

Legal deposit of the French Web: harvesting strategies for a national domain

France Lasfargues, Clément Oury, Bert Wendland

▶ **To cite this version:**

France Lasfargues, Clément Oury, Bert Wendland. Legal deposit of the French Web: harvesting strategies for a national domain. International Web Archiving Workshop, Sep 2008, Aarhus, Denmark. hal-01098538

HAL Id: hal-01098538
https://bnf.hal.science/hal-01098538

Submitted on 26 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Legal deposit of the French Web: harvesting strategies for a national domain

France Lasfargues, Clément Oury, and Bert Wendland
Bibliothèque nationale de France
Quai François Mauriac
75706 Paris Cedex 13
{france.lasfargues, clement.oury, bert.wendland}@bnf.fr

ABSTRACT

According to French Copyright Law voted on August 1st, 2006, the Bibliothèque nationale de France ("BnF", or "the Library") is in charge of collecting and preserving the French Internet. The Library has established a "mixed model" of Web archiving, which combines broad crawls of the .fr domain, focused crawls and e-deposits.

Thanks to its research partnership with the Internet Archive, BnF has performed four annual broad crawls since 2004. The last one has been made with noticeably different features: one of the most important was the use of the all-comprehensive list of the .fr domain names, given to BnF by the AFNIC ("Association française pour le nommage Internet en cooperation", the registry for the .fr) after an agreement was signed between both institutions in September 2007.

The technical choices made before and during a crawl have a decisive impact on the future shape of the collection. These decisions must therefore be taken according to the legal and intellectual frame within which the crawl is performed: for BnF, it is the five-centuries-old tradition of the legal deposit. To assess the consequences and the outcomes of the different technical solutions available, we propose to analyze the results of the BnF's last crawl and to compare them to those of previous harvests. These studies also prove to be useful in our attempt to characterize the 2007 French Web.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Information search and retrieval – *Information filtering, Query formulation, Relevance feedback, Retrieval models, Search process, Selection process.*

General Terms

Measurement, Documentation, Experimentation.

Keywords

Web archiving- Internet legal deposit – French national Library – BnF – International Internet Preservation Consortium – IIPC – Internet Archive – Digital Heritage.

1. THE FRENCH CONTEXT

1.1 Defining the scope of the legal deposit

On August 1st, 2006, a new Copyright law was voted by the French Parliament. One of the titles of this law, long-expected by BnF, extended the legal deposit to the Internet. The legal deposit is the obligation for every publisher to send copies of his production to the Library. First established in 1537 for printed materials, the legal deposit has been extended through centuries to every new published form of intellectual creation, from engravings to software and video games. As the World Wide Web was becoming the favorite place to create and spread out knowledge and information, it was necessary to give French heritage institutions a legal frame to organize its preservation.

The law is not explicit on what the French domain of the Internet is, but a decree is expected to clarify this issue in a near future. In the meantime, the Library has forged its own doctrine of the online national domain, which is likely to be consistent with the decree whenever it will be enforced. In order to refine the scope of its Web archiving policy, BnF combined its five-centuries-old practice of legal deposit together with its more recent experience of Web harvesting technical challenges. Our approach therefore needed to be both pragmatic (that is, compliant with bulk harvesting exploratory tools) and consistent (with the French legal deposit tradition).

This tradition is based on three criteria:

- *Publication*: documents to be collected are aimed at an audience and serve a public purpose, they should not fall in the fields of private or corporate internal communications;

-*Media:* all previous legal deposit dispositions were based on the physical existence of a media: prints, scores, photographs, tapes, disks, etc.;

- *Territory*: documents are to be published or distributed within the borders of the national territory.

In short, traditional legal deposit is applicable to any publication embedded in a media produced or distributed on French territory.

Unfortunately, none of these criteria is easily applicable to Web harvesting because websites:

- Tend to merge and mix public and private communications rather creatively;

- Are not a media as such, but rather a platform where all existing media tend to migrate (we can find books, photographs, films, scores, etc. on the Web);

- Cannot easily be localized on a specific territory, at least not on a very large scale.

Moreover, we cannot use the French language as selection criteria either, since our legal deposit applies regardless of the language of publication: BnF legal deposit collections include many items in foreign languages provided they are published, printed, or distributed in France. Likewise, we cannot expect our legal deposit strategy to focus on specific themes, authors or levels of publications. An important aspect of our legal deposit tradition is that collections should mirror the French society and culture in all its diversity regardless of the scientific value of the publications or their popularity. If we are to select, this is rather to sample than to choose. It is the next generations to decide what will be valuable one day, not the Library. In BnF stacks, unknown writers and dirty magazines stand side by side together with the Great thinkers and we expect the same philosophy to apply to our Web collections. Bulk harvesting represents a great opportunity to extend this approach at the scale of the Web. Last, our legal deposit tradition is both about content and form – or media -, which means that BnF pays attention to build collections which reflect trends in publishing models: we try to grasp a great diversity of objects and formats representative of the practical conditions met by the people who use information.

As a result, we needed to find a definition for our national domain which would reflect the "spirit" of this tradition, yet remain flexible and easily applicable. Defining a French "focus" while allowing for flexibility was indeed the only way to facilitate exploratory harvesting methods on a large scale. Language, geography, names or topics having proved to be either irrelevant or too challenging to be used for scoping discrimination on a large scale, what was left as a possible starting point for our exploration of the national domain was therefore the use of our national Top Level domain, the .fr, as a core starting list for our exploration, to be combined with other strategies.Our online legal deposit doctrine, as we expect it to be clarified in the decree, was therefore defined as follows. We consider to be "French":

- As a core, any website registered within the .fr TLD or any other similar TLD referring to the French administrative territory (for instance, the .re for the French island of La Réunion);

- Any website (possibly outside of .fr) whose producer is based on the French territory (this can usually be checked on the website or using DNS);

- Any website (possibly outside of .fr) which can be proved to display contents produced on French territory (this last criterium is more challenging to check but leaves room for interpretation and negotiation to the Library and the Web producers).

Of course, none of these criteria is expected to be strictly met before we build our seed lists (this would forbid exploration, and would involve checking every website before crawl, which is simply not scalable). However they are aimed at serving as legal and intellectual framework in order to:

- Define and explain the general policy of our national Web archiving policy to the public, stakeholders and librarians involved in the project;

- Clarify for ourselves the missions and directions we need to keep in mind when monitoring our crawls;

- Provide explicit elements of decision when we need to find out whether a website is definitely in scope or out of scope. This should prove especially useful whenever the Library is being asked by a webmaster to stop crawling a particular website or even face a law suit: if the website does not meet any of the above criteria, then it shall be excluded from future crawls and from the collections.

This approach of our national domain therefore reflects a compromise between our legal deposit tradition and the challenging characteristics of the Web. It is also a compromise between a totally open and absurd approach which could possibly lead us to consider the whole World Wide Web as potentially French and, on the contrary, a restrictive approach reduced to the sole .fr TLD while it is known to contain only a limited portion of the French websites. In short, it is aimed at providing both focus and flexibility.

