When press is not printed: the challenge of collecting digital newspapers at the Bibliothèque nationale de France

Clément Oury
Head of Digital Legal Deposit, Legal Deposit Department
Bibliothèque nationale de France

Abstract:

Since its birth in the early seventeenth century, the press has played a prominent role in the political and social life of France. Over the two last decades, the economic and even cultural pillars on which the press ecosystem is built has been challenged by the growing use of digital technologies, and by the increasing role of the Internet as a way to distribute and access information.

Heritage libraries are affected by these major changes. They need to address the accelerating shift from analogue to digital in order to maintain the continuity of their objectives and of their missions: being able to collect and preserve cultural items, and being able to document the way these items were produced, distributed and used. Many aspects need to be taken into account: legal, scientific, technical, economic and organizational issues have to be identified and addressed.

This paper looks at the example of the National Library of France (Bibliothèque nationale de France or BnF), and at the way it has dealt with collecting newspapers in digital form. During the ten last years, the BnF has launched several experiments, testing different approaches, with varying degrees of success:

- *Direct deposit of electronic publications on physical media (CDs and DVDs) or through FTP. This way of collecting has been experimented with by BnF for some regional newspapers whose local versions were not kept in their paper form; and for which a digital substitute was searched out. This paper explains why the experiments were not conclusive.*
- *Fully automated web harvesting. Since December 2010, almost 100 news websites (national and daily newspapers, pure players, news portals…) are collected on a daily basis. This harvest gives a very good overview of the kind of information available to French Internet users, but does not allow the collection of publications for which payment is required.*
- *Web harvesting through agreements with producers. It will be showed how this third approach may act as an improvement of both previous solutions.*

Since its birth in the early seventeenth century, the press has played a major role in the political and social life of France. In the seventeenth century, struggles between the domestic and foreign press created an open field for a first kind of public debate. During the Enlightenment, diffusion of a free press was considered a key condition to achieve the philosophers' demands for freedom, political equality and justice. Newspaper and journals were indeed widely used by leaders, parties, and activists during the French Revolution, and up to the present day. In the nineteenth and twentieth centuries, newspaper owners and journalists considered that no subject was beyond their reach, from political, diplomatic or military issues to economic, cultural or sporting topics.

This prominent role explains why the press is widely used by researchers working on the history of France and other countries, even for the most recent periods. And this is also the reason why acquiring, promoting and giving access to press collections is a major objective for heritage institutions.
At the Bibliothèque nationale de France, this mission is performed for printed journals thanks to the framework of legal deposit. Legal deposit is the obligation for every producer of cultural content to send exemplars of its works to the national library. It was introduced by King François I, at a time where the invention of the printing press radically enhanced the possibility of producing and distributing books. When the first periodicals were published, in the 1630s, they were automatically, as printed elements, submitted to legal deposit, thus theoretically allowing the library to gather all French press titles – even though this objective of comprehensiveness has never been perfectly reached. Legal deposit has been progressively extended to all kind of cultural items, from engravings (1672) to radio, television and software (1992), including also music (1745), sounds (1938) and videos (1975).

The digital shift: a threat for heritage institutions' missions?

However, over the two last decades, the economic and even cultural pillars on which the press ecosystem was built has been challenged by the growing use of digital technologies, and by the increasing role of the Internet as a way to distribute and access information. Some major press titles are encountering financial difficulties. New stakeholders are emerging: some of them, the "pure players" (or newspapers that only exist online) still come from the realm of the printed press, whereas others are companies related to computing or information technologies. Finally, all press titles are thinking about their business models, ranging from fully free content to subscription-based access, with many possible variations.
Heritage libraries are necessarily affected by the major changes that affect the journals, as well as all other actors in the cultural domain. Their mission remains the same: being able to gather and preserve all cultural items, and being able to document the way these items are produced, distributed and used. These institutions need to address this accelerating shift from analogue to digital in order to maintain the continuity of their objectives and of their missions. However, they are faced in this regard with two apparently contradictory problems:

- on one hand, radically new kinds of documents are appearing, that need to be gathered and preserved. The web allows a far larger number of people to publish journals online, hence multiplying the number of titles whose memory should be kept.
- on the other hand, digital technologies also simplify and make easier the ways people produce printed documents. The number of printed titles is therefore growing (even though at a slower pace than their online equivalents), challenging the libraries ability to acquire, index and store them. For example, 40 000 different titles are currently received by the periodicals legal deposit service at BnF, representing a total number of 330 000 issues.

