La situation du dépôt légal de l'Internet en France : retour sur cette nouvelle législation, sur sa mise en pratique depuis cinq ans, et perspectives pour le futur

Peter Stirling and **Gildas Illien**
*(*Auteurs principaux*)*

Pascal Sanz and **Sophie Sepetjan**
*(*Autres auteurs, intervenants à la conférence de l'IFLA*)*
Bibliothèque nationale de France
Paris, France

Traduction en français par :
Sophie Derrot, Bibliothèque nationale de France

Meeting: *193* — e-Legal deposit: from legislation to implementation; from ingest to access — *Bibliography Section with IFLA-CDNL Alliance for Digital Strategies Programme (ICADS), Information Technology, National Libraries and Knowledge Management*

Résumé :

Cet article décrit la situation juridique française en matière de dépôt légal des contenus numériques, et montre comment celui-ci a été mis en œuvre à la Bibliothèque nationale de France. L'accent est mis sur l'archivage du Web, dans le domaine duquel la BnF bénéficie d'une expérience de presque dix ans. D'autres aspects du dépôt légal numérique sont cependant abordés, comme ses développements possibles et ses défis futurs. Des comparaisons sont faites avec la situation dans d'autres pays tout au long du propos.

Le dépôt légal de publications électroniques en ligne est un développement relativement récent, mais il s'inscrit dans une tradition de dépôt légal établie en France depuis longtemps. Cet article vise à démontrer que le dépôt légal numérique est une continuation et une évolution naturelles de la situation juridique préexistante, en même temps qu'il crée de nouveaux défis et demande un nouvel examen de certaines idées reçues concernant le dépôt légal. Il présente la situation juridique en France et la manière dont elle est mise en pratique. La responsabilité du dépôt légal est distribuée entre plusieurs institutions, mais cet article se concentre particulièrement sur la Bibliothèque nationale de France (BnF).

Nous commencerons par un bref résumé de l'histoire de la législation du dépôt légal en France, qui pose les objectifs et l'esprit de cette législation. Seront ensuite exposés les lois et les règlements spécifiques qui gouvernent le dépôt légal, et particulièrement celui des publications électroniques. La partie principale de l'article énonce ensuite les aspects spécifiques de la loi et de la pratique en quatre points :

l'acquisition, la conservation et la description des documents, ainsi que les moyens d'y accéder. Pour chaque partie, les possibilités et les restrictions légales sont posées dans le contexte des pratiques actuelles ; des comparaisons sont faites avec la situation dans d'autres pays, et nous développerons les questions ouvertes et défis à venir. La conclusion résume la situation actuelle et suggère des perspectives de développement.

I. L'histoire et le contexte du dépôt légal numérique en France

Le dépôt légal en France a été créé en 1537 par le roi François Ier, par l'ordonnance de Montpellier. Ce texte oblige les imprimeurs et les libraires à déposer à la Bibliothèque royale future Bibliothèque nationale une copie de chaque livre imprimé publié ou distribué en France. Au cours des siècles, plusieurs textes juridiques ont été mis en place pour réglementer le dépôt légal, et la législation a évolué pour couvrir les différents types et formes de publication, pour permettre une adaptation à tous les changements technologiques et sociaux majeurs. C'est particulièrement vrai en ce qui concerne les XXe et XXIe siècles, quand le développement de plusieurs innovations médiatiques a créé un grand nombre de nouvelles formes de publications, qui ont été progressivement incluses dans le champ de la législation du dépôt légal. Avec la loi de 2006 sur les droits d'auteur et droits voisins dans la société de l'information (DADVSI), l'addition la plus récente est celle des publications électroniques et de l'Internet.