1.2 Where we are for now

The Library is allowed to use various ways to collect the French Internet: "Mandated institutions may collect material from the Internet by using automatic techniques or by setting specific agreements and deposit procedures together with the producers" (Clause 41 II). In line with these dispositions, BnF defined a "mixed" model, which combines three strategies.

- Bulk harvesting of the French Internet. The goal is to collect at least the .fr domain on a yearly basis. These broad crawls allow the Library to archive snapshots of the French Web. It is the less expensive approach if we compare the costs of the harvest (machines and humans) and the retrieved amount of data. However, due to resources and tools limitations, it is not possible with such crawls to harvest the deep Web (very big websites, databases…).
- Focused crawls of a restricted number of websites. These sites are discovered by a collaborative network of librarians and researchers, inside or outside the Library. Focused crawls are dedicated to big websites and to frequently modified websites.
- E-deposits for a limited number of electronic publications.

The Library did not wait for the law to experiment crawling techniques. The Web archiving project began in 1999. A first event-based focused crawl was tested in 2002, when the Library harvested nearly 2 000 websites related to the national elections (presidential as well as parliamentary elections). This work was renewed two years later, for European and for local elections: BnF then collected 1 162 websites[1].

[1] The crawler used for the 2002 and 2004 elections was HTTrack [12]. See [19] for more information on these two crawls.

However, the technical means (hardware and software), the skills and the experience necessary to realize large-scale crawls of the French Internet were still lacking within the Library. This is the reason why BnF agreed on a partnership with the Internet Archive, a not for-profit foundation involved in world-wide Web archiving since 1996. On November 2004, both institutions signed a research agreement named « Research project: Selection of a National Domain for Web archiving ». Its goal was to assess several methods and tools to be applied to a national domain Web crawl. The agreement specified that the broad crawls necessary to test these tools and methods would be performed by Internet Archive, and that the data collected during these crawls would be delivered on storage racks to BnF.

The first broad crawl done in the frame of this agreement occurred at the end of 2004. Three other .fr broad crawls followed in 2005, 2006 and 2007 (this last crawl will last until 2009 thanks to an extension of the research agreement).

Smaller-scale crawls were also performed – directly or not – by BnF. Two focused crawls, on a limited number of websites (about 4 000), were completed by IA for the research project [15]. As of 2007, these websites have been harvested by BnF by its own means. Other thematic or event-based focused crawls were performed the same year, such as the websites related to the French national elections of 2007.

Up to now, BnF has thus performed – directly or thanks to its partnership with IA – four broad crawls besides a large number of focused crawls. The Web archiving is not a *project* within the Library anymore instead it has become a daily activity and a permanent unit within the Library's Legal Deposit Department. As it is the best approach to face the challenge of collecting a growing number of digital objects on the Web, bulk harvesting is still considered to be the top priority.

However, bulk harvesting does not mean blind harvesting. Even when they try to discover a maximum of files on the Web, robots conform themselves to a set of rules and settings [20]. The technical decisions made before, during and after the crawl have a decisive impact on the outcome of the harvest.

This paper describes the strategies that BnF has developed together with the Internet Archive in order to run large-scale crawls that could match this vision of the French national domain.

2. CRAWL DESIGN

2.1 What is the goal?

It seems unavoidable, when performing a broad crawl over hundreds of millions of files, to face a lot of technical issues. However, the first question one should answer before starting a crawl is not a technical one: what is the goal of this crawl? Answers will be different if they come from a corporate, a research or a heritage institution. A broad crawl will not be done the same way if the data is to be indexed (by a search engine), analyzed (for Web domains characterization, for example) or archived and displayed for the long-term.

On the other hand, it is necessary to take the technical limits into account when defining the goals of a broad crawl. That is the reason why a constant dialogue should be established between the librarians (in charge of defining the collection policy) and the engineers (in charge of performing the crawls). The methodology described in this section describes the meeting point of their respective concerns.

Broad crawls are done at BnF within the frame of legal deposit. This frame is not only a convenient one to address intellectual property protection issues. It also introduces Web archiving as the continuation of a long-lasting mission. Web archives collection policies should comply with those of previous publishing forms.

The automatic discovering of websites by robots is a way to match the "non-discriminatory" feature of the French legal deposit, to harvest the "best" (literature, scientific publishing) as well as the "worst" (from advertisings to pornography) of French publications. However, even robots are subject to bias. The hyperlink structure of the Web leads to discover and to harvest the most "cited" websitesfirst. Less popular websites should however not be forgotten by our archiving robot. To avoid such a bias, BnF signed on September 2007 an agreement with the AFNIC, a body responsible for the management of .fr and .re domains. According to the terms of the agreement, AFNIC shall give the complete list of domain names registered on the .fr and .re domains (currently more than one million domain names)twice a year. On the other hand, BnF shall guarantee the confidentiality of this very valuable data.

The goal of a large domain crawl is therefore to collect a representative sample of the national domain and to illustrate the French production at the time of the harvest. This sample is often designated as a snapshot – a way to record and to freeze a moving space. As it is not possible to harvest everything, we prefer harvesting few documents on every website rather than collecting entirely few websites, at the expense of the others.

This is why we did not experiment ways to crawl deeper into the "most important" websites, as the National Library of Australia did by putting a high priority on government and academic sites (a list of these sites had been established by the librarians [17]). It was also not necessary, as our broad crawls are complemented by focused crawls.

On the other hand, the Web is likely to be quickly affected by technological evolutions. New publication forms appear and spread within months. The broad crawl should reflect those evolutions. One of our goals, as a legal deposit institution, is to illustrate the new forms of publishing and therefore to make sure that the robot is able to harvest these documents. This is why we paid more attention to blogs and personal websites in 2006 while we focused on videos in 2007 (see below).

According to the pre-defined goals, it is possible to shape the future collection before, and during the crawl. Two important elements decided before the beginning of the harvest play a major part in the constitution of future collections: the design of the seed list, and the crawl settings.

2.2 The seed list

The crawler begins its tasks with a list of URLs called the seed list. The seeds are the doors giving access to the websites; that is why the quality of the harvest depends for a large part on the quality of the seed list.

As our partner for performing our broad crawl remained the same for four years, it was possible to enrich the seed list progressively. Different sources of seeds were added year after year.

- 2004: For the first broad crawl, the seed list came from an extraction of the .fr domains of the last Alexa crawl[2].
- 2005: Seeds came from an extraction of .fr domains of the last Alexa crawl and of the host discovered during the previous BnF broad crawl. The goal was to give to the crawler an opportunity to go further and discover new hosts.
- 2006: The seed list was generated in the same way as the previous year.
- 2007: Thanks to the agreement with the AFNIC, the all-comprehensive list of the .fr and .re domain names was used as seed list. To ensure continuity and consistency with previous broad crawls, this list was merged with the Alexa crawl extractions and the host extractions from the previous crawls.

The AFNIC list consisted of:

- 890 064 .fr domains;
- 1 516 .re domains;
- out of which 21 753 were second level domain names (see below, 2.3.3).

It was, however, not possible to use the AFNIC list as a simple seed list: it was a list of domain names, not of URLs. There was not necessarily a website behind each domain name.

