Submitted on: 17/02/2015

Une information à préserver : la collecte de la presse en ligne à la Bibliothèque nationale de France

French translation of the original paper: "All we need is news preservation: harvesting digital newspapers at the Bibliothèque nationale de France".

Clément Oury
Département du Dépôt légal, Bibliothèque nationale de France, Paris, France.
clement.oury@bnf.fr

Traduction assurée par Ange Aniesa
Département du Dépôt légal, Bibliothèque nationale de France, Paris, France.
ange.aniesa@bnf.fr

Résumé

Acquérir, valoriser et donner accès aux collections de presse est un objectif majeur pour les instituions patrimoniales, qui doivent aborder la transition accélérée vers la documentation numérique pour maintenir la continuité de leurs missions. A la Bibliothèque nationale de France, cette mission a été principalement menée dans le cadre du dépôt légal. En 2006, une nouvelle loi sur le droit d'auteur a étendu ce dépôt légal à l'internet : son champ recouvre tous les types de sites de presse, des versions numériques des journaux imprimés aux blogs de journalistes et aux portails d'actualités.

Au cours des dix dernières années, la BnF a expérimenté deux approches différentes pour assurer la préservation de la presse en ligne : le dépôt direct des publications électroniques et la collecte de sites de presse accessibles gratuitement, cette dernière approche ayant été la plus efficace. Afin de couvrir tout le contenu accessible sur inscription, la BnF expérimente actuellement une troisième voie, une combinaison de ce qui fonctionne dans les deux premières approches : une collecte web reposant sur des accords avec les producteurs. Cet article vise à présenter cette troisième approche et expliquer comment la BnF essaie de l'installer au travers d'un projet dédié, le « projet Presse payante ».

Ce projet commencé fin 2012 repose sur la possibilité de donner au robot un identifiant et un mot de passe pour qu'il s'identifie en tant qu'abonné. Dès lors, ce robot est en capacité d'accéder au contenu et de le copier. Même si, d'un point de vue technique, l'activité de collecte s'est avérée la plus cruciale, ce projet a couvert l'ensemble du cycle de vie du document : de sa sélection à sa préservation à long terme, en passant par son contrôle qualité et sa mise à disposition dans les salles

de lecture. L'article présente les différentes étapes du projet, ses réussites (en terme de collection, d'innovation technique et de ressources humaines), ses limites, et envisage ses évolutions futures.

Mots-clés : dépôt légal de la presse en ligne par collecte web

---

Les origines du projet « presse payante »

Depuis sa naissance au début du XVII^e^ siècle, la presse a joué un rôle éminent dans la vie politique et sociale française et compte à l'heure actuelle parmi les plus précieuses sources pour les historiens et les sociologues. Par conséquent, acquérir, promouvoir et donner accès à des collections de presse est un objectif majeur pour les instituions patrimoniales. Cependant, durant les deux dernières décennies, les piliers économiques sur lesquels la presse reposait ont été remis en cause par l'usage des technologies numériques et par le rôle croissant de l'internet comme moyen de distribuer l'information et d'y accéder. Un nombre croissant de titres de presse sont désormais multi-supports, c'est-à-dire qu'ils sont publiés en version imprimé et en version web. En raison de contraintes budgétaires, d'autres sont uniquement publiés en ligne, tandis que certains y ont été directement créés.

Les bibliothèques patrimoniales sont évidemment concernées par ces changements majeurs. Leurs missions n'ont pas changé : être en capacité de rassembler et de préserver tous les types de documents et documenter la façon dont ils sont produits, distribués et utilisés. Ces institutions doivent aborder la question de la transition accélérée de la documentation imprimée vers la documentation numérique afin de maintenir la continuité de ses missions. Dans la poursuite de cet objectif, elles font face à deux problématiques *a priori* contradictoires :

