetarchiveSuite which are not listed in the manifest, and therefore have not yet been indexed. For each new file, a checksum is computed and every indexing machine is queried if it is willing to accept another ARC file. A copy of the ARC file is then placed into the index working directory of the machine with the fewest number of files in the queue.

The harvest ID from NetarchiveSuite, which shows which harvest definition the crawl belongs to, is part of the name of the ARC file. This ID can be used to define a sub-chain in the indexing process to create a specific index file for one or more harvest definitions. As these harvest definitions are linked to the collection policy, the Wayback Machine can use these different index files to define entry points to the collections.

index

This module creates the CDX index entry for each URL present in the ARC file. The **ARC file** format is defined of a version block, describing the ARC file itself, followed by a sequence of records, each one consisting of URL metadata, HTTP response, and the requested document as it has been fetched:



The module takes every record from the ARC files and creates a line in a **CDX file** which contains the URL in a normalised form, the timestamp of the fetch, domain, HTTP response code, checksum, offset within the ARC file, and the name of the ARC file:

```
dryswamp.edu:80/index.html 19961104142103 dryswamp.edu text/html 200  
7c08de3980044c0edcb5579bef71be9 99552170 IA-001102
```

This will tell the Wayback Machine in what ARC file and at what offset it can find a given URL collected at a given time.

transfer

The ARC file is copied into its final location in the file repository. Several possible locations can be given in a configuration file, and the module checks if sufficient space is left in the location. A **PATH file** is generated which has a line for every ARC file, indicating the filename and its physical location in the repository:

```
IA-001102 /net/storagenode001/IA-001102/IA-001102.arc.gz
```

This will tell the Wayback Machine at what location it can find a given ARC file.

backup

If desired, an immediate backup of the ARC file can be made in a centralised backup service.

delete

The ARC file is deleted in a controlled manner from the indexing machine. A deletion indicates also the successful termination of the indexing process for that ARC file.

merge

This module is executed once a day and merges together all the CDX and PATH files produced during the day into new files. To be used in the Wayback Machine, the new files must be sorted alphabetically.

bigmerge

This module is executed on the master machine and merges the CDX and PATH files of the day from the indexing machines. These files are then merged with the “big” CDX and PATH files in the repository.

cleanarcrepository

As a final step, this module deletes all the ARC files in the ArcRepository of NetarchiveSuite which have been indexed so far. This is to make sure that only files which have passed the indexing process are deleted.

Except for the first module (ingest) and the last two (bigmerge and cleanarcrepository), all the other modules can be run in multiple instances in parallel on several machines. The chain is thus scalable and its capacity can be adapted according to the number of ARC files that have to be indexed in a certain time. The maximum capacity of the chain is determined by the capacity of the ingest module which is at the BnF currently capable of treating 20,000 ARC files per day.

4.2. Access to the collections: “Archives de l’Internet”

The consultation of Internet legal deposit collections at the BnF is provided via an interface based on the Open Source Wayback Machine, which has been customised for the BnF’s needs⁶. The application is accessible through a Terminal Services (TSE) environment which provides two advantages: it prevents interference from live web content, thus maintaining the integrity of the archives for users, and prevents users from downloading material from the collections, which is forbidden by the law. The TSE environment runs an embedded version of Firefox browser with associated plugins to allow the playback of different file formats that may be found in the archives (PDF, video and audio files, etc.) The access application, known simply as “Archives de l’Internet”, runs automatically within Firefox as soon as the session is opened.



The main entry point into the archives is search by URL. Each search provides a chronological list of results from 1996 to the present, with each capture shown individually. webpages are then reconstructed as close as possible to their original appearance, with navigation by hyperlink possible within and between sites, as long as the linked content is present within the archives. An orange banner at the top of the screen shows the date and time of the capture.

There is currently no full text indexing process at the BnF. A small part of the collections has been indexed with NutchWAX by Internet Archive, corresponding to the broad crawls performed by them for the BnF in 2006 and 2007. A full-text search for this part of the

⁶ For more information, see “Introducing web Archives as a New Library Service: the Experience of the National Library of France”, AUBRY S. In: *Liber Quarterly*, 2010, vol. 20, n° 2. [\[http://liber.library.uu.nl/index.php/lq/article/view/7987\]](http://liber.library.uu.nl/index.php/lq/article/view/7987)

collections is therefore proposed, but its utility remains limited and the lack of a fully-functional full-text search for users is a major shortcoming in the current access interface.