To tackle these issues, heritage institutions – and especially national libraries – have to find groundbreaking solutions that should at the same time be consistent with the way they deal with analogue collections. Many aspects need to be taken into account: legal, scientific, technical, economic and organizational issues have to be identified and addressed.
We propose in this paper to look at the example of the National Library of France (Bibliothèque nationale de France or BnF), and at the way it deals with collecting newspapers in digital form. During the ten last years, the BnF has launched several experiments, testing different approaches, with varying degrees of success.

Deposit of digital substitutes for printed versions

Issues and methods

In the first approach, the publishers perform a deposit of the digital version of the journal they distribute on paper form. BnF librarians started considering this solution in the early 2000s. At this time, digitization of older press collections was considered a priority for the BnF online digital library, Gallica[1] – and it is still a priority. It therefore appeared logical and necessary to also be able to keep the memory of the newer ones.
Regional daily press titles were considered the best candidates in order to test this kind of deposit. Indeed, each of these regional titles generally proposes many local versions according to the district where they are distributed. Only a few pages vary between the different local versions, but strictly following the principles of legal deposit, all of them must be kept. This represents a huge storage cost compared to a rather small benefit in terms of new content. This is the reason why, at this date, BnF was microfilming the local editions of around 20 regional titles. As there was a threat to the maintenance of microfilm companies and microfilm reading devices, digital technologies were seen as a good replacement solution. However the goal was not to digitize the paper version (as it was done in the microfilming process), but to get directly the digital version used by the publisher and the printer.

There was no legal basis to ask publishers for their digital masters. This is the reason why BnF started working with volunteers. Two regional newspapers answered positively: the *Populaire du Centre* (located in the centre of France) and the *Union de Reims* (located in eastern France). First discussions occurred in 2002, and an agreement was signed with the *Populaire du Centre* in December 2003. A few months later, discussions started with a third title, *Ouest France* (located in western France); the agreement was signed in 2005. These agreements allowed the retrieval of files from the publishers but also authorized the consultation of these files in BnF reading rooms. As they were experiments, they had a suspension date (even though it was possible to renew them).
At the same time, technical teams from BnF and from the publishing companies examined the processes to be put in place in order to get the data. Very thorough analyses were performed. It was first decided to have one PDF per page; and to use a FTP platform to exchange data between publishers and BnF. However, getting the PDF is just the easiest part. Collecting data is useless if the preservation and access issues are not taken into account.

- From a preservation point of view, it is necessary to validate the format of the files that are retrieved. This supposes automated identification and characterization of PDFs, and a way to send back the files that aren't considered satisfactory.
- from an access point of view, each delivery has to be accompanied by the metadata that will help in recreating the structure of the newspaper, and will allow the end-user to navigate within the document. This set of metadata is called the flatplan (in French *chemin de fer*, or "railway").

Many other questions were discussed: naming conventions for files, delivery schedules…
Some hardware (servers) and software (access interfaces) were to be bought or developed on the BnF side.
However, it was decided to follow as strictly as possible the procedures used for mass digitization, in order to manage both kinds of documents in the same workflows. For example,

[1] See http://gallica.bnf.fr (consulted on June 14th, 2012).

the needs of this "ingest track" were taken into account in the specifications of our digital repository, SPAR (Scalable Preservation and Archiving Repository), whose development was already scheduled (it started actually in 2006) [1].
Finally, the legal framework for this kind of deposit was strengthened in June 2006. A decree was published that allowed the library to collect a digital equivalent instead of the paper version for all "printed, graphical or photographical documents" [2]. This decree establishing legal deposit of a digital "substitute" was an important step: it meant that the digital version had the same legal value as the original, notably regarding the rights related to access and preservation (whereas the rights related to the documents received previously depended on the texts of the agreements). On the other hand, the decree stated that the digital version had to be strictly identical to the one distributed to the public: indeed, as previously said, the objective of legal deposit is not only to retrieve content, but also to document the way publications were received and used by the public.