- D'une part, des documents d'un genre radicalement nouveau font leur apparition et doivent être rassemblés et conservés. Le web permet à un nombre beaucoup plus important d'acteurs de publier de l'information en ligne et donc de multiplier le nombre de titres dont la mémoire doit être préservée.
- D'autre part, les technologies numériques simplifient également la production de documents imprimés. Le nombre de titres de presse imprimés est par conséquent en progression (même s'il s'agit d'une progression moins rapide que celle de ses équivalents en ligne), mettant au défi la capacité des bibliothèques à les acquérir, les indexer et les ranger. Par exemple, 40 000 titres différents sont actuellement traités par le dépôt légal des périodiques de la BnF, ce qui représente un total de 330 000 exemplaires chaque année. En résumé, l'augmentation spectaculaire des sites de presse n'implique pas nécessairement une baisse de ce qui est disponible sur papier. Cependant, les deux types de médias font partie de notre patrimoine culturel national et doivent être conservés pour les générations futures.

A la Bibliothèque nationale de France, cette activité est principalement menée dans le cadre de la mission de dépôt légal. Le dépôt légal est l'obligation pour chaque producteur de contenu culturel de fournir des copies de ses œuvres à la bibliothèque nationale. Il fut instauré par le roi François I^er^, à une époque où l'invention de l'imprimerie augmenta la possibilité de produire et de distribuer des livres. Il fut progressivement étendu à tous les types de documents culturels, des gravures jusqu'à la radio, la télévision et les logiciels. En 2006, une nouvelle loi sur la propriété intellectuelle a créé un dépôt légal de l'internet, qui couvre

notamment tous les types de sites de presse en ligne, des équivalents numériques des journaux imprimés aux blogs de journalistes et portails d'actualités.

Avant et après la publication de cette loi (qui fait désormais partie du code du Patrimoine), la BnF a lancé plusieurs projets et testé différentes approches pour assurer la conservation de la presse ligne :

- le dépôt légal direct de publications électroniques sur supports physiques ou par FTP. Cette manière de collecter a été expérimentée par la BnF pour quelques titres de presse quotidienne régionale dont les versions locales n'étaient pas déposées sous leur forme papier et pour lesquelles un substitut numérique était requis ;
- la collecte web entièrement automatisée ;
- la collecte web avec l'autorisation des producteurs.

Ces approches ont atteint des degrés variés de réussite. Les essais en faveur de la première solution n'étaient pas concluants au moment où ils ont été conduits. La deuxième approche, par collecte, était plus efficace mais ne permettait pas de couvrir l'ensemble de la presse en ligne. La troisième approche est en un sens un mélange de ce qui a fonctionné dans les deux premières.

Cet article entend présenter cette troisième approche et expliquer comment la BnF a essayé de la mettre en œuvre par le biais d'un projet dédié, le projet *Presse payante*. Cette contribution s'inscrit dans la continuité d'un article, publié dans le cadre de la session satellite de la section « presse » de l'IFLA, à Mikkeli (Finlande) en août 2012, qui présentait les résultats des deux premières expérimentations[1].[1]

La collecte des sites de presse par robots

L'utilisation des technologies de collecte par la BnF

Afin d'assurer sa mission de conservation de la mémoire en ligne, en particulier dans le champ de la presse, la BnF a choisi de se fier aux technologies de collecte du web. Avec ce système, la BnF utilise des logiciels de moissonnage, également qualifiés de robots, se comportant comme des internautes automatiques : partant d'une liste d'URL fournie par les administrateurs humains, les robots suivent les liens et copient toutes les pages ou fichiers, (PDF, vidéos, ou autres) qui se présentent à eux. Néanmoins, ils obéissent également à des règles strictement définies qui leur permettent, ou non, de collecter certains contenus. Il est ainsi possible de restreindre la collecte à un domaine ou un ensemble de domaines spécifiques (par exemple de collecter uniquement des pages hébergées par le domaine bnf.fr). Ces règles sont les « paramètres » du robot. Un paramètre important est la fréquence de la collecte : il est possible d'attribuer différentes fréquences selon le rythme de modification des sites sélectionnés.

