Harvesting digital newspapers at the Bibliothèque nationale de France

Clément Oury

Submitted on: 12/26/2014

All we need is news preservation: harvesting digital newspapers at the Bibliothèque nationale de France

Clément Oury
Legal deposit department, Bibliothèque nationale de France, Paris, France.
clement.oury@bnf.fr

Abstract

Acquiring, promoting and giving access to press collections is a major objective for heritage institutions, which need to address the accelerating shift from analogue to digital documentation in order to maintain the continuity of their missions. At the National library of France (Bibliothèque nationale de France or BnF), this mission has mainly been performed in the framework of legal deposit. In 2006, a new law on copyright extended this legal deposit to the internet: its scope covers all kinds of news websites, from digital equivalents of printed newspapers to journalists' blogs and news aggregation portals.
During the last ten years, the BnF has experimented two different approaches to ensure the preservation of online news: direct deposit of electronic publications and web harvesting of freely accessible new websites; the latter has been more successful than the former. In order to cover subscription based content, the BnF is experimenting currently a third solution, as a mix of what worked in the two first approaches: web harvesting through agreements with producers. This paper intends to present this third approach, and to explain how the BnF tried to implement it through a dedicated project, the "subscription-based press project".
This project launched in late 2012 relies on the possibility of giving the robot a login and a password, in order to let it identify itself as a subscriber. Then, the robot is able to access and copy the protected content. Even though the crawling part was technically the most critical one, this project covered all parts of the documentary lifecycle: from selection to long term preservation, including quality control, cataloguing and access in reading rooms. The paper presents the different steps of the project, its successes and achievements (in terms of collection, technical innovation and human resources), its limits, and considers its future evolutions.

Keywords: digital newspapers web legal deposit

The origins of the "subscription-based press project"

Since its birth in the early seventeenth century, the press has played a prominent role in the political and social life of France, and it is currently one of the most useful sources for historians and social scientists. Therefore, acquiring, promoting and giving access to press collections is a major objective for heritage institutions. However, over the last two decades, the economic and even the cultural pillars on which the press ecosystem was built have been challenged by the use of digital technologies, and by the increasing role of the Internet as a way to distribute and access information. A growing number of news titles are now "bi-media", i.e. they are published on paper and on the web. Due to financial difficulties, some others are now only published online; whereas some titles have been created directly on the internet.

Heritage libraries are obviously affected by these major changes. Their mission remains the same: to be able to gather and preserve all cultural items, and to be able to document the way these items are produced, distributed and used. These institutions need to address the accelerating shift from analogue to digital documentation in order to maintain the continuity of their objectives and 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 news online, hence multiplying the number of titles whose memory should be kept.
- on the other hand, digital technologies also simplify the way 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 each year.

In short, the dramatic increase in news websites does not necessarily lead to a decrease in what is available on paper. But both kinds of media are part of national cultural heritage and should be preserved for future generations.

At the Bibliothèque nationale de France, this mission is mainly performed in the framework of legal deposit. Legal deposit is the obligation for every producer of cultural content to send copies 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. It has then been progressively extended to all kinds of cultural items, from engravings to radio, television and software. In 2006, a new law on copyright created a legal deposit of the internet, which covers all kinds of news websites, from digital equivalents of printed newspapers to journalists' blogs and news aggregation portals.

Before and after the publication of this law (now part of the Heritage Code), the BnF has launched several experiments and tested different approaches to ensure the preservation of online news:

- direct deposit of electronic publications on physical media (CDs and DVDs) or through FTP. This way of collecting has been experimented with by the BnF for some regional newspapers whose local versions were not kept in their paper form; and for which a digital substitute was sought;
- fully automated web harvesting;
- web harvesting through agreements with producers.

These approaches reached varying degrees of success. The experiments towards the first solution (direct deposit) were not conclusive at the time they were conducted. The second approach, harvesting, was more successful, but it didn't manage to cover the whole scope of online news. The third solution is in a sense a mix of what worked in the two first solutions.

This paper intends to present this third approach, and to explain how the BnF tried to implement it through a dedicated project, the "subscription-based press project". This contribution is a continuation of a previous paper which was given during the IFLA satellite session of the newspapers section in Mikkeli (Finland), in August 2012 and which presented the outcomes of the two first solutions[1] [1].