An instructive experiment… but not a conclusive one

Everything was apparently in place for the deposit, but nevertheless the experience was not conclusive. Files were in effect delivered by publishers, but BnF teams were not able to manage them. Several problems occurred:

- First, the files received by BnF were frequently different than those that were actually printed. In fact, due to the necessity for daily newspapers to take into account important news that has just arrived (for example for sport events), changes occur between the moment the publisher validates the version and the moment the newspaper is printed; it may be challenging to get the very final version.
- Second, all modifications in the fabrication processes of the publisher resulted in a modification of the metadata delivered to BnF. It was very resource-consuming for BnF to quickly adapt to these changes, especially for modifications related to the structure of the documents (the flatplan). As each publisher had its own fabrication processes, this problem was multiplied by the number of publishers in the experiment…
- Third, developments for access interfaces were more difficult than expected on BnF side.

As a consequence, BnF teams were not able to integrate the retrieved files neither in the mass digitization workflow, nor in a specific workflow. The agreement signed with the *Populaire du Centre* ended in 2007, the one with *Ouest France* in 2008.
Even though these experiments were not successful, they were not useless. First, they were a chance for BnF teams to establish strong relationships with publishers. Second, some important lessons were learnt. It is critical to be able to have a digital version as similar as possible as the distributed one, so the file sent to BnF should be chosen at the very last step of the publisher production workflow. Moreover, the workflow designed by BnF has to be as generic as possible, in order not to depend on the evolution of the publisher's internal processes. And the process should be as automated as possible; otherwise it is too time consuming to individually check each newspaper issue. In this regard, this experience was very helpful when working on another way of collecting newspaper: the web archiving approach.

The continuation of collections by other means: the web archiving approach

From project to mission: the building of a whole document processing workflow

BnF reflections and experiments regarding web archiving started in the same years as those concerning direct deposit, in the early 2000s. As explained above, legal deposit has always been extended over time to new kinds of cultural items. At this period, it was becoming obvious that the Internet was becoming the main publishing platform, and that it was necessary to organize the safeguarding of documents distributed on the web. In order to do so, BnF (as well as many other heritage institutions that were working together in the framework of the International Internet Preservation Consortium[2]) decided to rely on the technologies already developed by search engine companies: harvesting robots. This software acts as an automated web user: starting from a list of URLs given by the human administrator, the robots follow hyperlinks and copy all pages, files, PDFs, videos, etc. that they may discover. However, the robots also follow strict rules that allow them, or not, to collect certain contents; it is thus possible to restrict the crawl to a specific domain name or set of domain names (e.g. to crawl only files hosted on the bnf.fr domain name), or to a specific Top Level Domain (e.g. to crawl only files hosted on the .fr TLD).
The first "production" crawls (i.e. the first crawls that produced content that was kept) were performed on political websites, during the presidential and parliamentary elections of 2002. The volume of archived data was modest (a few hundred gigabytes), especially compared to 2012 standards, but this project was viewed as a success because the websites were properly archived and accessible (if only for librarians at this time). Moreover, the importance of this documentation for the history of French political life was obvious – in fact, it is nowadays considered one of the most valuable web archive collections, because the data captured by BnF do not exist anymore elsewhere.
A similar project was launched in 2004, during the European and regional elections. This time BnF experimented also with cooperation with regional libraries, where the latter selected local websites and the former was in charge of selecting national websites, and of harvesting all the sites selected. The same year, thanks to an agreement with Internet Archive, BnF launched its first annual .fr domain crawl.

2006 was a decisive year from the legal point of view. In June, as written above, a decree was published allowing deposit of a digital substitute instead of the paper version. But an event of greater importance occurred a few months later. On August 1st, a law was passed that established a legal deposit of "signs, signals, writings, images, sounds or messages of any kind communicated to the public by electronic means" – i.e. a legal deposit of the Internet [2]. The experiments made by BnF (and INA for websites related to television and radio) towards the safeguarding of the memory of the Internet were thus recognized and encouraged by the legislators, who entrusted them with a dedicated mission. However, the law also states clearly that access to the collection is restricted to readers who have a demonstrable need to use them. In that sense, access is not only open to scholars but also to any person having a need to use the archive for personal or professional reasons.