Après les essais conduits entre 2000 et 2005, la BnF a installé sa première infrastructure permanente de collecte en 2006, qui suit un rythme de développement constant depuis lors. En décembre 2010, la BnF était en mesure de collecter à une fréquence quotidienne environ 80 sites de presse (les quotidiens nationaux, journaux tout en ligne ou « *pure players* », les portails d'actualité) ; une centaine de sites sont aujourd'hui collectés quotidiennement. Cette

[1] On pourra se reporter à cet article pour une présentation plus détaillée du cadre juridique, scientifique et organisationnel de la conservation de la presse en ligne à la BnF et pour une description plus complète des différentes solutions expérimentées depuis dix ans.

collecte (simplement appelée « collecte actualités ») couvre, pour chaque site, la page principale et les pages web directement accessible depuis cette page. Cela donne un très bon aperçu du type d'information mis à disposition des internautes français. Il était donc primordial de trouver un moyen d'accéder et de collecter ce contenu sous protection.

Les objectifs et le champ du projet Presse payante

Le projet presse payante a été lancé fin 2012, afin de se confronter à ce problème. Ce projet repose sur la possibilité de fournir au robot un identifiant et un mot de passe, pour qu'il puisse être identifié en tant membre inscrit. Le robot peut dès lors obtenir et copier le contenu protégé. Même si la partie concernant la collecte est techniquement la plus compliquée, ce projet couvre chaque étape de traitement documentaire :

- l'identification et la sélection des sites à collecter en priorité,
- le contact avec les producteurs du site afin d'obtenir l'identifiant et le mot de passe,
- les tests et la collecte effective des parties protégées des sites web,
- le contrôle de la qualité de la collecte,
- la mise en accès et la valorisation,
- la préservation à long terme.

Très vite, il est apparu que le projet devait se concentrer sur la presse quotidienne régionale et viser plus spécifiquement les équivalents PDF des versions imprimés quand ceux-ci existent. Cette décision a été prise afin d'assurer la continuité des collections de presse de la BnF. En général, les titres de presse quotidienne régionale proposent de nombreuses éditions locales en fonction de la zone géographique où ils sont distribués. Par exemple, le plus gros titre de presse régionale, *Ouest France,* a actuellement plus de 50 éditions locales. Pour des raisons d'espace de stockage physique, il n'est pas possible de collecter toutes les versions papier de ces éditions locales. Afin de respecter les principes du dépôt légal, la BnF en microfilme certains. Cependant il s'agit d'une activité coûteuse et probablement peu durable à cause de la menace qui pèse sur les fabricants de microfilms et la maintenance des appareils de lecture. Ainsi, la collecte des équivalents numériques (au format PDF) de ces éditions locales apparaît comme une solution de substitution raisonnée, d'où cette nouvelle priorité pour les robots de la BnF.

Le projet associe tous les services qui traitent de la presse sous forme imprimée ou numérique : le département Droit, économie et politique (service de la presse), le département du Dépôt légal (service des entrées et de gestion des périodiques imprimés) et le département des Systèmes d'information (service études et développements qui mène les études techniques et améliore les outils informatiques, et service support et production qui gère les robots et les serveurs). Au niveau opérationnel, le projet est piloté par le service du dépôt légal numérique. Des représentants de chaque service participent à chaque étape du projet.

Les différentes étapes

Identification et sélection

L'identification et la sélection des titres de presse ont évidemment constitué l'élément de départ. Une forte orientation vers la presse quotidienne régionale a été décidée. Les critères de priorité étaient :

- les titres dont le microfilmage allait cesser ;

- les titres ayant un grand nombre d'éditions locales ;
- les titres avec lesquels la BnF avait déjà des contacts. Assurer une diversité régionale était également considéré comme important.

Cependant, les titres de la presse nationale et les *pure players* n'étaient pas à exclure : certains ont été sélectionnés en fonction de leur diffusion et de leur importance dans le paysage de la presse. Une diversité, éditoriale comme technique, était également recherchée.