Histoire du dépôt légal en France

Imprimés	1537
Estampes, cartes et plans	1648
Musique imprimée	1793
Photographies et enregistrements sonores	1925
Affiches	1941
Vidéos et documents multimédias	1975
Cinéma	1977
Multimédia, logiciels et bases de données	1992
Internet	2006

L'objectif de sauvegarde du patrimoine culturel national sous-tend la législation du dépôt légal dès ses débuts : en 1537, le libellé de l'ordonnance de Montpellier énonçait déjà l'idée de préserver les livres d'une perte pour la postérité. D'autres objectifs ont certes été suggérés, plus ou moins officiellement, comme le contrôle de l'Etat sur les publications ou la protection du droit d'auteur. Dans le premier cas, le dépôt légal aurait été considéré comme étant essentiellement un moyen de contrôle de l'État sur ce qui est publié : ce n'est pas entièrement exact, notamment parce qu'il existait des lois sur la censure assurant une surveillance des publications plus efficace que celui du dépôt légal. Néanmoins, au fil du temps le but perçu de ce dernier a changé, la notion de patrimoine culturel se mêlant à celle de surveillance par l'État. Par ailleurs, durant la période 1793-1925, le statut d'un ouvrage soumis au dépôt légal a également été utilisé pour protéger le droit d'auteur. Depuis 1925, le dépôt légal ne joue plus ce rôle en France et le code de la propriété intellectuelle, suivant la Convention de Berne, spécifie aujourd'hui que le

droit d'auteur est inhérent aux ouvrages publiés[1]. Dans certains pays, notamment aux Etats-Unis, le dépôt légal reste cependant étroitement lié à la législation du droit d'auteur[2].

La vocation patrimoniale du dépôt légal a été réaffirmée en 1992 lors d'une révision de la loi en application, dont les clauses relatives au dépôt légal furent ajoutées au code du patrimoine qui rassemble la législation relative au patrimoine culturel[3]. Le rôle culturel du dépôt légal est aussi abordé dans le décret fondant la nouvelle Bibliothèque nationale de France en 1994. Les deux premières missions de la Bibliothèque sont ainsi définies[4] :

> 1. Collecter, cataloguer, conserver et enrichir dans tous les champs de la connaissance, le patrimoine national dont elle a la garde, en particulier le patrimoine de langue française ou relatif à la civilisation française ;
>
> 2. Assurer l'accès du plus grand nombre aux collections, sous réserve des secrets protégés par la loi, dans des conditions conformes à la législation sur la propriété intellectuelle et compatibles avec la conservation de ces collections.

Plus loin dans le texte, le dépôt légal est clairement identifié comme étant un des moyens par lesquels ces missions peuvent être remplies. En tant que vocations fondamentales de la Bibliothèque, ces clauses illustrent l'esprit du dépôt légal, qui s'applique aussi bien aux publications électroniques qu'aux autres supports : le dépôt légal doit collecter tous les documents publiés en France, indépendamment du contenu, de la langue ou de la valeur, il doit les conserver sans limite de temps, et doit les rendre accessibles au public, dans des conditions qui respectent la propriété intellectuelle et ne présentent aucun risque à la conservation des matériaux.

Les articles concernés dans le code du patrimoine contrôlent avec plusieurs autres textes la façon par laquelle les documents sont collectés, conservés et rendus accessibles. Nous allons voir dans la partie suivante de quelle manière le cadre législatif applicable aux publications électroniques.

---

[1]. Code de la propriété intellectuelle, article L111-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006278868&cidTexte=LEGITEXT000006069414&dateTexte=20110520 (consulté le 20 mai 2011).

[2]. Copyright Law of the United States of America and Related Laws Contained in Title 17 of the United States Code; Chapter 4: Copyright Notice, Deposit, and Registration; Article 407. Deposit of copies or phonorecords for Library of Congress. http://www.copyright.gov/title17/92chap4.html#407 (consulté le 20 mai 2011).