[2] The mission of the International Internet Preservation Consortium (IIPC) is to acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations. It currently groups more than 40 heritage institutions. See http://netpreserve.org/about/index.php.

The same year, another step was reached. To harvest websites related to the presidential and parliamentary elections of 2007, BnF team built a permanent web archiving workflow (whereas 2002 and 2004 campaigns were run in project mode). In this framework, BnF teams launched frequent crawls of major French newspapers that were of course commenting on the electoral campaign.
In 2008, access to web archive collections was opened in the reading rooms of the "research library"[3]. In 2010, for the first time, the .fr annual domain crawl was not performed by Internet Archive, but by BnF own machine and human resources. This in-house domain crawl was for the BnF digital legal deposit team the most important event in 2010, as it showed that the library was able to tackle the challenge of harvesting millions of websites and hundreds of millions of files in a few weeks. But the mass isn't the only difficulty that faces web archiving activities; the frequency of harvest is also another issue.

Harvesting news on the web

As a matter of fact, some websites propose content whose lifecycle is very short. Harvesting this content at a rapid pace is critical in order to be able to catch the essence of the Web: its rapidity in publishing and removing information. In order to demonstrate that its workflow was also able to harvest short-lived content, BnF decided to consider harvesting news websites as a priority[4].
In the list of these news websites, it has been chosen to include the major national and local newspaper titles, but also sites that documented the way people were accessing news on the web. To this end, the Press Service (the team in charge of managing and giving access to press collections at BnF, and which belongs to the Law, Economics and Politics Department) prepared a list of 80 titles, including:

- Press agencies (Reuters, Agence France Presse…);
- National daily newspapers (*Le Monde, Le Figaro…*);
- Regional daily newspapers (*Ouest France*, *La Montagne*…);
- Magazines (*Le Nouvel Observateur*, *Le Point*);
- Portals (*Google news*, *Yahoo news*);
- Pure Players (*Rue89*, *Médiapart*);
- Internet information (data journalism such as *Owni.fr* or sites documenting the Internet as *PCInpact*).

After a few weeks of tests, the first harvest of this list of 80 websites was launched in December 2010 – and those sites have been harvested each day up to now. A few months later, 20 new titles were added in order to get a comprehensive list of national and regional newspapers online. In 2011, more than 110 millions of files have been harvested through these daily crawls, for a total sum of 0.8 TB.

Successes and limits

The web archiving workflow offers several advantages:

[3] The BnF is divided in two parts: a "public library", proposing books specifically acquired for this purpose, and a "research library", proposing all legal deposit collections.
[4] The French approach towards newspaper websites harvesting has been presented at the General Assembly of the International Internet Preservation Consortium in 2010. See [3] for BnF presentation, others are available here: http://netpreserve.org/events/singapore.php (consulted on June 14th, 2012).

- Similarity between the published content and the collected content (as required by legal deposit principles). The robot acts as an automated web user, so what is captured by the robot is what is seen by the human user. Note however that the robot may encounter technical difficulties when harvesting files online; in that case this similarity may not be ensured.
- Continuity of collections. Most French newspapers have now an equivalent on the web. In recent years, some of them have ceased their printed edition and decided to keep only their online version, generally because web publication is less expensive. This has been the case for example for French daily newspapers *France Soir* (a journal intended for a broad audience that has existed since 1944) or *La Tribune* (an economic newspaper launched in 1985). Without web archiving, BnF collections for those titles would have ended in 2012.
- Integration in a wider context. After its publication, a printed newspaper is a stable and discrete element. On the other hand, newspapers on the web are only a part (even though the most visible part) of a whole publication space. Almost all newspapers offer a way for their readers to react to the articles thanks to online comments. Archiving these comments will help future researchers understand how news was received by the public, and compare reactions to the same news in different newspapers. Moreover, newspaper companies frequently propose blog platforms for their journalists or even for their readers. They now have Facebook or Twitter accounts that change the way they communicate. All these kinds of documents should be archived in order to let future researchers write the history of the press in the early 2010s.
- Integration in an existing workflow. Archiving web versions of newspapers does not imply running specific processes; news websites are crawled the same way as online scientific publications, literary blogs or the whole list of websites in .fr. The current web archiving workflow at BnF covers harvesting of content, quality assurance, indexing, and access in reading rooms. The last step, ingest in our digital repository for long-term preservation, will be effective in the second half of 2012.
- From this point of view, this solution offers obvious economic advantages, as all technical processes are fully mutualised – only the selection process remains specific.