Harvesting news websites with robots

The use of harvesting technologies at the BnF

To ensure its mandate of preserving online heritage, especially for the field of press and newspapers, the BnF decided to rely on web harvesting technologies. In this system, the BnF uses crawling software, also called robots. They act as 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 obey 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). These rules are the robots "parameters". An important parameter is the frequency of the crawl: it is possible to assign different harvesting frequencies according to the rate of modification of the targeted websites.

After experiments between 2000 and 2005, the BnF installed its first permanent crawling infrastructure in 2006, and continuously developed it. In December 2010, the BnF was able to collect on a daily basis around 80 news websites (national and daily newspapers, pure players, news portals…); 100 websites are currently crawled every day. This harvest (called the "news" collection) covers, for each website, the main page and web pages directly accessible from the home page. It 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. It was therefore necessary to find a way to access and crawl this protected content.

Objectives and scope of the subscription-based press project

The subscription-based press project was launched in late 2012, in order to tackle this issue. This project relies on the possibility of giving the robot a login and a password, in order to let it identify itself as a subscriber. Then, the robot is able to access and copy the protected content. Even though the crawling part was technically the most critical one, this project covered all parts of the documentary lifecycle:

- identification and selection of the news websites to be harvested as a priority;
- contacts with the website publishers in order to get ID and password;
- tests and actual crawls of the protected parts of the websites;
- quality control;
- access and promotion;

[1] We refer to this paper for a detailed explanation of the legal, scientific and organizational framework of the preservation of online news at the BnF, and for a thorough description of the different solutions experimented by the BnF from the beginning of the 21st century to 2012.

- long term preservation.

Very quickly, it appeared that the project should focus on the regional daily press, and target more specifically the PDF equivalent of the printed versions (when applicable). This decision has been taken to ensure the continuity of BnF press collections. In fact, the regional daily titles generally propose many local versions according to the district where they are distributed. Only a few pages vary between the different local versions. For example, the largest regional title, *Ouest France*, has currently more than 50 local editions. For storage reasons, it is not possible to collect the paper version of these local editions. In order to respect the principles of legal deposit, the BnF is microfilming some of them – but it is a costly activity, and probably not a long-term one as there is a threat to the maintenance of microfilm companies and microfilm reading devices. So, harvesting the digital online equivalents of these local editions (as PDF files) appeared as a reasonable replacement solution – and hence a priority for BnF robots.

The project associates all entities that deal with press titles oin printed or digital form: the Law, Economics and Politics Department (Press service), the Legal deposit department (Service of management of periodicals and digital legal deposit service), and the IT department (development service to perform the technical studies and enhance the tools, and production service to run the robots and the machines). At an operational level, it was led by the digital legal deposit service. Representatives of all entities participated in each step of the project.

The different steps

Identification and selection of the titles were of course the first part of the process. A strong focus was given to the regional daily press. Priority criteria of were: titles whose microfilming was about to cease; titles which have a large number of local editions, titles with which the BnF had existing relationships. Ensuring a regional diversity was also considered important. On the other hand, national press titles and titles existing only online were not excluded: some of them were selected according to their audience and their importance in the publishing field. Editorial and technical diversity was also sought.

Contact with publishers. To obtain IDs and passwords for their robots, BnF teams needed to identify and contact a representative of the title. Depending on the cases, this might be the director of the title, the person responsible for subscriptions, the head of digital services… In several cases, identifying and contacting the right representative has been a long process but all of them understood the legitimacy and the usefulness of digital legal deposit and agreed to give us technical information.

Collection of the content. The first attempts to collect each website met with varying degrees of success. Many tests were sometimes necessary in order to successfully gather the content. In fact, three types of outcomes were seen:

- successful crawl of the content (easily or after different technical attempts);
- unsuccessful crawl after the first tests; further developments were still required but were not carried out;
- crawl of a dedicated space, put online by the publishers for the BnF, containing the targeted content (i.e. PDF equivalents of local editions).

An unsuccessful crawl does not necessarily mean that the website was absolutely impossible to crawl; it only means that it has been decided not to use too many resources on a specific website and to turn to other important websites.

Quality control. Two kinds of quality control are performed for the web archiving workflow:

- statistical quality control: the reports and metrics of each crawl are analyzed in order to identify if something went wrong: e.g. if too few or too many URLs have been collected for a website;
- visual quality control: the archived website is visually checked; against its online equivalent when possible.