At reception of the AFNIC list, several analyses were thus processed.

The first one was to quantify how many domains were still active and should actually be used as seeds. This large test has been done by Internet Archive. Each domain was tested with 2 different addresses (that is testing 1 780 128 URLs), to check if they answered online:

http://domainname.fr

http://www.domainname.fr

Both versions of a domain have DNS	79%
One version of a domain has DNS	14%
None of the two versions has DNS	7%

Figure 1: DNS responses of AFNIC domain names

A list of valid URLs was created from those results, so that "domainname.fr" became either "http://www.domainname.fr" or "http://domainname.fr" if the www version did not answer.

Seeds with no DNS were included and randomized in the seed list.

We checked how many domain names listed in the AFNIC registry were also available in our web archives (and especially in our 2006 broad crawl collection). It was quite a surprise for us to discover that only 30% of the AFNIC domains were available within our collection.

This is probably related to the significant increase of the .fr domain size. The .fr has been steadily increasing since 2004, thanks to the successive easing off of the .fr attribution rules.The major one was the opening of the TLD to individuals in June 2006 (only administrations and private societies were allowed to own a .fr domain name before this date):the .fr grew up to63% in one year. This must be related to the positive image of a ccTLD on Internet users. Individuals represent for now 30% of the registered .fr domains, and 50% of new registrations [2].

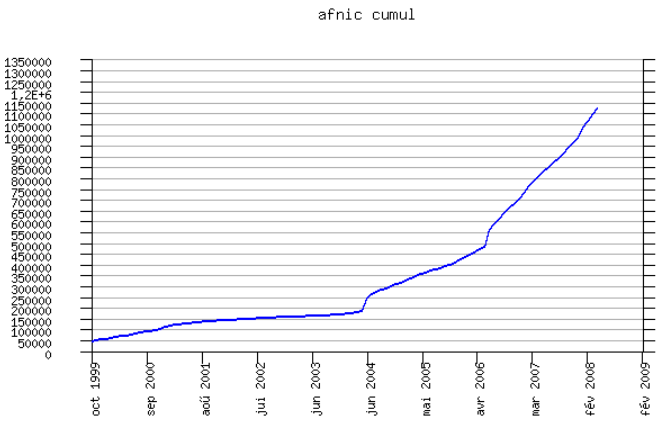


Figure 2: .fr domain size evolution[3]

This phenomenon might also be partly due to the presence, in the AFNIC list, of domain names that were not or were poorly-linked to other websites, and had not been discovered by robots the previous years.

This tremendous growth largely explains the considerable extension of our seed list from 2006 to 2007.

	2006	2007
Seeds extracted from Alexa crawls	207 046	58 224[4]
Seeds extracted from previous crawls	427 476	2 295 890
AFNIC list	-	890 064
Total after duplicate reduction	562 634	2 888 723

Figure 3: seed list size evolution from 2006 to 2007.

[2] Alexa Internet is a for-profit company that crawls the Web to provide data for a browser toolbar (plug-in) offering data about sites being viewed, and based on data gathered from other users, suggestions of related pages that might also be of interest. It provides its crawl archives to Internet Archives since 1996 [16].

[3] Available at: http://www.afnic.fr/actu/stats/evolution[Accessed: March 16, 2008].

[4] Total number of Alexa seeds that are not in the AFNIC list.

2.3 Crawler settings

The harvest was made by Heritrix, the open-source archival quality Web crawler developed by Internet Archive with contributions from the members of the IIPC Consortium (especially Nordic national libraries) [11 and 21][5]. It is used for the BnF broad crawls since 2004 and for its focused crawls since 2006.

As it is highly configurable, Heritrix allows modifying a large number of settings, including scope, crawling priorities, filters, robot politeness…

These settings tell the crawler what it should harvest, and how. That is why they have a great impact on the collection, and why the robot should be configured according to the purposes of the crawl.

2.3.1 Scope

The scope given to the crawler defines which discovered URLs should be included in the harvest and which should be discarded.

BnF crawl scope hence includes every website in every domain name belonging to:

- The .fr TLD
- The .re TLD
- Any other domain the crawler comes across because it has been redirected from a .fr or a .re domain (398 548 redirections from AFNIC domain names were noticed during the test crawl). The crawler should take everything that is part of http://fr.yahoo.com because http://yahoo.fr redirects to http://fr.yahoo.com. However, in such a case, the crawler should remain on the same host. So, the crawler should take everything that is part of http://fr.yahoo.com but not of http://de.yahoo.com or even http://fr.news.yahoo.com.

This scope seems very large on the one hand – it includes more than one million domains – and restrictive on the other hand, as the French Web is not only hosted on .fr domains. A report published by the AFNIC quotes that less than 30% of French websites are hosted on .fr [2]. This figure is confirmed by our analysis of the 2007 election sites collection. Only 36 % of the URLs from this collection proved to behosted on the .fr domain. Even if we outline that the political websites are not fully representatives of the whole French Web (the wide use of the .org by parties and trade unions websites, for example, leads to an over-representation of this TLD), we should recognize that the .fr is not used by the majority of French websites.

[5] The International Internet Preservation Consortium was founded in 2003 by 12 institutions (Internet Archive and several national libraries) to find common solutions in order to archive and ensure a long-term access to electronic publications on the Web. Since 2007, the membership is open to new institutions. See [13] for more information on the Consortium goals and activities.

TLD	URLs	%
.fr	22938947	36.11%
.com	18373574	28.93%
.org	15955225	25.12%
.net	3690655	5.81%
.de	882656	1.39%
.info	733634	1.16%
.eu	257769	0.41%
.tv	110944	0.17%
.us	106753	0.17%
.re	99012	0.16%
other TLDs	368148	0.58%

Figure 4: Number of URLs per TLD, 2007 elections focused crawl

However, by following the redirected links our purpose was to to harvest domains other than .fr and .re. It was a way of adopting a flexible scope. A more exploratory approach would probably have led us to collectmore foreign websites: it would cause legal (BnF would have been out of the frame of the legal deposit) and economic problems: collecting non-relevant (according to our mission) websites takes up resources that would be better used harvesting surely French sites. Focusing on .fr was therefore a pragmatic choice. Moreover, as the .fr is dramatically growing, we can hope it will represent a larger part of the French Web year after year. On the other hand, we may experiment in the future new ways to discover French sites out of the .fr domain, for example by using automatic DNS (Geo-location) lookup.

2.3.2 Crawling priorities

Every URL that is considered to be in the scope of the crawl is placed in the "queue" of the crawler, that is in the list of files waiting to be crawled. As the robot will probably not be able, to harvest all the URLs it finds on its way when performing a broad crawl, the management of such a queue, and the crawling priorities given to the crawler, are key issues.

A first major decision was to choose between a "per-domain" and a "per-host" approach. With the per-host approach, used for the previous broad crawls, the URLs in the queue waiting to be crawled are grouped per hosts, and each host is treated separately. This setting leads to crawl more websites with several hosts. Commercial platforms hosting blogs or personal pages as different hosts take a lot of space within the collection. With a per-domain approach, however, these websites are treated as a single entity.