This rather positive overview should not hide some considerable shortcomings.
- The most obvious is the fact that robots are not able to capture automatically content that is password protected. The parts of newspaper websites that are only accessible to people that pay a subscription or a specific fee are the crawlers' *terra incognita* – or, in other words, are part of the "deep web".
- Consequently, the PDF versions of most of the newspapers (that are generally proposed in the payment-based part of the websites) are not archived. The problem mentioned in the first section of this paper – the difficulty of managing the local paper editions of regional newspapers – is therefore not solved by the web archiving solution.
- Finally, the indexation of the captured documents may be considered insufficient. Websites archived by BnF are not catalogued (unlike what is done in some other heritage institutions). Web archive access interfaces are separate from the BnF general catalogue, so newspapers captured online are not referenced in the latter. This is a problem in order to show the collection continuity: a reader accessing the bibliographical record of *France Soir* or *La Tribune* does not necessarily know that its online equivalent is still accessible in the Library.

Building an up and running web harvesting workflow, able to face both the challenges of mass and frequency, was a major objective for BnF and a considerable outcome. Harvesting

newspaper websites was only a part of the wider goal of archiving the whole French web, but it appeared that processes designed for large-scale operations were also relevant for smaller targets.
At least, up to a certain point. The goals followed when designing direct deposit processes are not fully reached by the integration of online press collections in a generic workflow; and work is still to be done in this regard. However, as direct deposit did not show the expected results, BnF has changed its approach: instead of designing specific processes, it may be more beneficial to implement specificities in an existing generic workflow.
This is the essence of the new way BnF is keen to work. It will combine some characteristics of both previous solutions: close relationships with the publisher in order to get protected content but integration in the existing web harvesting workflow.

Implementing specificities in a generic workflow: the "press project"

A pilot project with *Ouest France*…

Ouest France was still one of the best candidates to test this new approach. It is a regional daily newspaper, but its circulation is bigger than the sum of the circulations of the three best-known French national daily newspaper[5]. It is indeed the biggest circulation in France: more than 800 000 copies per day. *Ouest France* also holds a European record, with 50 local editions printed each day. These few figures explain why, on one hand, it is important to ensure *Ouest France*'s legal deposit; and on the other hand why it is very resource-consuming to get all the paper editions. Finally, BnF and *Ouest France* have an agreement in order to digitize and publish online the BnF collection of the *Ouest Éclair* newspaper, which is the ancestor of *Ouest France*. Therefore, even though BnF decided in 2008 not to get DVDs from *Ouest France* anymore, the goal of testing legal deposit of online newspapers starting from its example was maintained.
The project was actually launched again late 2011 / early 2012. BnF and *Ouest France* teams held discussions, by e-mail and face-to-face, in order to design a new workflow. Different issues were raised: what kind of content to collect, how, at what frequency. It was decided to have in parallel two kinds of harvests. First, each day, BnF harvesting robot should crawl the entirety of the web version of the newspaper (http://www.ouest-france.fr/). In fact, all content proposed in the paper edition is also freely available online. This kind of harvest is not different from that of the "press collection" described above.
However, it doesn't allow keeping the memory of the editorial layout of the print edition, so another crawl is launched at the same time. Each day, BnF harvesting robot should also crawl the PDFs editions available online but password-protected (http://www.ouestfrance-enligne.com). To this end, a specific feature of the Heritrix robot[6] has been used. It is indeed possible to indicate in the robot settings some passwords in order to access protected pages and content.
Several months of tests have been necessary in order to define the best settings:
- frequency of harvest: it has been decided to get the content every day;
- time of harvest: in the middle of the morning because *Ouest France* is a morning newspaper, and because BnF teams are available if anything goes wrong;

[5] Namely *Le Figaro*, *Le Monde*, and *Libération*. These figures were given by *Ouest France* team.
[6] Heritrix has been developed by Internet Archive in the framework of the International Internet Preservation Consortium. See https://webarchive.jira.com/wiki/display/Heritrix/Heritrix (consulted on June 14th, 2012).