[3]. Code du patrimoine, article L131-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845515&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulté le 20 mai 2011)

[4]. Décret n° 94-3 du 3 janvier 1994 portant création de la Bibliothèque nationale de France, article 2. http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006082797&dateTexte=20110520 (consulté le 20 mai 2011)

II. La législation régissant le dépôt légal numérique en France aujourd'hui

A. Le code du patrimoine, intégrant la loi sur le dépôt légal (1992) et la loi DADVSI (2006)

Le texte majeur régissant le dépôt légal en France est le code du patrimoine : au cours de l'exposé des différents aspects du dépôt légal, nous ferons régulièrement référence aux articles du Titre III, spécifiquement consacré à ce sujet. Dans la loi française, un code est un recueil de différents lois et règlements sur un domaine spécifique ; les articles sur le dépôt légal et leur intégration dans le code du patrimoine viennent principalement de la loi sur le dépôt légal de 1992. Le dépôt légal numérique quant à lui a été instauré par la loi de 2006 sur le droit d'auteur et les droits voisins dans la société de l'information (DADVSI)[5]. Cette loi est une transposition de la directive européenne sur le droit d'auteur (2001/20/CE)[6]. Elle introduit la possibilité d'un dépôt légal numérique comme exception au droit d'auteur au bénéfice des établissements dépositaires. Comme il est issu d'une directive européenne, cet acte a des similarités avec d'autres textes législatifs que l'on peut trouver dans d'autres pays européens, comme la Finlande ou le Danemark. Par conséquent, la situation légale décrite ici n'est pas spécifique à la France et peut être considérée comme assez représentative d'autres législations nationales applicables en Europe, même si des différences existent d'un pays à l'autre.

Dans l'article définissant la liste des publications sujettes au dépôt légal, la loi DADVSI introduit la phrase suivante :

> « Sont également soumis au dépôt légal les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique[7]. »

La définition des publications électroniques est énoncée en termes délibérément généraux, pour éviter de limiter la législation à des technologies spécifiques qui pourraient bientôt devenir obsolètes. La législation établit le dépôt légal de tout ce qui est publié sur l'Internet, tout en excluant la correspondance privée (courriels, intranets, parties privées de réseaux sociaux). Cela peut aller des sites Web dans le sens général du terme, aux vidéos et aux enregistrements sonores, ou bien à toute forme d'« e-publication » (« e-journaux », *e-books*, blogs, etc.) mise à disposition par voie électronique, « immatérielle ». Les publications sur un medium physique comme un CD-ROM sont déjà couvertes par le même article du code du patrimoine, ayant été inclues dans le dépôt légal par la loi de 1992.

D'autres articles couvrent la responsabilité du producteur, notamment concernant la fourniture des informations techniques nécessaires à la collecte et à la

---

[5]. Loi n° 2006-961 du 1 août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information. http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006054152&dateTexte=20110520 (consulté le 20 mai 2011)

[6]. Directive 2001/29/CE du Parlement européen et du Conseil du 22 mai 2001 sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information http://eur-lex.europa.eu/Notice.do?val=259975:cs&lang=fr&list=524158:cs,483585:cs,454158:cs,272233:cs,262958:cs,259975:cs,&pos=6&page=1&nbl=6&pgs=10&hwords= (consulté le 18 juillet 2011)

[7]. Code du patrimoine, article L131-2. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000020905828&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulté le 20 mai 2011)

conservation des documents, ainsi que les détails pratiques[8] et les conditions d'accès[9]. La loi spécifie également que les détails exacts concernant son application seront fixés par un décret. Il est important de noter qu'au moment de la rédaction de cet article, ce décret est encore en cours de validation et n'a pas été publié ; la mise en œuvre du dépôt légal numérique tel qu'il est présenté ici, bien que mis en pratique par la BnF depuis plusieurs années, doit par conséquent être considérée comme encore expérimentale. Certains développements et détails concernant sa mise en œuvre seront confirmés ou clarifiés seulement quand le décret sera publié. Référence est quelquefois faite dans cet article aux avant-projets les plus récents du décret, mais ceux-ci ne peuvent être considérés comme définitifs et les possibilités énoncées en relation avec ce décret restent de l'ordre de l'hypothèse.