Contact des éditeurs/producteurs.

Afin d'obtenir des identifiants et des mots de passe pour les robots, les agents de la BnF avaient besoin de connaître et de contacter la personne ressource du titre de presse concerné. En fonction des titres, cela pouvait être le directeur de publication du journal, le responsable des abonnements, le chef du service numérique… Dans plusieurs cas, identifier et contacter la bonne personne fut un travail de longue haleine mais toutes ont compris l'utilité de ce dépôt légal et ont accepté de fournir des informations techniques.

Collecte du contenu

Les premières tentatives de collecte de chaque site ont rencontré des degrés de succès divers. Des tests importants étaient parfois nécessaires afin de réussir à recueillir l'ensemble du contenu. En réalité, il y eut trois types de résultats :

- la réussite de la collecte du contenu (immédiatement ou après différents essais techniques) ;
- l'échec de la collecte du contenu après des premiers tests; des développements supplémentaires étant encore requis mais n'ont pas été menés à terme ;
- la collecte d'un espace dédié, spécialement conçu par les éditeurs pour la BnF, contenant le document souhaité (soit le PDF des éditions locales).

Une collecte n'ayant pas abouti ne signifie pas forcément que le site était impossible à collecter ; cela signifie uniquement qu'il a été décidé de ne pas consacrer trop de ressources pour un site spécifique afin de pouvoir traiter d'autres sites importants.

Contrôle qualité

Deux types de contrôle qualité sont effectués dans le processus d'archivage du web :

- Un contrôle qualité statistique : les rapports et les volumes de chaque collecte sont analysés afin d'identifier si quelque chose ne fonctionne plus, c'est-à-dire si trop ou trop peu d'URL ont été collectées pour un site.
- Un contrôle qualité visuel : les pages archivées sont vérifiées dans l'application dédiée avec parfois, si possible, un comparatif entre l'archive et la page dans le web vivant.

Puisque des millions de sites sont archivés chaque année, un contrôle qualité à l'unité n'est pas réalisable et des statistiques à grande échelle sont privilégiées. La procédure de contrôle individuel est uniquement effectuée sur un échantillon représentatif de la collection. Cependant, les collections de presse en ligne diffèrent des autres archives web car elles correspondent (en particulier pour les équivalents PDF des imprimés) à des ressources qui étaient par le passé collectées sous leur forme imprimée. Ainsi, il a été décidé d'appliquer un

contrôle qualité très soutenu, similaire à celui qui était en vigueur pour les imprimés. En outre, le modèle de publication des sites d'actualité peut évoluer très rapidement et provoquer un échec de la capture. Par exemple, si le propriétaire du site modifie l'URL où se trouve le PDF à télécharger, le robot ne peut plus rien récupérer. Il était donc *de facto* nécessaire d'effectuer un contrôle qualité quotidien sur l'ensemble des sites. Pour les titres proposant plusieurs éditions locales, un système d'échantillonnage a été adopté : une édition locale différente est testée chaque jour.

Comme ce contrôle qualité peut requérir beaucoup de temps, tout en ne nécessitant pas un savoir technique poussé, il a été décidé de ne pas confier cette responsabilité aux professionnels qui effectuent déjà un contrôle statistique. Afin de transférer cette tâche à d'autres agents, il a été décidé de confier le contrôle visuel des documents numériques aux magasiniers qui par ailleurs manipulent les périodiques physiques entrant par la voie du dépôt légal des imprimés.

Préservation à long terme

Puisque les parties réservés aux abonnés des sites sont collectées *via* les chaînes d'entrée du dépôt légal du web, aucun autre développement supplémentaire n'a été nécessaire pour permettre l'entrée dans l'entrepôt numérique de la BnF, SPAR (Système de Préservation et d'Archivage Réparti) [2].