For the 2007 crawl we decided to adopt a per-domain approach. The main reason was to comply with the use of the AFNIC list, which handles only domains. As a legal deposit institution, we also wanted to give every domain the same "chance" to be harvested. Moreover, as the 2007 seed list was much bigger than the previous one, and as the amount of data to be retrieved by IA was not increasing accordingly (see below, 2.5), we feared we would obtain globally a lesser depth from the crawl. Therefore we did not want to give too big an importance to commercial or

institutional websites, that often have several hosts, and could be better harvested with focused crawls.

Big domain names like free.fr, skyblog.fr, orange.fr had to be managed as all other domains at the beginning of the crawl. The decision of a special treatment for these websites was postponed: it seemed easier to make the relevant choices during the crawl, when analyzing the reports which were to be sent by IA engineers.

It is also to avoid the over-representation of big websites that we fixed a maximum "budget" for each site: The robot was forbidden to harvest more than 10 000 URLs from the same domain. This did not mean that the crawler would be definitively stopped when this limit would be reached. The robot could be allowed to run over it if the total budget of the crawl would have not been spent.

At last, to ensure that the robot would have enough resources to harvest the entire seed list, we chose to lower the "replenish-amount": each "thread" of the robot received the order to harvest, when connecting to a seed, the 100 first URLs it discovered within the corresponding domain, and then to go further to another seed. It would come back to this seed after having finished harvesting the 100 first URLs of each seed.

This approach is a mid-term one between the breadth-first approach described in [22] and [6] which is intended to collect as many different websites as possible, and a pure archiving approach that would lead to crawl deeply into one site before harvesting the next one.

Making those decisions represented a risk for the Library. In fact, the collection could be very different from the one obtained the previous years. However, it was very important for us to experiment this way of collecting, in order to explore new ways to match the needs of our legal deposit mission.

2.3.3 Other settings

The per-domain approach we adopted had a shortcoming: it was not possible for the robot to detect automatically the second level domains (SLD). In our case, the second level domains are specialized sub-parts of the .fr. They are either "second level sector names" designed to identify an industry or a regulated sector (such as .aeroport.fr for airports, or .gouv.fr for governmental websites) or "second level descriptive domains" designed to identify an activity or title of some kind (for example, .asso.fr for federations or .tm.fr for holders of brands).
Without any specific setting, different websites hosted on the same second-level domains would have been considered by the robot as the same entity and wouldn't hav received the same budget. To avoid this problem, we gave the robot the entire list of domains hosted on SLDs (also provided to BnF by the AFNIC) with the instruction to treat them as individual websites.

Special attention was paid to content which was embedded in a page but hosted on a different domain than the one of the page itself. The robot was allowed to follow up to three "max-trans-hops", that is links embedded in Web pages. This setting was necessary to harvest a great number of video files, which was a important goal for this crawl.

Indeed we tried to find additional solutions for harvesting video files. The number of video is steadily increasing on the Web, but it is challenging for robots to collect them. The problems of harvesting video files is related to their size, and to their broadcasting mode (see [4] and [7] for an analysis of the difficulties to collect streaming media). The lack of video files, because of crawler technical limitations, would have been contradictory with our goal of archiving a "representative" image of the Web. We decided to concentrate our efforts on the two major video-broadcasting platforms used by French netsurfers: YouTube [25] and its French equivalent, Dailymotion [9]. We asked IA engineers to study Dailymotion technical architecture (they had already studied the one of YouTube very well) and to modify Heritrix scripts in order to allow the crawler to collect video files on these websites.

	2004	2005	2006	2007
Scope	Host	Host	Host	Domain
max-hops	---	100	100	100
max-trans-hops	---	3	3	3
Replenish-amount (in URLs)	---	1 000	500	100
Budget (in URLs)	---	200000	200000	10 000
delay-factor	---	5	4	4
min-delay (between two requests to the same host, in milliseconds)	---	5 000	5 000	5 000
max-delay (between two requests to the same host, in milliseconds)	---	10 000	10 000	10 000
redirect	ok	ok	ok	ok
Maximum size of downloaded URLs	---	100 Mo	100 Mo	100 Mo

Figure 5: crawler settings, from 2004 to 2007

2.4 Robots exclusion protocol

The 2006 Copyright law allows the Library to disobey the robots exclusion protocol (REP): "[Producers or publishers] shall not use codes or access restriction to prevent mandated institutions from harvesting their websites"[6]. That is the reason why BnF usually does not respect the robots.txt when performing in-house focused crawls. Robots exclusions are indeed frequently used by webmasters to prevent crawler from collecting pages that are not supposed to be indexed: images directories or CSS pages. Yet these documents can be critical for archiving crawlers, as they are necessary to display the archived Web pages in their original formin the future.

However, BnF decided to obey robots.txt rules for its 2007 .fr crawl, as it did for its previous broad crawls. The focused crawl experiences had shown that webmasters might not like it when they discover that a robot is crawling through their site, violating their own robots.txt rules. For example a French blogger, crawled by BnF sent furious e-mails to the Library and generated a crawler-trap to slow down BnF robot, and posted a message on his

[6] This sentence also means that BnF has legally the right to ask a website owner for the passwords and the codes necessary to crawl his site, for example when access to the data held by the website is not free.

blog to encourage other webmasters to create their own crawler traps[7]! In fact, Web pages forbidden by robots.txt files frequently hold crawler traps. Moreover, exclusion rules are sometimes intended to prevent crawlers from searching URLs that could overload websites (e.g. submissions to a forum). Technical problems as well as relationship difficulties with website producers can be easily managed during a focused crawl, when it is possible to monitor the harvest of each individual website. But BnF did not want to manage them on a large scale, and above all did not want the IA engineers to be overwhelmed by protests of angry French webmasters (IA's policy is to respect robots.txt exclusions anyway).

Most of the institutions performing broad crawls through the Web, especially national domain harvest, chose to respect the robots exclusion [17, 10]. Netarchive.dk, the virtual center (from the Royal Library and the State and University Library) in charge of archiving the Danish domain, prefers to overlook robots.txt, as it is mostly used for the "truly important net sites" [3][8]. In fact, websites hosting very valuable content, such as newspapers or political parties websites, are those using robots.txt exclusions most commonly [24].

2.5 Working schedule with Internet Archive

For the 2007 broad crawl, IA and BnF agreed on the targeted size of the collection. An amount of 300 000 000 URLs seemed reasonable to match the requirements of the crawl. If needed, IA could decide an up to 10-15% extension.

Both institutions agreed also on the day-to-day organization of the crawl. The scope and the major settings of the harvest should be discussed between IA and BnF, and decided by the Library. IA engineers should monitor the crawling machines, and reports should be sent twice a week to BnF (see below). The harvest should be terminated before the end of the year, and the data should be indexed (for Wayback Machine access and NutchWAX access) in the first months of 2008. The archives should then be sent on their storage rack to the Library[9]. Two IA engineers should also come to Paris to help the Library install the racks, advise BnF team and insure the quality of the collection.