The third and final method of accessing the archives is via selections of sites, known as “Guided tours”, prepared by BnF subject librarians, sometimes with external partners. These are intended to provide a user-friendly way of discovering the collections and also to showcase the work done by selectors, and especially projects. There are currently six “Guided tours” (Elections 2002 and 2007, Literary and personal blogs, web activism, Sustainable development, Arab spring in Tunisia, Amateur images and videos).

As well as providing access to users, the application also allows digital curators and librarians to perform visual quality control of the archives. Comments can then be noted in BCWeb with the aim of improving future crawls, either by including additional URLs or by trying to improve crawling tools to capture difficult content such as rich media.

Under French legal deposit law consultation of the BnF web archives is only possible for accredited users within the confines of the library, for reasons of copyright and data protection. The access interface was first put in place in 2008, and is now provided on all computers in the research library at the main François-Mitterrand site of the BnF and at the other sites. Levels of use of the archives remain limited but stable: 30-50 users per month, but with an increasing number of sessions lasting 30 minutes to 1 hour or more. Librarians in the reading rooms provide support and mediation for users.

4.3. SPAR: the BnF digital repository

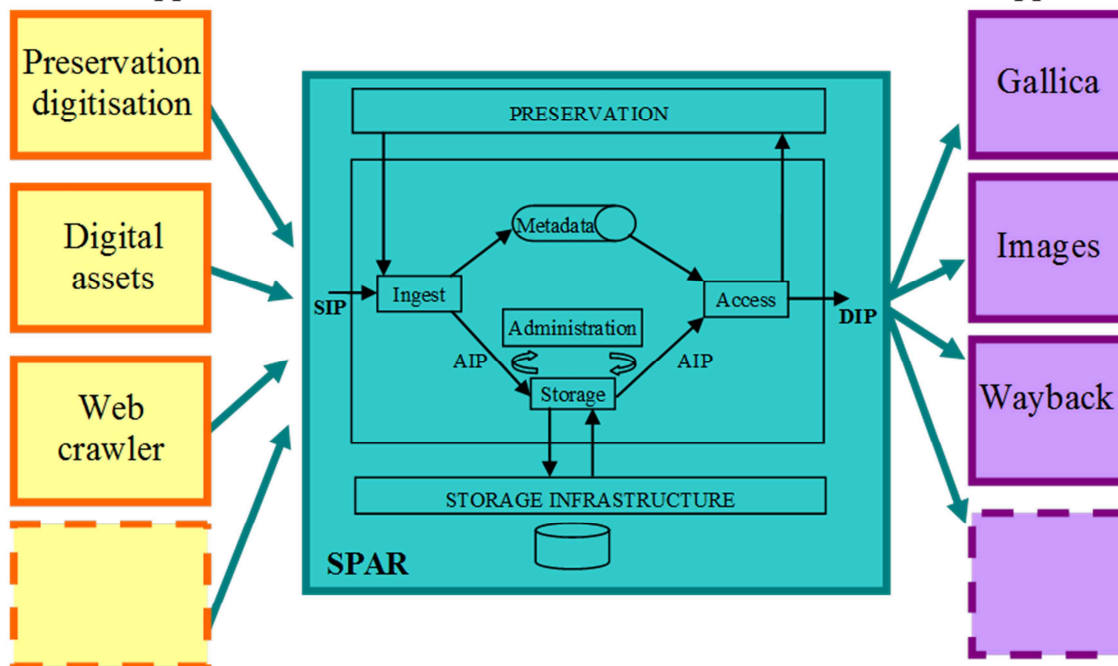
After indexing, the ARC data and metadata files are ingested into the SPAR system (Système de Préservation et d’Archive Réparti – Scalable Preservation and Archiving Repository). SPAR is a long-term preservation system for digital objects, compliant with the OAIS (Open Archival Information System) standard, ISO 14721. The system guarantees the independence of storage units from the physical infrastructure, which is itself strongly secured (distributed between two geographically distinct sites, each site owning its own tape library). It has been put in place at the BnF to face the challenges of the volume of digital collections as well as the diversity of their formats⁷.