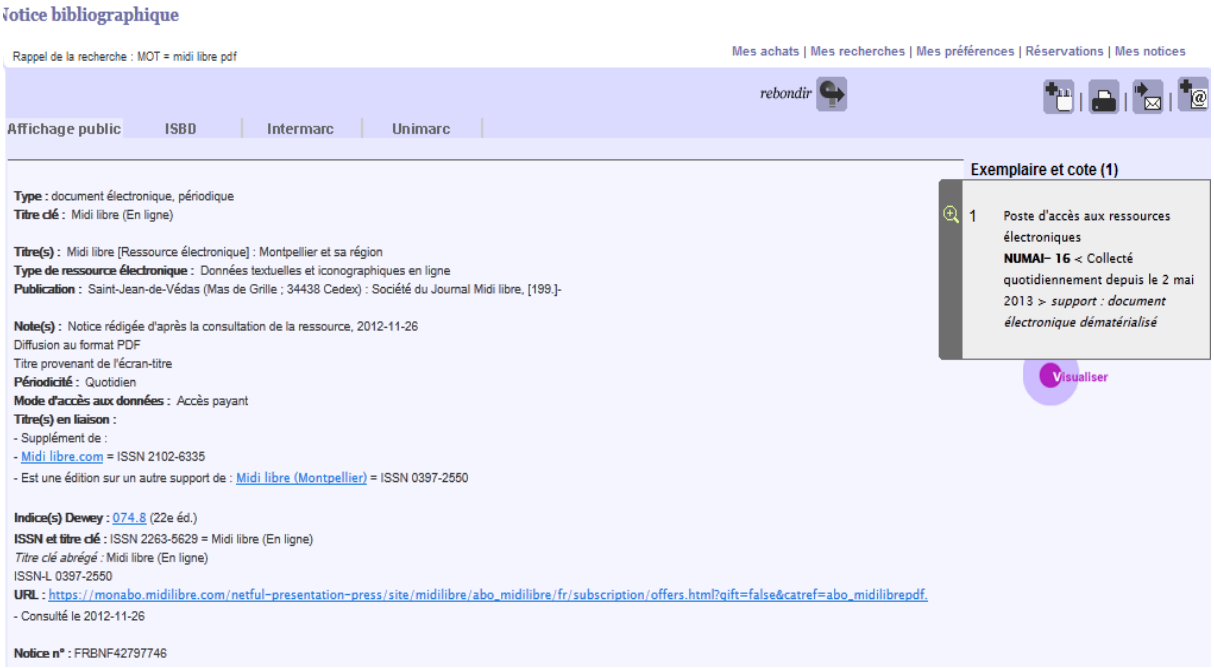
Catalogage

A la BnF, les archives de l'internet ne sont pas cataloguées, et ne sont donc pas signalées dans le catalogue général. Cette décision a été prise car l'ampleur des archives du web rend difficilement possible leur signalement complet dans le catalogue. De plus, la granularité de la description de ces archives peut s'avérer très complexe : en fonction des cas, on pourrait vouloir cataloguer un corpus complet, un site web, un hôte de ce site, voire une seule page. Cela est difficilement compatible avec des techniques de catalogage initialement pensées pour les livres et les périodiques. Ainsi, jusqu'à récemment, il n'y avait absolument aucun lien entre le catalogue général, où des documents sur support physique sont signalés, et les archives de l'internet.

Cependant, le cas des sites de presse est différent. Tout d'abord, leur nombre est limité et la problématique de la granularité est simple à résoudre : le niveau de description est celui du titre de presse. Ensuite, élément plus important, il apparaît primordial d'aider les lecteurs à comprendre le fait que différents types de ressources de presse sont disponibles dans les emprises de la BnF. Comme les archives du web sont un moyen pour la BnF d'assurer la continuité des collections, il est important que le catalogue indique que certains titres sont d'abord disponibles sous forme imprimée, puis en microfilms et enfin en version numérique.