[7]The Library answered quickly to the blogger protests. After having exchanged a few e-mails with BnF, the blogger recognized the utility of the web legal deposit, and decided to suppress his trap.

[8] Note that this decision leads the Danish web archivist to choose restrictive politeness rules for their robot, to avoid overloading requested servers and potential lawsuits.

[9] The racks are the Petaboxes, high density, low cost, low power storage hardware designed by Capricorn Technologies. See: http://www.capricorn-tech.com/ [Accessed: April 17, 2008].

3. CRAWLER AT WORK

3.1 Testcrawls

Before running a crawl, test crawls must be done in order to experiment robot and machine reactions and to anticipate problems. Thesetasks are actually the most important part of the quality assurance process for a crawl, along with monitoring the robots during the crawl (see 3.2) and collection characterization after the harvest (see section 4). This methodology was used for our four broad crawls.

The main task before the beginning of the crawl was to merge the different seed lists and to test them, as previously explained. Another critical step was to performthe "test crawl": the robot was launched on the seed list, to parse it and to discover a large number of URLs. The idea is not to run the crawl from the beginning to the end, but to perform it during enough time to identify non-relevant domain names, or URLs likely to be dangerous for the crawl.

For example, URLs sending a "404 error" code were discarded. Moreover, a lot of domain names that redirected to a small number of common hosts (in most cases registrars or domain squatters[10]), were removed from the crawl. Domain names used for domain farming (i.e. use of multiple domain names corresponding to one single IP address, in order to increase the PageRank of a website) were also identified.

3.2 IA / BnF relationship during the crawl

The broad crawl was launched on October 11, 2007 and was definitively terminated on November 29 (that is after the "patch-crawl"). During the crawl, IA/ BnF relationship was based on the analysis of the "frontier report". This report lists all domains in the queue, showing the number of URLs already harvested for each of them, the amount of budget expended and the URLs to be crawled. The purpose of this work was to complete the traditional expertise of IA engineers in this area with the BnF team knowledge of French seeds. This work was conducted at BnF by a librarian and an engineer.

Special attention was paid to domains reaching the maximum "budget" of 10 000 URLs. We took this number as a threshold over which we should check if the collected data was relevant or not. Non-relevant data is not, in our definition, documents of poor scientific value, but redundant files generated by pathological websites features. For example, robot traps (due to calendars or javascripts) created an infinite number of URLs for the same Web pages. Mirror websites are also a problem, as several domains host the same content under different names. In most cases, these domains were discarded.

We identified the main websites using domain farming and we excluded them. Online tests were also processed to identify which character strings corresponded to calendars, to filter them and to prevent them from generating robot traps.

The goal of this weekly control was not to ensure a 100% quality crawl – it was a broad crawl anyway. It was to manage the biggest queues as best as possible and to prevent the crawler from wasting too much time and resources.

[10] More than 100 000 domain names redirected to only three websites – two registrars and one domain squatter.

3.3 "Patch-crawl"

After 3 weeks of non-stop 24 hour crawling, IA engineers decided to stop the crawl. They launched a QA analysis on the retrieved data, to identify the domains for which the robot had not reached the fourth level of depth (that is, three hops from the seed page). These domains were crawled again (as a "patch-crawl"), when possible.

Another patch crawl was launched to retrieve video files identified during the harvest but not downloaded for some technical reasons (problems of files hosted on different domains…).

4. CRAWL OUTCOMES

Several analyses of the harvested collection were conducted after the crawl. The IA engineers who came to Paris to assist us when we installed our racks were very helpful. The first goal of this range of analyses was to make a quality control of the delivered data. We also wanted to characterize, our collections on a large scale,: what different kinds of documents were actually in the racks, what were the shape and the depth of the harvested websites… The goal was thus to quantify and to qualify our 2007 collection. At last, analyzing the outcomes of the 2007 broad crawl, and comparing them to those of previous crawls (especially the 2006 crawl), was necessary to assess the impacts of the new crawl settings and to decide if they were compliant with our legal deposit mission.

4.1 Key-figures

Number of	2006	2007
URLs	271 697 456	337 322 200
Hosts	2 928 364	1 589 458
Domains	382 540	1 062 317
(of which .fr domains)	131 136	791 940
URLs per domain	710	318
URLs per host	93	212
Unique ARC files	73 073	91 745
Compressed size of unique data (in Tb)	7,2	8,8

Figure 6: key-figures of the 2007 .fr broad crawl

The growth of the number of harvested domains was predictable, due to the dramatic increase of the seed list. Collaboration with AFNIC allowed the Library to discover and harvest six times the number of .fr domains collected in 2006.

On the other hand, fewer hosts were crawled in 2007 than in 2006, in spite of the growing number of URLs harvested by IA. This is very likely due to the per-domain approach of our last broad crawl.

The number of URLs per domain or per host is a convenient way to evaluate the "mean depth" of a crawl. However, this figure hides too many differences between websites to be significant: more detailed figures are examined in 4.6.

4.2 Distribution per MIME-type

The MIME type report of the 2007 broad crawl shows a total of 1604 different types. It was not a surprise to notice that one MIME typeonly, text/html, represents two thirds of the harvested files. Moreover, 97% of the downloaded URLs have one of the five most used MIME types: HTML, JPEG, GIF, PNG and PDF. If one looks at the MIME types of the documents harvested during the 2007 broad crawl, one could think that the French Web of 2007 mostly consisted of text and images!

We should however be very cautious with these figures. We must take into account the technical limitations of therobot. The crawler is not able to parse and to harvest every file format it discovers on the Web, even if its performances are continuously improved. Complex file formats are under-represented or simply absent from the collection.

Another good reason for being cautious is that the MIME type information used for our calculation is the one that is sent by the server. It is not really reliable. Sometimes, the server even sends a MIME type that does not exist (we discovered a surprising "application/x-something" in our collection). Out of these 1604 different MIME types, 1400 are associated with less than 500 files – we can deduce that these types are badly specified.

This problem seems to be more critical year after year. The 2004 broad crawl URLs had a total of 554 different MIME types; this figure turned to 1024 in 2006 and to 1604 in 2007.

MIME Type	URLs	%
text/html	229 257 942	67.96
image/jpeg	64 222 287	19.04
image/gif	25 376 262	7.52
image/png	3 955 885	1.17
application/pdf	3 955 463	1.17
text/plain	2 256 759	0.67
application/x-shockwave-flash	1 594 342	0.47
text/css	1 432 809	0.42
application/x-javascript	1 415 230	0.42
application/xml	1 083 991	0.32
other	2 771 213	0.82

Figure 7: the ten most-ranked 2007 broad crawl MIME types[11]

But if we cannot fully trust the MIME type of an individual file, the broad repartition given for hundreds of millions of documents is most likely to be reliable. MIME types evolution may be viewed as a way to analyze the changes and trends of the Web, on a large scale. We can observe for example, from 2004 to 2007,

[11] Note that these figures sometimes group several MIME type: for example, the number of JPEG document is given by adding the number of documents having as a MIME type "image/jpeg", "Image/jpeg" or "image/JPEG".

adecreasing use of the GIF format (the percentage of GIF images is almost divided in two in four years), in favor of JPEG and of the PNG open format[12]. The rate of XML documents is multiplied by five in the same time. The augmentation of XML files on the Web is even more obvious when we look at the number of harvested documents: from 88 000 in 2004 to one million in 2007. This is probably partly due to the growing use of RSS feeds (the correct MIME type for a RSS file is "application/rss", but "application/xml" or even "text/xml" are often used instead).