B. Le décret sur le dépôt légal (1993, modifié en 2006)

Autre texte en rapport avec le dépôt légal numérique, ce décret met en œuvre la loi sur le dépôt légal de 1992, et a été modifié en 2006 pour autoriser la BnF à proposer aux éditeurs de fournir à la place du document physique un fichier numérique identique, d'une manière qui restait à établir entre la Bibliothèque et l'éditeur[10]. Comme nous le verrons en détail (III.F.), cette possibilité a été jusqu'à présent employée uniquement dans le cas des affiches publicitaires de grand format ; peu maniables et difficiles à gérer et à consulter dans leur format physique, elles sont désormais déposées sous forme de fichiers PDF. Des expériences de e□dépôt ont également été conduites avec l'un des quotidiens régionaux les plus importants, *Ouest-France*. La possibilité d'une substitution numérique peut permettre d'envisager à l'avenir beaucoup d'autres options ; néanmoins il est important de noter que cette disposition exige que la version numérique soit rigoureusement identique à celle diffusée sous forme imprimée et qu'elle autorise uniquement le remplacement du dépôt d'un document physique. Ainsi, cette alternative ne pourrait pas être utilisée pour collecter à la fois les versions électronique et papier d'un roman : cela nécessite obligatoirement qu'un choix radical soit fait par la Bibliothèque, en abandonnant la version imprimée.

C. Le décret fondant la BnF (1994)

Comme nous l'avons vu, le décret portant création de la nouvelle Bibliothèque nationale de France[11] donne au dépôt légal une place centrale parmi les missions de la bibliothèque. En fait, le cadre législatif principal pour cette mission est encore la loi sur le dépôt légal de 1992 et le décret relatif de 1993. Cependant, ce décret établissant les priorités de la nouvelle BnF renforce le statut des collections du dépôt légal comme appartenant au patrimoine national, ce qui a des implications notamment pour des questions de conservation sur le long terme.

---

[8]. Id., article L132-2-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845522&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulté le 20 mai 2011)

[9]. Id., article L132-4. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845526&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulté le 20 mai 2011)

[10]. Décret n° 93-1429 du 31 décembre 1993 relatif au dépôt légal, article 9. http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006082758&dateTexte=20110520 (consulté le 20 mai 2011)

[11]. Décret n° 94-3 du 3 janvier 1994 portant création de la Bibliothèque nationale de France, article 2, *op. cit.*

D. Le code général de la propriété des personnes publiques, le code de la propriété intellectuelle et la loi relative à l'informatique, aux fichiers et aux libertés

Bien que ne concernant pas directement le dépôt légal, trois autres textes législatifs sont importants à mentionner pour l'application pratique et la mise en œuvre de ce dernier. Le code général de la propriété des personnes publiques[12] et le code de la propriété intellectuelle[13] sont tous les deux des recueils législatifs de grande portée et plusieurs de leurs dispositions encadrent la collection, la préservation et la consultation de documents soumis au dépôt légal numérique. Enfin, une autre loi de 1978, relative à l'informatique, aux fichiers et aux libertés[14] a une relation importante à l'accès aux collections du dépôt légal numérique et à leur usage, car elle impose de strictes restrictions en regard de la protection des données personnelles qui pourraient y être inclues.

E. Résumé des possibilités législatives pour le dépôt légal numérique

Ces différents textes de loi permettent donc trois mécanismes possibles pour la collecte de documents numériques par dépôt légal :

- d'après le code du patrimoine :
  - la collecte automatique de contenus sur l'Internet (par moissonnage),
  - le dépôt de fichiers par l'éditeur (par « e-dépôt ») ;
- d'après le décret de 1993, modifié en 2006 : le dépôt d'un fichier strictement identique comme substitut au dépôt papier.

Pour des raisons à la fois économiques et patrimoniales, la BnF a jusqu'ici accordé la priorité à la collecte automatique de documents en ligne, et cet article examine en particulier cet aspect du dépôt légal numérique. Cependant, nous étudierons l'éventail complet des possibilités afin d'explorer également d'autres approches.