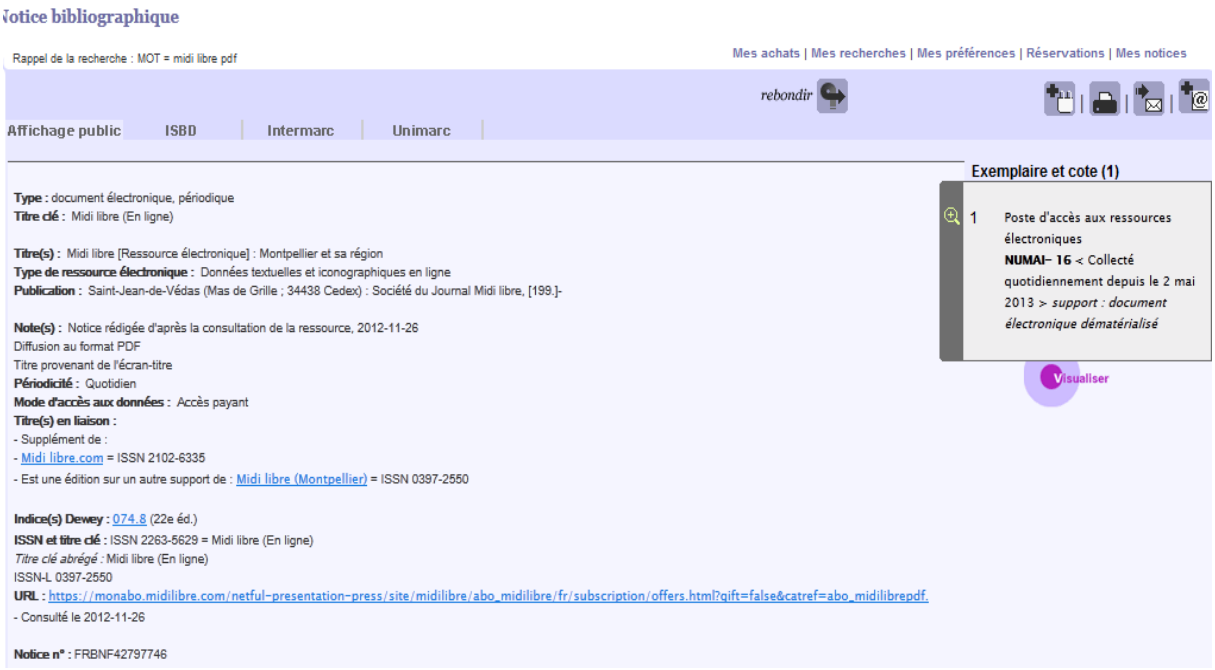
As web archiving processes retrieve millions of websites per year, an individual quality control for each website is not feasible; large-scale statistics are preferred and individual quality assurance procedures are only performed on a representative sample of the collection.
However, online press collections differ from other web archives as they correspond (especially for PDF equivalents of printed files) to resources that were previously gathered in their print form. So, as there is a very strong quality check for printed documents, it was decided to also apply strong quality assurance levels to their online equivalents. Moreover, the publication model of news websites may evolve very quickly, leading to an unsuccessful crawl. For example, if the website owner changes the URL where the PDFs are to be downloaded, the robot cannot retrieve anything. It was thus necessary to perform a daily quality control for all titles. For the titles proposing several local editions, a sampling system has been adopted: a different local edition is tested each day.
As this individual quality control procedure is very time consuming, and as it doesn't require very technical knowledge, it has been decided that it should not be the responisbility of the curators that perform statistical quality control, but rather to let other professionals be responsible for it. The library assistants that physically handle periodicals that enter through printed legal deposit are therefore now in charge of managing digital collections, as they are performing this visual quality control.

Long term preservation. As subscription-based parts of the websites are crawled on the technical channels of the "web legal deposit" track, no further developments have been necessary in order to ingest them in the digital repository of the BnF, SPAR (Scalable Preservation and Archiving Repository) [2].