Another file format whose rating within the broad crawls collections is increasing is application/x-shockwave-flash. Two reasons could explain this augmentation: the increasing popularity of the flash format on the Web, or the better ability of Heritrix to harvest such files.

MIME Type evolution	2004	2005	2006	2007
text/html	68.11	67.22	70.15	67.96
image/jpeg	14.04	15.79	15.13	19.04
image/gif	12.70	11.09	8.05	7.52
application/pdf	1.36	1.39	1.19	1.17
image/png	0.79	0.73	0.87	1.17
text/plain	1.0833	1.19	1.01	0.67
application/x-shockwave-flash	0.2488	0.34	0.35	0.47
application/xml	0.07	0.16	0.50	0.32

Figure 8: evolution of few MIME types from 2004 to 2007

This information is very valuable from a long-term preservation perspective. It indicates on which format we should concentrate our efforts – both on the national and on the international scale. The increasing use of open formats, such as PNG or XML, is good news from this point of view.

4.3 Video files

Adding the four most used video formats (Windows media video, Quicktime, Flash video and MPEG video) makes a total of 40 000 harvested video files in 2004, against 120 000 four years later (that is 0,04% of the collection). We notice the decline of the MPEG video format, in favor of flash. In 2006 Heritrix harvested only one hundred flash video files. One year later, the script allowing our crawler to harvest content hosted on video broadcasting platforms was implemented: Heritrix collected thirty thousand documents. Our decision to focus on the video files harvesting had reached its goals: even through most of our archives contains video "holes", we did get a bigger sample of videos in 2007.

[12] Note that similar rates are observed for the .au domain (Australia, from 2004 to 2007): the percentage of GIF images is divided by two (from 10 to 5%). However, png images are less used in Australia (0,56%) than in France (1,17%) [18].

MIME Type	2004	2005	2006	2007
video/x-ms-wmv	4 408	7 705	33 936	39 218
video/quicktime	22 020	26 687	39 073	36 294
application/x-flv	0	0	104	31 556
video/mpeg	11 408	17 304	28 413	14 992
Total	39 840	53 701	103 532	124 067

Figure 9: evolution of video files MIME types from 2004 to 2007

4.4 Distribution per TLD

It was not a surprise either to discover that three-quarters of the crawled documents belonged to the .fr Top Level Domain (we can add to this figure the .re domain, which is also managed by AFNIC). Figure 10 proves however that it is possible, with the settings we chose, to start from a .fr seed list and to exceed its borders. The two types of TLDs also present in the collection are general and country codes top level domains. Most of the sites hosted under general Top Level Domains (.com, .net, .org, .info) are probably produced by French webmasters. The country code TLDs ranked in the figure 10 belongs to French-speaking countries (Belgium and Switzerland) or to France's neighbours with whom France has its main business relationships (Germany, United Kingdom): the same phenomena was noticed for Spain in [5]. The .eu domain stands for Europe.

TLD	Number of URLs	%
fr	259 869 452	77.12
com	59 843 624	17.76
net	4 951 932	1.47
org	3 171 196	0.94
de	2 808 359	0.83
eu	993 546	0.29
info	900 544	0.27
be	660 834	0.20
ch	461 021	0.14
uk	434 315	0.13
re	381 746	0.11
other TLDs	2 471 064	0.73

Figure 10: Number of URLs per TLD, 2007 broad crawl

These figures are quite similar to those obtained for the previous collections, from 2004 to 2006. Note however the progressive fall of the .biz (which was in the 10 higher-ranked TLDs the years before), and the sudden appearance of the .eu.

We notice however very different repartitions if we look at the number of domains hosted in a specific TLD.

TLD	% of domains, 2005	% of domains, 2006	% of domains, 2007
com	44,59%	42,78%	17,10%
fr	26,86%	34,28%	74,55%
net	5,37%	5,95%	2,06%
org	4,39%	4,69%	1,46%

Figure 11: percentage of domains per TLD (restricted to .com, .fr, .net and .org), from 2005 to 2007.

The 2005 and 2006 collections, shaped with similar settings (seed list as well as crawl settings) show a predominance of general TLDs over the .fr. This feature may be explained by the large number of domains, not-included in the seed list, which had been "touched" by the robot (that is where only one URL, linked to an in-scope URL, has been crawled). A lot of links to .com or .net domains were available on .fr pages, and slightly collected for this reason. They represent a large percentage of domains available within the 2005 and 2006 collection, but not in the 2007 broad crawl, because of the significant increase of the seed list size[13].

4.5 Robots.txt

In 2007, robots Exclusion Protocol prevented us from archiving 15 million files, i.e. 4.5% of the discovered files – and obviously prevented us from archiving all the documents the robot could have discovered starting from these files. These figures are quite inferior to those of the 2006 broad crawl, when robots.txt files blocked the downloading of 6% of the discovered documents. It is difficult to interpret these rates, as they are in contradiction with the latest studies on this protocol, which identify a growing use of the robots.txt [for example 23].

It is not possible, when analyzing what was not crawled, to use the MIME types of the files, as they were not sent by the server. However, we may use the file extension of the requested URLs. Nearly 40% of these files are images[14]. Many webmasters probably wanted to prevent robots from crawling these files because they were not to be indexed by search engines. This supposition is confirmed by the way our robot discovered these files: half of them (7 378 578)were found when following an embedded link.

Thus obeying robots.txt kept us from harvesting very relevant data, unnecessary for search engines robots but very useful for our archiving robot. In many cases, REP even prevented us from accessing a whole site: more than 150 000 URL used as seeds were protected by robots.txt.

4.6 Crawling depth

To assess the depth of the crawl, we may calculate the number of URLs per .fr domains.

[13] Note that the number of .com or .net domains harvested do not vary so much between 2006 and 2007. 163 632 .com domains were harvested in 2006 to 181 626 in 2007; 22 746 .net domains in 2006 to 21 853 in 2007.

[14].jpg (26%), .gif (8%), .JGP (2%), .png (2%). Note additionally that 104 109 files (1%) were CSS files.

Number of URLs	Number of domains
<10	498777
10-100	146356
100-1000	103370
1000-10000	43101
>10000	334

Figure 12: number of URLs per .fr domains, 2007 broad crawl

Nearly 50% of the harvested domains contain 10 or less URLs. This can be caused by persistent server unavailability during the crawl. However, we made several "hand-made" tests, i.e. we clicked on a dozen of domain names under the threshold of 10 URLs, to discover what was available online. These tests showed us that these websites were empty (their owner bought them to ensure that they will not be used, but do not use them) or that they were used for link-farming.

These figures were very different from those from the previous broad crawl.