Dans l'ordre, nous examinerons les quatre buts du dépôt légal tels que définis par le code du patrimoine : la collecte des contenus (III), leur préservation (IV), la création de bibliographies nationales (V) et la consultation des collections (VI). Dans chaque cas, les restrictions et possibilités légales seront étudiées en relation avec les mesures pratiques déjà en place et avec celles qui pourraient être imaginées à l'avenir.

---

[12]. Code général de la propriété des personnes publiques, article L2112-1. http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006070299&dateTexte=20110520 (consulté le 20 mai 2011)

[13]. Code de la propriété intellectuelle. http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006069414&dateTexte=20110520 (consulté le 20 mai 2011)

[14]. Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068624&dateTexte=20110520 (consulté le 20 mai 2011)

III. Moyens d'acquisition de contenu numérique par le biais du dépôt légal

Le dépôt légal de publications numériques, tout en s'inscrivant dans la tradition des autres formes de dépôt légal, génère des défis inhérents à la nature de ces contenus. Comme nous l'avons vu dans la partie précédente, les textes de loi gouvernant le dépôt légal permettent la prise en compte d'un large panel de contenus électroniques ; cependant, la nature de ceux-ci signifie que les deux principes sous-tendant l'approche française du dépôt légal – l'idée de publications *rendues disponibles sur le territoire français* et le *caractère exhaustif* du dépôt légal – doivent être réinterprétés.

A. Champ des contenus soumis au dépôt légal numérique en France

Selon le code du patrimoine, tout ce qui est publié sur l'Internet en France est sujet au dépôt légal. Cela soulève la question de la circonscription de l'« Internet français » : par principe, toute information accessible sur le Web est disponible en France ; par conséquent, une définition du dépôt légal basée sur ce qui s'applique aux livres imprimés, où les publications importées sont collectées, deviendrait rapidement ingérable. La règle qui devrait être énoncée dans le décret à venir et qui est déjà appliquée par la BnF, est basée sur l'idée du lien au territoire français. Trois critères sont utilisés pour juger si une publication entre dans les limites nationales du dépôt légal numérique :

- si elle est mise à disposition sur le .fr qui est le domaine de premier niveau (en anglais « Top Level Domain », ou TLD) national, ou sur tout autre TLD enregistré dans une liste de noms de domaines basés en France (par exemple les extensions .com enregistrées en France) ;
- et/ou si le producteur du site Web (ou de tout autre document) est une personne physique ou morale domiciliée en France ;
- et/ou si le site Web est produit en France (ce dernier critère étant sujet à davantage d'interprétations que le précédent, il autorise aussi plus de flexibilité).

Il est important de noter que dans les pratiques actuelles et étant données l'échelle à laquelle la BnF opère et la limite des ressources disponibles, de telles conditions ne sont pas systématiquement vérifiées par la Bibliothèque avant la collecte des sites. Cette définition générale de l'envergure nationale est cependant prise en compte pour définir la politique générale et les paramètres techniques des collectes du Web ; ainsi, le point d'entrée principal des collectes larges nationales est constitué par les adresses de sites Web enregistrées sous l'extension .fr. Les conditions listées peuvent aussi être opposées, pendant ou après la collecte, en cas de réclamations individuelles de producteurs par exemple (voir *infra* C.).

Cela représente un nombre significatif de noms de domaine et un énorme volume de données. La question peut être cependant posée de savoir si un cloisonnement national de l'Internet a vraiment du sens, puisque les hyperliens ne respectent pas les frontières, et que l'Internet est par sa nature international. Il reste que la législation nationale du dépôt légal est un puissant moyen d'assurer la préservation de l'Internet sur une grande échelle, en donnant des moyens légaux de copie et de préservation des contenus, en mobilisant les ressources des bibliothèques et Archives nationales et en considérant l'archivage de l'Internet comme une part constitutive de la préservation du patrimoine national. Cette

division par pays n'en pose pas moins la question de la collaboration internationale et de l'interopérabilité entre les collections (voir F.).