Enfin, il était possible de mettre à profit le fait que certains sites de presse étaient déjà catalogués. Comme il est nécessaire de créer des notices bibliographiques dans le catalogue pour attribuer un numéro ISSN à un titre de presse en ligne, de nombreuses notices avaient déjà été créées par le centre ISSN France. Cependant, ces notices ne sont pas visibles par les usagers, car elles ne correspondent pas à des ressources effectivement conservées par la BnF. Il a donc été décidé de rendre visible ces notices dans le cas où elles correspondent à un titre archivé par la BnF, et de créer un lien automatique entre la notice de catalogue et l'interface des archives de l'internet.

Un dernier développement était requis : il était nécessaire d'établir un système de liens pérennes pour chaque titre. Le problème est que les sites web changent fréquemment leurs URL : par exemple, http://www.metrofrance.com collecté à cette adresse depuis 2010 est devenu http://www.metronews.fr en mai 2013. Afin de maintenir une continuité dans la visibilité du titre de presse, quels que soient ses changements d'URL, chaque titre se voit attribuer un identifiant ARK. L'identifiant ARK regroupe toutes les URL à partir desquels un titre de presse a été diffusé. Par exemple : cliquer sur la notice de catalogue du *Midi Libre* mène à une page dans l'interface des archives de l'internet : celle-ci présente à la fois l'ensemble des dates de captures de l'URL http://journaux.midilibre.fr/journauxjdm2013 au moment où cette page était la page principale, puis l'ensemble des dates de capture de l'URL http://profil.midilibre.fr/telechargement/ lorsque cette page devient la nouvelle page d'accès au titre.



En définitive, il a été décidé d'étendre cette procédure à tous les sites de presse collectés quotidiennement (collecte « Actualités ») et pas uniquement aux titres collectés dans le cadre du projet Presse payante. Cela permet d'avoir un signalement cohérent de tous les titres de presse en ligne collectés de façon systématique par la BnF.

Afin de pouvoir signaler l'ensemble des titres de presse, la procédure a finalement été étendue à tous les sites d'actualités quotidiennement collectés et plus aux seuls titres du projet Presse payante.

Valorisation des collections par le biais de l'interface des archives de l'internet

Il existe au sein même des archives de l'internet un espace spécifique de valorisation, appelé « parcours guidés ». Il s'agit de sélections de sites constituées par les bibliothécaires de la BnF, parfois avec des partenaires extérieurs, qui ont pour but de fournir à l'utilisateur une interface conviviale de découverte des collections et de mettre en avant le travail effectué par les sélectionneurs. Un nouveau parcours, nommé « Presse et actualités » a été ajouté aux sept déjà existants[2]. À l'inverse des autres parcours, dont le principe est de constituer un échantillon représentatif, le parcours « Presse et actualités » se veut exhaustif, en regroupant l'ensemble des 115 titres de presse (gratuits ou payants) collectés tous les jours par la BnF.

Résultats et leçons du projet

Les collectes de 2013

Entamé en 2012, le projet Presse payante était destiné à ne durer qu'un an, pour permettre d'étudier la faisabilité de la collecte de sites protégés par mot de passe par les robots de la BnF et de tester la fiabilité et la durabilité de cette approche.

[2] Les élections de 2002 et de 2007, la littérature et les blogs, le web militant, le développement durable, le Printemps Arabe en Tunisie, les images amateurs et les publications officielles.

Un bilan des résultats du projet a été préparé à la fin de 2013 afin de décider des prochaines étapes. Les chiffres utilisés dans cet article sont issus de ce bilan.

En 2013, des responsables de 12 titres différents ont été contactés. Sur ces 12 titres :

- sept sont actuellement collectés ;
- deux sont techniquement impossibles à collecter, en raison des technologies qu'ils utilisent (contenus embarqués sur des pages en flash ou utilisation de DRM) ;
- les analyses techniques des trois titres restants n'ont pas été complétement effectuées ; les premiers tests techniques n'ont pas été concluants et leur collecte n'a pas été considérée comme prioritaire.