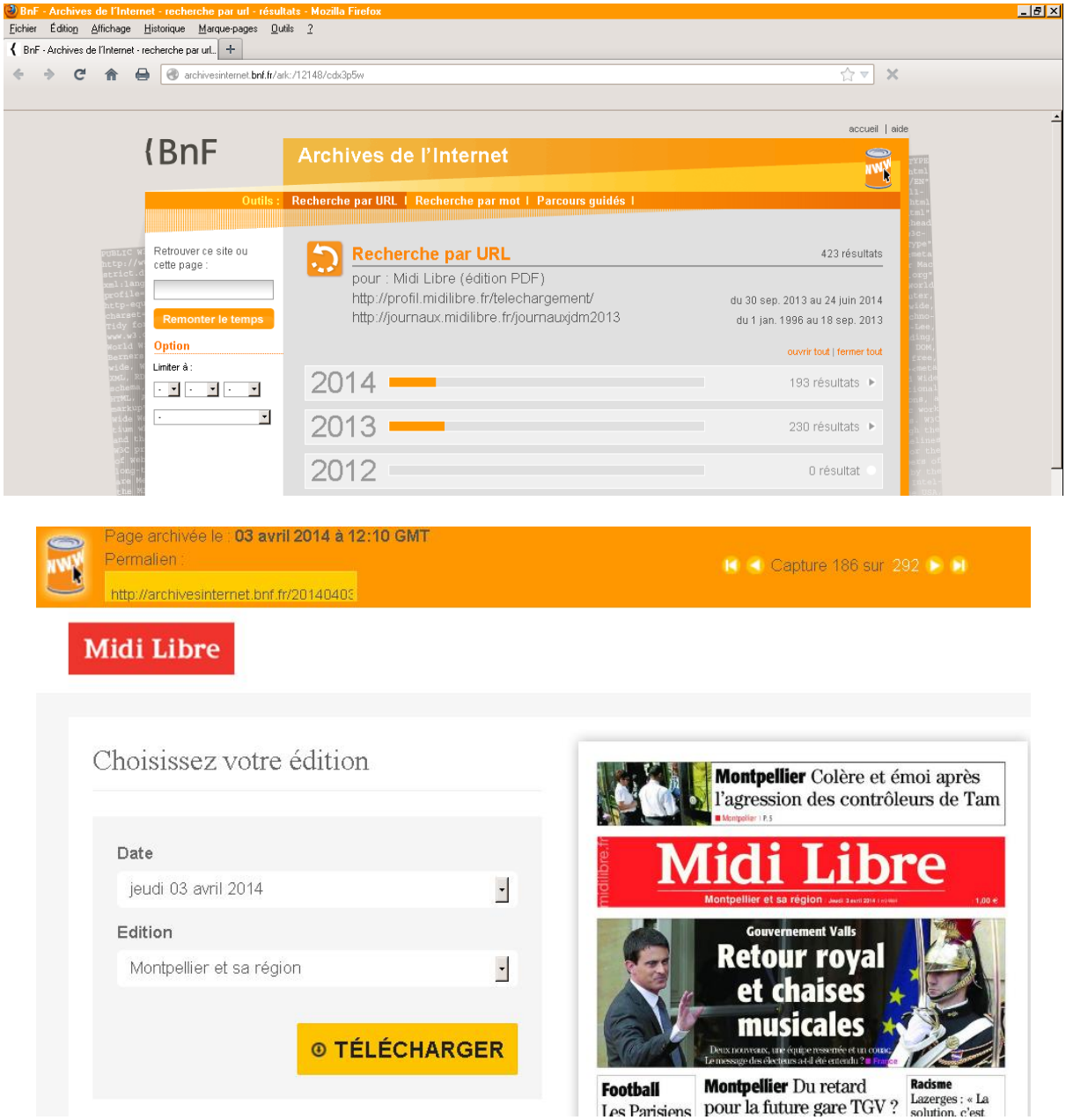
Cataloguing. At the BnF, web archives are not catalogued and therefore not searchable through the General Catalogue. This decision was taken as the scale of web archives makes it hardly possible to record all resources in the catalogue. In addition, the granularity of the description of web archives may be very complex: depending on the case, the level can be a whole collection, a website, a sub-site or even a page. It is hardly compatible with the cataloguing techniques originally designed for books and periodicals. So until recently, there was absolutely no relationship between the general catalogue, where the documents on physical media are described, and the access interface of the web archives.
However, the case of press websites is different. First, their number is limited and the question of the granularity is easier to solve: the level of description is that of the press title. Second, and more importantly, it seems critical to help readers understand the fact that different kinds of press resources are available in BnF holdings. As the web archives help the BnF ensure the continuity of their collections, it is important that the General Catalogue explains that some titles are first available on print, then on microfilms, and then as web archives.
Third, it was possible to leverage the fact that some news websites were already catalogued.The French ISSN Center was already creating records for online press titles in the BnF catalogue: it is technically necessary to do so in order to give them ISSN numbers. So, in most cases cataloguing records were already available, but they were not visible by non-professional users as they did not corrspond to resources actually held by the BnF. It was

decided to make these records visible for each title that was archived by the BnF, and to make an automated link between the cataloguing record and the web archives access interface.