B. Les institutions responsables du dépôt légal numérique

Le code du patrimoine partage la responsabilité du dépôt légal entre trois institutions culturelles : la Bibliothèque nationale de France, l'Institut national de l'audiovisuel (INA) et le Centre national du cinéma et de l'image animée (CNC)[15]. Concernant le dépôt légal numérique en particulier, le décret à venir devrait définir la distribution des responsabilités entre la BnF et l'INA. Entre-temps, une division *ad hoc* a été acceptée entre ces deux institutions, suivant la logique de la continuité de leurs mandats et collections respectifs : l'INA collecte les publications en ligne émises en France relatives à la télévision et à la radio et la BnF collecte tous les autres contenus. Cette division devrait être fixée plus précisément dans le décret à venir. Dans notre article, l'attention portée aux aspects pratiques de la collecte des contenus en ligne est basée sur l'expérience de la BnF ; l'INA a une approche différente basée sur des collectes bien plus fréquentes d'un nombre de sites plus restreint, avec une focalisation importante et complémentaire sur les médias en *streaming*[16].

Tandis que la responsabilité légale des collections et de leur diffusion reviennent à la BnF et à l'INA, d'autres institutions et organisations peuvent être impliquées dans le processus, particulièrement lorsqu'il est question d'une sélection de contenus à collecter. La BnF a déjà mis en place une coopération expérimentale avec les 25 bibliothèques régionales chargées de recevoir le dépôt légal des imprimeurs (bibliothèques de dépôt légal imprimeur, BDLI) : ces bibliothèques ont été impliquées dans la sélection de sites de leurs régions respectives, destinés à être archivés lors des élections nationales ou régionales[17]. Les chercheurs et les spécialistes de différentes organisations (universités, associations, etc.) ont également été impliqués dans la sélection de sites pour d'autres thématiques ou dans le cadre de projets et d'ensembles de données sur des événements tels que le Web militant, la littérature en ligne ou le développement durable. De telles possibilités devraient être explorées plus avant, bien qu'il y ait des implications en termes de possibilités d'accès aux collections, qui seront abordées plus loin.

Il existe en France d'autres initiatives dans le domaine de l'archivage du Web, en dehors du contexte législatif du dépôt légal. Certains chercheurs et universités sont activement engagés dans des projets de recherche et de développement concernant le Web, et cela peut impliquer l'archivage de contenus en ligne. Les démarches plus actives sont conduites par la fondation Internet Memory[18] (anciennement fondation European Archive), à but non lucratif, qui œuvre pour la préservation de l'Internet, et plus récemment, le Médialab de Sciences Po[19] dans le

---

[15]. Code du patrimoine, article L132-3. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000020967935&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulté le 20 mai 2011).

[16]. Pour plus d'informations, voir le site Web de l'INA : http://www.institut-national-audiovisuel.fr/nous-connaitre/entreprise/statut.html (consulté le 26 juillet 2011).

[17]. HUCHET Bernard, ILLIEN Gildas, OURY Clément, « Le Temps des moissons. Le dépôt légal du Web : vers la construction d'un patrimoine coopératif », *Revue de l'Association des Bibliothécaires de France*, 2010, n° 52, p. 28-31.

[18]. Fondation Internet Memory, http://internetmemory.org/fr/ (consulté le 26 juillet 2011).

[19]. Le médialab de Sciences Po, http://www.medialab.sciences-po.fr/index.php?page=accueil (consulté le 26 juillet 2011).

domaine des sciences sociales. Cependant, seuls l'INA et la BnF peuvent bénéficier des dispositions spécifiques attachées à la législation du dépôt légal, en particulier la possibilité de collecter des sites Web sans demander la permission des producteurs.