Number of URLs	Number of domains
<10	38 439
10-100	32 258
100-1000	41 352
1000-10000	15 159
>10000	3 928

Figure 13: number of URLs per .fr domains, 2006 broad crawl

The three main discrepancies between 2006 and 2007 for .fr domains depth are:

- The very large number of almost empty domains in 2007, which could be explained by the increase of the seed list size.
- The number of domains where more than 10 000 URLs were archived is divided by 10 between the two crawls. This is the consequence of the "budget" restriction to 10 000 URLs; it is also probably due to the "per-domain" approach: domains containing several hosts are under-represented.
- On the other hand, this approach allowed deeper crawling of many more "small" or "medium" domains, between 100 and 10000 URLs.

These figures are probably biased by robots traps, but we currently do not know to which extent. Other quality controls, mainly visual control of individual websites, should be necessary to assess it. However it may be, these figures seem satisfying in light with our legal deposit tradition and with initial goals for this crawl: ensuring that every .fr website is included in the collection, having a strong commitment to the harvest of small and medium-size sites, possibly at the expense of bigger websites.

4.7 Big websites

We may focus on the biggest websites to refine our analysis of websites depth. The objective is to assess the reason why these sites have been more crawled than others – is it only because they are actually bigger online?

Different information were extracted from he crawl index (CDX): the list of the 50 biggest domains (from 2005 to 2007); and the list of the 1000 biggest domains in 2006 and 2007.

4.7.1 Domains

There are again huge discrepancies between the different collections. The first one is the size of big domains: only three domains have more than 1 000 000 URLs in the 2007 collection, to 25 in the 2006 collection – we recognize the features identified in the previous section, and the same reasons may explain them. The content of the 50 biggest site list is also very different: only 20% of the 50 biggest domains of the 2007 broad crawls are present in its equivalent for 2006.

The 2006 biggest website (free.fr), which was holding more than seven millions URLs, "weighs" only 40 000 URLs in 2007!

Domain name, 2007	Number of URLs	Domain name, 2006	Number of URLs
asso.fr	3 984 821	free.fr	7 405 987
com.fr	1 760 957	amiz.fr	5 194 030
tm.fr	1 270 244	asso.fr	4 036 224
gouv.fr	534 599	1rencontre.fr	3 482 657
cci.fr	495 753	sportblog.fr	2 547 231
co.uk	408 881	gouv.fr	2 360 960
nom.fr	179 256	promovacances.fr	2 113 314
presse.fr	144 207	football.fr	1 895 302
dailymotion.com	108 222	mbpro.fr	1 885 549
notaires.fr	102 018	com.fr	1 856 720

Figure 14: the 10 biggest domains, from 2005 to 2007.

In the 2007 collection, the most ranked domains are mostly second level domains, for which special settings were applied. On the other hand, we discover two kinds of websites in the 2005 and 2006 crawls: platforms hosting blogs and personal websites (free.fr, sportblog.fr) favoured by the per-host approach, and commercial websites advertising on numerous pages (e.g. promovacances.fr, an online travel agency). The 2005 broad crawl shows also several academic websites (16 on the 50 biggest domains), such as jussieu.fr or cnrs.fr. These websites almost disappeared in the 50-biggest list the following years.

This large representation of commercial websites explains the number of general top level domains within the 50 biggest domains: even for the 2007 collection, only 32% are in .fr. As it is, this collection reflects the French Web of 2007: mainly a space for business, services and social relationships.

4.7.2 Second Level Domains

	2006	2007	Evolution
asso.fr	4 036 224	3 984 821	↙
com.fr	1 856 720	1 760 957	↙
tm.fr	1 150 555	1 270 244	↗
gouv.fr	2 360 960	534 599	↙

Figure 15: evolution of some second level domains, from 2006 to 2007.

The special attention paid to second level domains during the 2007 harvest allowed the Library to crawl amounts of data slightly inferiors yet similar to those of the 2006 harvest. The most significant exception is the .gouv.fr second level domain, falling from 2.3 millions to 500 000 URLs. A way to explain this feature is the common use of third and even fourth level domains by governmental websites (for example www.rhone.pref.gouv.fr, or www.auvergne.culture.gouv.fr).

Focusing on video files brought a better harvest of video-broadcasting websites. Dailymotion entered the 10 highest-ranked domains (only 30 000 URLs were harvested on this domain in 2006). The number of files collected on YouTube doubled.

5. CONCLUSION

There is no single way to harvest a national domain. A range of technical choices (done before and during the crawl, and even after if URLs are to be discarded) shapes the collection. Even at a very large scale, collection policy applies. It is necessary to identify these choices, and to assess their consequences, in order to perform a domain crawl complying with its legal frame and its goals.

The French legal deposit mission traditionally uses three criteria to decide if a document is in or out of the scope of the collection: it should be made available to a public, on a specific form, within the borders of the French territory. All those should be adapted to the new features of the Web, and should be taken into account when performing a broad crawl.

The goal of harvesting the "French" Web explains the focus on the .fr. This is not really satisfying if we think that 50 to 60% of French websites are outside the .fr, but it is for now a pragmatic and economic choice: not all the French content is on .fr but anything on .fr is French. Moreover, this focus was flexible as the robot was allowed to follow redirects from .fr to other TLDs. At last, the .fr domain is rapidly growing with the easing offof the .fr domain attribution rules, and it will hopefully soon represent a larger part of the French domain. She should hope that this trend will not be stopped by ICANN's new ccTLD creation rules.

Focusing on .fr is also very convenient, as we are sure to be able to harvest all-comprehensively this top-level domain, thanks to the agreement with the AFNIC. This is a way to match the second major principle of our legal deposit tradition: collecting the whole intellectual production of the country, whatever its "quality" or "value", as soon as it is made available to the public. Starting from a very large number of seeds is the guarantee not to neglect poor-linked or unpopular websites.

These "non-discriminatory" principles do not mismatch with another feature of the legal deposit: the will to encompass every emerging form of publication. The Library got used to handle different supports: texts, images, sounds or videos. With Web archiving, it is not conceivable anymore to treat them separately, as they are linked elements on the same network. However, it may be necessary to find appropriate solutions for the different types of media: collecting, indexing, preserving and giving access to textual, audiovisual or interactive files do not always raise the same issues.

Some questions still remain : it is for example hard to say, up to now, if an approach focusing only on domains is better than handling each host separately, to get a representative "snapshot" of the French Web. Further analyses should be made to answer such a question and BnF would be most interested to hear about international reports on this topic. The problem, at last, is how to define a website, as this intellectual entity often does not match with the technical hosting. If we define a website as a domain name, a per-domain approach is to be adopted, as it leads to a better crawl of small or medium websites. But if we define a website as the intellectual entity created by the same author or the same editor (one or several persons, a public or private institution), sites and domains do not match anymore. Several, or even a huge number of websites can indeed be hosted under the same domain name, such as blogs hosted on a commercial platform. These sites – although relevant – became underrepresented with the strategy we used in 2007: for example, the numerous personal pages of free.fr harvested in 2006 all but disappeared a year latter. To avoid this problem, it might be possible, for future broad crawls, to adopt specific approaches for some very popular platforms hosting many personal websites or blogs, as we took care of video broadcasting platforms in 2007.