- depth of harvest;
- "politeness": number of request sent by the robot in a minute. If there is a high politeness, the load on server is low but the harvest may take longer. So it has been decided to have for *Ouest France* a lesser politeness than the generic one, because *Ouest France* website was able to stand a stronger load.

Finally, these tests and the good relationship with *Ouest France* team helped in choosing crawler settings that allow the harvest of the whole content in 4 hours, every day. Each daily crawl harvests around 2 000 files, on which 1 600 PDFs, for a total sum of 850 MB (i.e. around 0.3 TB per year against around 1 TB for the other newspaper websites). This solution will enter in production mode in the middle of 2012.

This way, it is possible to tackle the two main problems that led to the abandon of the direct deposit solution:

- as the PDFs are the ones that are published online, and as this harvest is performed under the mandate of the web legal deposit (not under the mandate of the deposit of a substitute of the printed form), it is possible to be sure that the crawled version is identical to the version distributed online;
- as the PDFs are crawled by a robot, they are inserted in a whole workflow that comprises automated indexing, preservation and access.

…that has to move forward

This first experiment is only a beginning. It is now necessary to test the harvesting of protected content by robot on other examples, with other publishers. BnF team is currently defining a list of potential candidates, trying to design a representative sample: for example a national daily title, another regional title, a "pure player"…
When more newspapers are archived, it will be time to think about better ways to index them. For now, there is absolutely no relationship between the general catalogue, where the printed versions are described, and the access interfaces of the web archives. This represents an important shortcoming: if, for example, there is no way for a reader to know that the local editions of *Ouest France* are available on the web archives after a certain date, collection continuity will apparently (even if not in reality) not be ensured. Therefore a new set of technical developments will be necessary: on the web archive side, offering a way to propose permalinks to harvested content; on the catalogue side, activating hyperlinks from the catalogue interface to web archive interfaces.

In the early 2000s, BnF teams started to think about ways to perform a legal deposit of digital publications. They proceeded in two parallel and complementary directions. On one hand, the possibility of letting the publishers directly deposit their content was explored, while on the other hand the archiving of online content, offered by the technology of harvesting robots, was tested. The goals of these two experiments were different. On one hand, the goal was to help BnF and publishers reduce the number of paper versions sent to the library; on the other hand it was to gather a fully new kind of heritage. Even the legal frameworks were different, because the former deposit was supposed to be performed under the decree regarding the deposit of a digital substitute, and the latter under the law on Internet legal deposit – surprisingly, both texts were passed the same year.
Comparing both experiments, one may wonder why the first went wrong and the second succeeded. A possible explanation would be that the first solution was too dependent on the

specificities of the production workflow of each publisher. When formats, metadata schemes and technologies differ from one publisher to another, each new publisher adds a new level of complexity. With web archiving, on the contrary, the collection is made on the publishing side: as all publishers are supposed to follow the standards and the rules of the web, there is theoretically nothing different between harvesting news websites, scientific publications or literary blogs. Practically, however, some kinds of documents are not well archived by robots: dynamic files, streaming sounds and videos, password-protected content. This is the reason why it is probably time to try to find a way to combine the advantages of both approaches, using on one hand the relationships built with publishers, and on the other hand the technologies and workflows of web harvesting: a third way for legal deposit.

Bibliography

[1] BERMÈS E., FAUDUET L., PEYRARD S., A data first approach to digital preservation: the SPAR project, *Proceedings of the 76th IFLA general conference and council (Gothenburg, Sweden, 2010)* [Available online: http://www.ifla.org/files/hq/papers/ifla76/157-bermes-en.pdf (consulted on June 14th, 2012)].

[2] ILLIEN G., STIRLING P., The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future, *Proceedings of the 77th IFLA general conference and council (Puerto Rico, 2011)* [Available online: http://conference.ifla.org/past/ifla77/193-stirling-en.pdf (consulted on June 14th, 2012)].

[3] ILLIEN G., Archiving news on the Web: current projects at BnF, *Proceedings of the International Internet Preservation Consortium (Singapore, 2010)*, [Available online: http://netpreserve.org/events/2010GApresentations/04a_Gildas singapour news panel[1].pdf (consulted on June 14th, 2012)].