Suivant la tradition française de centralisme administratif et culturel, ce sont les mandats respectifs de la BnF et de l'INA, tous deux localisés à Paris, qui conditionnent la distribution des tâches entre institutions patrimoniales. En revanche, d'autres pays doivent éclaircir la distribution des tâches selon une perspective différente. Par exemple, la division entre la bibliothèque nationale et les archives nationales (c'est le cas en Grande-Bretagne, par exemple) est une question fréquemment rencontrée. De même, dans le cas d'administrations fédérales il faut envisager des schémas organisés de manière coopérative, comme les réseaux de bibliothèques régionales et spécialisées (comme en Suisse ou en Allemagne). Indépendamment des législations spécifiques, d'autres formes de réseaux peuvent se développer, comme aux États-Unis : sous la tutelle du programme national d'infrastructure et de préservation numériques (National Digital Infrastructure and Preservation Program, NDIIPP)[20] mené par la Bibliothèque du Congrès, on peut trouver une variété d'institutions activement engagées dans l'archivage du Web, telle que la fondation à but non lucratif Internet Archive, la Bibliothèque numérique de Californie et l'université du Texas du Nord.

C. Questions d'échelle et d'exhaustivité

Le dépôt légal visait précédemment à un idéal d'exhaustivité : les collections en résultant devaient compter tout ce qui était publié ou importé en France, en tenant compte de critères définis. Toutefois, l'extension du dépôt légal aux contenus numériques implique une remise en question de cet idéal. La définition des contenus électroniques dans le code du patrimoine est formulée, comme nous l'avons vu, de manière à être indépendante de tout *format* précis (*e-books*, Web, etc.) ; elle met plutôt l'accent sur le *contenu* qui est communiqué par des moyens électroniques. Cela élargit le champ du dépôt légal pour inclure tout ce qui est publié sur le Web et qui rencontre les critères décrits, c'est-à-dire la notion de territoire et la nature publique de la communication. Mais la nature même du Web semble opposée à toute idée de collecte exhaustive et crée donc une difficulté.

Le problème vient d'abord de la quantité d'informations disponibles en ligne. En avril 2011, le nombre de noms de domaines enregistrés en .fr était d'environ 2 millions, auxquels doivent être ajoutés les sites également compris dans le spectre du dépôt légal et enregistrés avec d'autres TLD, notamment les .com, .org et .net. L'AFNIC, l'institution en charge de l'administration du .fr, estime que celui-ci ne représente qu'un tiers de l'« Internet français », utilisant une définition très semblable à celle appliquée par les lois sur le dépôt légal[21]. Alors que la collecte du domaine national (.fr) menée par la BnF en 2010 a montré qu'une large proportion de ces noms de domaine n'avait pas ou peu de contenu, certains sites importants contiennent plusieurs millions de fichiers.

---

[20]. Site Web du NDIIP : http://www.digitalpreservation.gov/ (consulté le 20 mai 2011).

[21]. AFNIC, Observatoire 2010 du marché des noms de domaine en France, p. 20-22. http://www.afnic.fr/data/actu/public/2010/afnic-observatoire-domaines-france-2010.pdf (consulté le 27 juillet 2011).

Distribution des domaines en termes de nombre d'URL collectées par domaine, collecte large 2010 de la BnF

Nombre d'URL collectées	Nombre de domaines
≤ 10	976 948
11-100	580 362
101-1 000	320 620
1 001-10 000	85 471
10 001-50 000	23 630
50 001-100 000	352
≥100 001	230

Le Web est moins un ensemble fractionné de publications individuelles et séparées qu'un espace d'information avec des frontières mouvantes, où il est difficile de définir des « items » ou « unités » distinctes et stables, comparables à un livre ou à un article dans un périodique. Un site Web peut contenir de multiples pages, images, fichiers vidéo ou audio, des documents sous la forme de PDF ou de documents Word, des applications… De plus, la nature du Web réside dans l'utilisation de liens à l'intérieur et entre les sites, de sorte que la majorité de l'information prend tout son sens au sein d'un réseau complexe de liens interconnectés. Pour ajouter à cette complexité, il y a un constant flux d'information, puisque les sites sont mis à jour à une fréquence qui varie entre les sites et en leur sein. Tout cela signifie que pour être véritablement exhaustif, il serait nécessaire de collecter tout en permanence ; la technologie de la collecte du Web et l'espace de stockage impliqué rendent cela impossible. En outre, les collections générées seraient énormes et ingérables, à la fois pour les bibliothécaires et les utilisateurs finaux.