⁷ “From the World Wide Web to Digital Library Stacks: Preserving the French Web Archives”, OURY C., PEYRARD S. In: *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES)*, Singapore, 2011, pp. 231-241. [<http://getfile3.posterous.com/getfile/files.posterous.com/temp-2012-01-02/dHqzmzicCGoexvmiBzJDCyhrhIgswoffzvsfnpEAXjHFEesarvwahEHrmyvj/iPRES2011.proceedings.pdf>]

“Preservation Is Knowledge: A community-driven preservation approach”, DERROT S., FAUDUET L., OURY C., PEYRARD S. In: *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES)*, Toronto (Canada), October 2012, pp. 1-8. [<https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf>]

Production applications

Diffusion applications



Different “tracks” are defined for different groups of digital objects. The tracks are characterised by the relations between the production of the digital objects (needs and requirements) and the archival system (service agreement). Preparation of the DL_AUTO track, for the Internet legal deposit collections from automated crawls, began in 2007. It concerns 370 TB of data from web harvesting with Heritrix (with NetarchiveSuite or as a standalone crawler) but also with other crawlers as HTTrack used by the BnF during its first period of web archiving experimentation. Each type of material is treated as a separate “channel”.

| Channels | Periods | Robots |
|---|----------------|-------------|
| Historical collections | 1996-2005 | ia_archiver |
| First election crawls | 2002 and 2004 | HTTrack |
| Internet Archive crawls | 2004-2008 | Heritrix |
| In-house crawls without NetarchiveSuite | 2006-2010 | Heritrix |
| In-house crawls with NetarchiveSuite | 2010 and after | Heritrix |

The ingest into SPAR is closely linked to the functioning of NetarchiveSuite: in addition to the crawled data produced by Heritrix, SPAR will also preserve the metadata ARC files produced by NetarchiveSuite, containing the configurations, reports and logs that describe the crawls. This allows SPAR to create coherent collections of data using a data model which is organised in three layers of granularity, based on the organisation of NetarchiveSuite: the ARC, the crawl job (containing ARCs of both data and metadata) and the harvest definition (containing the jobs). While the data model of SPAR is based on that of NetarchiveSuite, it will also be applied to previous kinds of crawls (such as standalone Heritrix crawls performed by the BnF, broad crawls by Internet Archive and historical collections extracted by Alexa Internet). The ingest has begun with the most recent crawls performed with NetarchiveSuite; the retrospective collections will enter SPAR after this channel is completed.

In practical terms, when a job is finished and indexed, it is ready to be stored in a temporary zone before being ingested into SPAR. A script is applied to push the metadata file first to the pre-ingest process of SPAR to make it possible to rebuild the three levels of data. When SPAR has accepted this file, another script takes all the data files, one by one. An ARK (Archival

Resource Key) number is given to each file (data or metadata) by the pre-ingest process of SPAR. If there is a problem at any stage, two types of solution can be applied: an automatic process if the error is not serious, or a manual one if a technical analysis is needed.

At this point, statistics are supplied by the scripts working on the temporary zone (number of jobs put into the temporary zone, their harvest definition names, etc.) and others by SPAR (ARK numbers of the information packets, ingest time, etc.).

The data ingested through the DL_AUTO track is not the only one to enter SPAR; there are other tracks as digitalisation or third party storage which could also generate a great amount of assets. As the volume of data is restricted, this means that the speed of ingesting could be lowered.

5. Conclusion

The unified web archiving workflow that has been put in place at the BnF has many advantages. It is flexible and scalable enough to handle the different technical needs of the different kinds of harvest performed by the BnF in its broad crawls and its different selective crawls. The workflow can thus respond to the different collection needs imposed on the BnF by its mission of Internet legal deposit. The tools put in place also reflect the organisation of web archiving at the BnF, as they allow a clear division of the different roles: the digital legal deposit team, the crawl engineers and the subject librarians, while allowing visibility and reporting for management.

This flexibility comes in large part from basic technical decisions: the choice of open source software solutions where possible means tools do not need to be bought or developed from scratch, while allowing the possibility to customise them for specific purposes and enrich them for the benefits of the user communities. From an infrastructure point of view, the use of virtual servers again allows a flexible response to the varying technical needs inherent in the different kinds of harvest put in place by the BnF.

Finally, the creation of a unified workflow is a demonstration of the integration of web archiving activity as part of the everyday activity of the BnF. Just as Internet legal deposit is now established as one of the missions of the Library, the web archiving workflow, in its technical and organisational aspects, has taken its place besides other workflows such as the treatment of books and periodicals.

There are of course still areas to improve and shortcomings in the current system that need to be addressed. In particular, the access interface remains limited and work needs to be done on optimising the indexing process, implementing a full-text indexing process and proposing new methods of consultation and analysis of the collections for users. Other projects underway or planned for the near future are the harvesting of subscription resources and e-books, which will have an impact on all aspects of the workflow. The system that has been put in place has proved its effectiveness, however we will continue to develop new processes and tools that will allow the existing workflow to evolve to meet the needs of the Library and its users.