One last development was required: it was necessary to establish a system of permalinks for each title. The problem is that websites frequently change their URLs (for example, http://www.metrofrance.com harvested since 2010 became http://www.metronews.fr in May 2013). In order to maintain the continuity of a press title, whatever its URL may be, each title was given an "ark" identifier[2]. This identifier refers to all URLs under which a press website has been available over time. As an example: clicking on the catalogue record of *Le Midi Libre*, a newspaper from southern France, leads to a page in the web archive interface, which refers to the dates where the URL http://journaux.midilibre.fr/journauxjdm2013 was archived, then to the dates when the newer URL of the title, http://profil.midilibre.fr/telechargement/ was archived.



[2] On the ark identifier and its use at the BnF, see http://www.bnf.fr/en/professionals/issn_isbn_other_identifiers/a.ark_en.html.



Finally, to describe all kind of news websites archived by the BnF, it was decided to apply this procedure not only for the titles crawled through the subscription-based press project, but also to all news websites harvested on a daily basis.

Promotion with the web archives access interface. Within the web archive interface itself, a specific way of promoting collections is available, called the "Guided tours". These are selections of sites prepared by BnF subject librarians, sometimes with external partners, that are intended to provide a user-friendly way of discovering the collections and also to showcase the work done by selectors. A new "tour", called "News and current events", has been added to the list of seven already existing (Elections 2002 and 2007, Literary and personal blogs, web activism, Sustainable development, Arab spring in Tunisia, Amateur images and videos, Online government services). Whereas previous guided tours presented a representative sample, this last one is currently comprehensive, listing all the 115 press titles (freely available or subscription-based) harvested everyday by the BnF.

Results and lessons learnt from the project

Collections harvested in 2013

Started in late 2012, the subscription-based press project was intended to last for one year in order to identify the feasibility of harvesting password-protected websites by robots; and to test the relevancy and the sustainability of this approach.
A summary of the outcomes of the project has been prepared at the end of 2013 in order to decide on the next steps. The figures used in this paper are drawn from this summary.
During 2013, the representatives of twelve different titles were contacted. Out of these twelve titles:

- seven are currently collected;
- two are technically impossible to collect, due to the publication technologies (embedding of content in Flash pages or use of DRM);
- the technical analyses of the last three was not fully performed; the first technical tests were not conclusive and their harvest was not considered a priority.

The negative outcomes were balanced by some very good news. On one occasion, when BnF teams worked with the technical team of a single title in order to set up a specific space where the robot could crawl the target resources, it proved that the same team was also in charge of the other titles of the same news group. Therefore, they were able to open for the BnF a space where all editions of seven different titles (instead of the one initially targeted) were made available.
At the end of 2013, the subscription-based parts of fifteen titles were currently harvested: thirteen regional daily titles (representing 112 local editions) and two national daily titles. The volume of data archived varies considerably from title to title, depending on the technical architecture of the website and on the number on local editions. It ranges from 4 URLs to 24 000 URLs and from less than 0,1 Mo to 3 Mo.

Main achievements

On the positive side, this project has represented a chance for the web legal deposit team to set up new techniques and processes that will be applied to other type of online publications:

- harvesting of password-protect content thanks to crawling robots has proved its efficiency, and may be relevant for any kind of content whose access is restricted. It is for example currently used to collect online music scores.
- cataloguing of web archives may be applied to any kind of harvested content; it may be especially relevant for blogs whose author is also the author of resources on other media.
- the system of links between the catalogue and the web archives has been reused: it is now possible to make a relationship between "BnF Archives and Manuscripts", the catalogue for BnF non-published content, and the web archives.

This project had also human resources aspects. In order to perform the visual quality control of the collections, it was necessary to involve professionals that were not used to dealing with digital material – the library assistants. A team of 3 assistants (out of 21) volunteered to experiment with these new kinds of tasks. After specific training, it appeared that this role was seen as a complement rather than a risk to the "traditional" work on periodicals. The number of agents involved is now due to increase. This experience also helped the legal deposit department to intensify its reflection on how to assist its entire staff to switch from a solely paper-based activity to a digital and paper-based activity. A general training program intended for the *ca* 140 professionals of the department was launched in the first semester of 2014, which takes into account the lessons learnt from this project.