Les résultats mitigés sont compensés par de très bonnes nouvelles. À une occasion, lorsque les équipes de la BnF ont travaillé avec l'équipe technique d'un titre, afin de ménager un espace spécifique où le robot pouvait collecter les contenus spécifiquement visés, il s'est trouvé que la même équipe était également en charge de la conception des sites de tout un groupe de presse. Ainsi, elle était capable d'ouvrir l'accès à la BnF d'un espace où toutes les éditions de sept autres titres étaient disponibles, au lieu du seul initialement visé.

Fin 2013, les parties à accès protégé de quinze titres étaient collectées : trente titres régionaux (représentant cent douze éditions locales) et deux titres nationaux. Le volume de données archivées varie considérablement d'un titre à l'autre, en fonction de l'architecture technique du site et du nombre d'éditions locales. Il peut aller de 4 URL à 24 000 URL et de 0,1 Mo à 3 Mo par jour.

Les principales réussites

Du côté positif, ce projet a représenté l'occasion pour les équipes du dépôt légal de mettre en place de nouvelles techniques et processus qui seront reproductibles pour d'autres types de publications :

- La collecte de contenus protégés par les robots a démontré son efficacité et peut s'avérer pertinente pour tous les types de contenus dont l'accès est restreint. Elle peut par exemple être utilisée pour collecter des partitions musicales diffusées en ligne.
- Le catalogage des archives du web peut s'appliquer à tous types de contenus collectés ; cela peut être particulièrement pertinent pour les blogs d'auteurs qui écrivent également sur d'autres types de supports.
- Le système de lien entre le catalogue et les archives de l'internet a été réutilisé : il est désormais possible de faire le lien entre le catalogue « BnF Archives et Manuscrits » et les archives de l'internet.

Ce projet avait également un versant ressources humaines. Pour mener à bien le contrôle qualité visuel des collections, des magasiniers ont été impliqués, soit des professionnels non initialement formés à travailler sur des contenus numériques. Une équipe de trois magasiniers (sur vingt et un) s'est portée volontaire pour expérimenter ce nouveau type d'activités. Après une formation spécifique, il a été perçu que ce travail intervenait plus en complément qu'en conflit avec le travail « traditionnel » sur les périodiques physiques. Le nombre d'agents impliqués va désormais augmenter. Cette expérience a aussi amené le département du Dépôt légal à intensifier sa réflexion dans le domaine de la sensibilisation de ses agents, pour les accompagner dans la transition du travail sur des collections uniquement physiques vers des collections à la fois papier et numériques. Des sessions générales de

formation pour les 140 professionnels du département ont été organisées au premier semestre de 2014, prenant en compte les conclusions issues de ce projet.

Le revers de la collecte

Comme nous l'avons vu, l'archivage du web ne résout pas toutes les difficultés techniques induites par l'entrée des journaux numériques dans les collections de la bibliothèque ; certains titres de presse sont impossibles à récupérer avec un système de collecte par robot.

Mais ce système a d'autres inconvénients. Il est très dépendant de l'éditeur du site. Chaque modification de l'architecture du site, c'est-à-dire l'utilisation d'une nouvelle technologie de publication, peut rompre la chaîne d'entrée et nécessiter le besoin d'une intervention des ingénieurs en informatique de la BnF. De plus, le producteur du site peut oublier de renouveler l'abonnement gratuit de la bibliothèque, empêchant donc les robots d'accéder aux contenus protégés.

Enfin, les collectes de rattrapage sont délicates à réaliser. Le contenu est généralement accessible sur les sites de presse pendant une semaine maximum ; l'équipe du dépôt légal numérique doit donc être très réactive afin de s'assurer de ne manquer aucun contenu. Certains sites de presse mettent également uniquement l'édition du jour à disposition : tout retard dans le processus de collecte de contenu crée donc un vide dans la collection.