These decisions indicate some improvements we could target for the Heritrix crawler. Better abilities to parse and harvest complex file formats are one of the most necessary – Heritrix already showed itself very configurable in this perspective. Three other features are already developed in the frame of the "Smart Crawler" project, supported by IIPC, IA, the Library of Congress, the British Library and BnF, whose goal is to enhance Heritrix robot. The first one is to avoid harvesting content that has not changed since the last crawl: this deduplication feature would lead to save computing resources and storage and thus to crawl deeper the websites. The second one is to allow the robot to give priorities to some URLs within the queue. It would be very useful to mix an all-selecting approach (at the beginning of the crawl) with better automatic crawl monitoring capacities. The third enhancement, automatic recognition of the websites change frequency, would also allow the robot to identify which sites should receive special attention. This would lead us, for example, to choose broad crawl dates and frequency accordingly, or to perform focused crawls on the most frequently changing sites.

In fact, if broad crawls are supposed to harvest every website once or twice a year, with a medium depth, we should define the goal of our focused crawls accordingly: focused crawls should be primarily intended to archive big and deep websites; non .fr French websites, or/and frequently changing websites – to archive them as best as possible, as even focused crawls are often insufficient to completely harvest huge websites, and to collect documents on the hidden Web. However it may be, we should remember that snapshots are only one – if the most economic one – way to collect French digital memory, and that every decision on this matter should take into account the other archiving methods of a mixed strategy.

6. ACKNOWLEDGMENTS

We wish to acknowledge Kris Carpenter and the Internet Archive team, our partner these last four years (already!), especially Igor Ranitovic who monitored all BnF crawls from 2004 to 2007, along with John Lee, Brad Tofel and Michael Magin who helped us install the racks and analyze these new forms of collection in Paris and in San Francisco.

We are also very grateful to Mireille Chauveinc for her careful proofreading. We wish at last to express our thanks to Gildas Illien, Head of Digital Legal Deposit, for his advice and support.

7. REFERENCES

[1] Abiteboul, S., Cobena,Masanès, J. and Sedrati, G. 2002. A First Experience in Archiving the French Web. In Proceedings of the Research and advanced technology for digital libraries: 6th European conference (Italy, 2002).

[2] AFNIC. 2007. French Domain Name Industry report. 2007 Edition. AFNIC, Saint Quentin en Yvelines. http://www.afnic.fr/data/actu/public/2007/afnic-french-domain-name-report-2007.pdf

[3] Andersen, B. 2005. The DK-domain: in words and figures. Netarkivet.dk, Aarhus, Copenhagen. http://netarchive.dk/publikationer/DFrevy_english.pdf

[4] Ashenfelder, M. 2006. Web Harvesting and Streaming Media. In Proceedings of the 6th International Web Archiving Workshop (Alicante, Spain). http://www.iwaw.net/06/PDF/iwaw06-proceedings.pdf

[5] Baeza-Yates, R., Castillo, C. and Lopez, V. 2005. Characteristics of the Web of Spain. In Cybermetrics, 9. http://www.catedratelefonica.upf.es/webes/2005/Characteristics_Web_Spain.pdf

[6] Baeza-Yates, R., Castillo, C., Marin, M. and Rodriguez, A. 2005. Crawling a country: Better Strategies than BreadthFirst for Web Page Ordering. In Proceedings of the 14th international conference on World Wide Web (Chiba, Japan).

[7] Baly, N. and Sauvin, F. 2006. Archiving Streaming Media on the Web, Proof of concept and Firsts Results. In Proceedings of the 6th International Web Archiving Workshop (Alicante, Spain). http://www.iwaw.net/06/PDF/iwaw06-proceedings.pdf

[8] Brin, S. and Page, L. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine. In *Computer Networks and ISDN Systems*, 30 (1-7), 107-117. http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

[9] Dailymotion. Dailymotion – Partagez vos vidéos. http://www.dailymotion.com [Accessed: May 10, 2008].

[10] Gomes, D and Silva, M. Characterizing a National Community Web. ACM Transactions on Internet Technology (volume 5, issue 3), New York, 508-531. http://xldb.fc.ul.pt/daniel/gomesCharacterizing.pdf

[11] Heritrix. Heritrix Home Page. http://crawler.archive.org [Accessed: May 22, 2008].

[12] HTTrack. HTTrack Website Copier. http://www.httrack.com [Accessed: April 17, 2008].

[13] IIPC. International internet preservation consortium – welcome. http://www.netpreserve.org. [Accessed: May 24, 2008].

[14] Illien, G., Aubry, S., Hafri Y. and Lasfargues, F 2006. Sketching and checking quality for web archives: a first stage report from BnF. Bibliothèque nationale de France, Paris. http://bibnum.bnf.fr/conservation/index.html

[15] Illien, G. 2006. Web archiving at BnF. In International Preservation News, Paris, BnF, 27-34. http://www.ifla.org/VI/4/news/ipnn40.pdf

[16] 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.

[17] Koerbin, P. 2005. Report on the crawl and Harvest of the Whole Australian Web Domain Undertaken during June and July 2005. National Library of Australia, Canberra. http://pandora.nla.gov.au/documents/domain_harvest_report_public.pdf

[18] Koerbin, P. 2008. The Australian Web domain harvests: a preliminary quantitative analysis of the archive data. National Library of Australia, Canberra. http://pandora.nla.gov.au/documents/auscrawls.pdf

[19] Masanès, J. 2002. Towards continuous Web Archiving: First results and an agenda for the future. In D-Lib Magazine, 8 (12).http://www.dlib.org/dlib/december02/masanes/12masanes.html

[20] Masanès, J. 2006. Selection for Web Archives. In Web Archiving, J. Masanès, Ed, Springer, Berlin, Heidelberg, New York.

[21] Mohr, G., Kimpton, M., Stack, M, and Ranitovic, I. 2004. Introduction to Heritrix, an archival quality Web crawler. Paper presented at the 4th International Web Archiving Workshop (Bath, United Kingdom, 2004). http://www.iwaw.net/04/Mohr.pdf

[22] Najork, M. and Wiener, J. L. 2001. Breadth-First Search Crawling Yields High-Quality Pages. In: Proceedings of the 10th international conference on World Wide Web. Elsevier Science, Hong Kong, 114-118.

[23] Sun, Y. Zhuang Z., CouncillI. and Giles C L. 2007 Determining Bias to Search Engines from Robots.txt. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society Washington, 149-155. http://www.personal.psu.edu/yus115/docs/sun_robotstxtbias.pdf

[24] Sun, Y. Zhuang Z. and Giles C. L..2007. A large-scale study of robots.txt. In WWW '07: Proceedings of the 16th international conference on World Wide Web, ACM Press, New York 1123–1124. http://www2007.org/posters/poster1034.pdf

[25] YouTube. YouTube - Broadcast Yourself. http://www.youtube.com [Accessed: May 15, 2008].