The dark side of the crawl

As already stated, web archiving does not solve all technical difficulties involved in in letting digital newspapers enter library holdings; some press titles are impossible to harvest with crawling technologies.

But this system has other shortcomings. It is highly dependent on the website owner. Each modification of the website architecture, e.g. the use of a different publication technology, may break the entry chain and call for modifications by crawl engineers. In addition, the website owner may forget to renew the free subscription of the library, hence preventing the robot accessing protected content.

Finally, "catch-up crawls" are not easy to perform. Content is not generally held on press websites for longer than one week; the web legal deposit team must therefore be highly reactive in order to be sure not to miss any content. Some press websites even propose the daily edition only: every delay in crawling the document means therefore a gap in the collection.

The future of the project – and its alternatives

An assessment of the project was performed after one year of experiment, at the end of 2013. It was first decided to proceed with new titles, still focusing on those which were microfilmed by the BnF. From January to May 2014, seven new titles were added. Twenty-two digital newspapers are thus currently entering the BnF collections every day; of which twenty are regional newspapers, representing 194 local editions.

It was also decided to work on organizational issues. The "press project" was considered a priority for the web legal deposit team in 2013, as it was in its launch phase; but other important matters were arising in 2014. Therefore, fewer resources could be dedicated to the project. The team is now figuring out an organizational scheme that allows reactivity to handle unexpected crawling problems of news websites, while not monopolizing the time of librarians and engineers.

These new collections should also be actively promoted. New indexing and access services have been developed; however BnF readers are not sufficiently aware of them. BnF communication tools (readers' journal, website, blogs) will be used. Communication should also be directed towards the librarians in charge of reference desks, as they will act themselves as mediators and promoters of these collections.

Finally, alternative solutions are being explored. Some newspaper publishers proposed to the library to deposit PDFs on an FTP platform. As already explained, the first attempts to set up an FTP deposit system were unsuccessful. However, this was less due to technical problems than to the lack of maturity of the publishers' printing workflows; this situation has now evolved and better procedures may be put in place. Lessons may be learnt from the current design of the process for legal deposit of ebooks [3]. For press titles, two stakeholders may be identified as potential partners: either the press publishers themselves (such as for the web harvesting approach) or distributors (who will be the main technical partners of the BnF for the ebooks legal deposit).

As a conclusion, we may state that this project is successful, as it has allowed BnF to quickly harvest some press titles that it was no longer able to gather on paper or on microfilm. It is cost-effective as it provides solutions for access and long-term preservation that are shared with other kind of digital documents. Moreover, robot harvesting techniques are the only

practical way to preserve some news websites: those which are only available in HTML form. However, this solution has shortcomings and cannot be considered relevant for all kinds of press titles. Some of them are technically impossible to harvest; some websites have a rate of change so high that it becomes very expensive to adjust the crawling parameters for every modifications. Therefore, new entry tracks – such as FTP deposit – should be experimented to complement the robot solution.

Acknowledgments
The author wants to thank Géraldine Camile, responsible for the subscription-based press project, and all the teams working on digital press collections, for their dedication.

References
[1] Oury C. 2011. When *press* is *not printed:* the challenge of collecting digital newspapers at the Bibliothèque nationale de France. In *Proceedings of the IFLA Preconference, newspaper section* (Mikkeli, Finland, August 2012).
[http://www.ifla2012mikkeli.com/getfile.php?file=154 or http://halshs.archives-ouvertes.fr/docs/00/76/90/84/PDF/LegalDepositNewspapersBnF_Oury_IFLA2012.pdf]

[2] Derrot S., Fauduet L., Oury C., and Peyrard S. 2012. Preservation is Knowledge: A community-driven preservation approach. In *Proceedings of the 9th International Conference on Preservation of Digital Objects* (Toronto, Canada, October 2012).
[https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf]

[3] Derrot S., Oury C. 2014 Ebooks: rather electronic or book? Extending legal deposit to ebooks at the Bibliothèque nationale de France. To be published in the *Proceedings of the 80th IFLA Conference* (Lyon, France, August 2014).