La poursuite du projet et ses différentes alternatives

Une évaluation du projet a été menée après une année d'expérimentation, à la fin de l'année 2013. Il a d'abord été décidé de poursuivre et d'ajouter de nouveaux titres, en se concentrant toujours sur ceux qui étaient microfilmés par la BnF. De janvier à mai 2014, 7 nouveaux titres ont été ajoutés. 22 versions numériques de quotidiens entrent ainsi continuellement dans les collections de la BnF, dont 20 titre de presse régionale, soit 194 éditions locales.

Il a aussi été décidé de travailler sur les questions d'organisation du travail. Le projet Presse payante était considéré comme une priorité pour le service du dépôt légal numérique en 2013, lors de sa phase de lancement, mais d'autres projets importants se sont ajoutés en 2014. Par conséquent, moins de ressources ont pu être consacrées au projet. L'équipe réfléchit donc à un schéma organisationnel qui favorise la réactivité par rapport aux problèmes de collecte des sites de presse, tout en ne monopolisant pas l'emploi du temps des bibliothécaires et ingénieurs.

Ces nouvelles collections doivent être aussi activement valorisées. De nouveaux systèmes d'accès et d'indexation ont été développés ; cependant les lecteurs de la BnF ne sont pas suffisamment informés. Les supports de communication de la BnF (le site institutionnel, le blog des lecteurs et la lettre aux lecteurs) vont être investis. La communication doit également se faire auprès des bibliothécaires en salle de lecture, qui doivent effectuer le travail de médiation et de valorisation de ces collections.

Enfin, des solutions alternatives ont été explorées. Certains titres de presse proposent à la bibliothèque de déposer leurs PDF sur une plateforme FTP. Comme déjà évoqué, les premières tentatives de mise en place d'un dépôt par FTP ont échoué. Cependant, cela était moins dû à des obstacles techniques qu'au manque de maturité des circuits numériques des éditeurs ; cette situation a désormais évolué et de meilleures procédures pourraient être mises

en place. La mise en place du dépôt légal des livres numériques à la BnF pourra aussi être riche d'enseignements [3]. Pour les titres de presse, deux acteurs peuvent être identifiés comme de potentiels partenaires : les éditeurs de presse eux-mêmes (à l'instar de l'approche du dépôt légal du web) et les distributeurs (qui seront les principaux partenaires techniques de la BnF pour le dépôt légal des livres numériques).

En conclusion, il peut être établi que ce projet est un succès, car il permet à la BnF de collecter rapidement des titres de presse qu'elle n'était plus capable de rassembler sous forme papier ou de microfilm. Il réduit les coûts car il s'appuie sur des solutions d'accès et de préservation à long terme utilisées pour d'autres types de documents numériques. En outre, les techniques de collecte par robot sont l'unique moyen de préserver certains sites d'information qui sont uniquement produits en HTML. Cependant, cette solution a des défauts et ne peut pas être considérée comme adéquate pour tous les titres de presse. Certains sont impossibles à collecter ; d'autres changent si souvent qu'il devient très coûteux d'ajuster les paramètres de collecte à chaque modification. Ainsi, de nouveaux canaux d'accès – tel que le dépôt par FTP – devront être testés pour compléter le travail des robots.

Remerciements

L'auteur souhaite remercier Géraldine Camile, responsable du projet Presse payante, et tous les agents travaillant sur les collections de presse.

Références

[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 ou http://halshs.archives-ouvertes.fr/docs/00/76/90/84/PDF/LegalDepositNewspapersBnF_Oury_IFLA2012.pdf]

[2] Derrot S., Fauduet L., Oury C., et 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 ou https://hal-bnf.archives-ouvertes.fr/hal-00925315v1]

[3] Derrot S. et Oury C. 2014 Ebooks: rather electronic or book? Extending legal deposit to ebooks at the Bibliothèque nationale de France. In *Proceedings of the 80th IFLA Conference* (Lyon, France, August 2014). [http://library.ifla.org/830/1/087-derrot-en.pdf ou https://hal-bnf.archives-ouvertes.fr/hal-01059549]