Confronté à cette impossibilité, la seule réponse est d'abandonner l'idéal d'exhaustivité et d'accepter que le dépôt légal du Web ne collecte qu'une partie de ce qui est disponible. Quant à la mise à jour des contenus en ligne, le décret à venir devrait prendre acte de ce problème, en spécifiant que les sites devraient être collectés « au moins une fois par an ». Cependant, même dans ce cas, la masse des contenus signifie qu'une collecte exhaustive, même une fois par an, reste infaisable. On peut envisager à partir de là deux approches : la *sélection* et l'*échantillonnage*. La première option implique une sélection des sites à collecter en amont, habituellement sur la base d'un jugement de la qualité ou de la valeur scientifique ou esthétique du site ; il pourrait être ainsi décidé que des sites publiant de la recherche scientifique, des publications officielles ou gouvernementales, ou bien des travaux littéraires ou artistiques sont de plus grande valeur et devraient par conséquent être au centre de la collecte. Cette approche est analogue à l'acquisition d'ouvrages choisis par un bibliothécaire, avec une logique de sélection d'unités qui enrichissent les collections de recherche. L'approche alternative, l'échantillonnage, est proche du principe du dépôt légal : des sites sont collectés sans jugement préalable sur leur « valeur » ou de leur intérêt potentiel pour les chercheurs d'aujourd'hui ou de demain. Le but est plutôt de préserver un échantillon représentatif de la production nationale née numérique, qui présente le « caractère » du Web national à un moment donné.

Chaque approche a ses limites : la sélection requiert la définition de critères et un investissement de temps pour les conservateurs, chercheurs et autres, avec la possibilité que les sites sélectionnés aujourd'hui ne soient plus considérés comme les plus importants par les utilisateurs du futur. L'échantillonnage, d'un autre côté,

signifie que des sites importants peuvent n'être collectés que partiellement ou pas du tout, alors que l'on pourrait argumenter que la majorité de ce qui est collecté n'a aucun intérêt pour les chercheurs, comme le contenu qui peut être vu comme du déchet (les spams, le cybersquattage, etc.) ou comme ayant peu de valeur (les blogs personnels, la publicité, les sites commerciaux, etc.).

À la BnF, la décision a été prise de combiner les deux approches et d'adopter un « modèle mixte » pour l'archivage du Web, qui combine la sélection et un échantillonnage à grande échelle.

D. La collecte du contenu en ligne

Pour respecter les obligations de dépôt légal tout en acceptant les réalités du Web, la BnF a mis en place depuis 2006 ce « modèle mixte » d'archivage du Web, qui combine deux types de collectes : les collectes larges ou de domaine, et les collectes ciblées ou sélectives. Les premières consistent en une collecte annuelle de tous les noms de domaine enregistrés en .fr ; cette liste est fournie annuellement grâce à une convention avec l'AFNIC. Dans le futur, la BnF espère être capable d'inclure les sites enregistrés en France sous d'autres TLD, comme les .org, .net ou .com, qui entrent dans les limites du dépôt légal et qui représenteraient environ deux tiers des sites enregistrés en France (voir *supra* C.). Cela impliquerait des accords supplémentaires directement avec les bureaux d'enregistrement[22]. Il n'y a donc pas de jugement concernant la qualité ou la valeur de ce qui est collecté ; dans la tradition du dépôt légal, tout ce qui tombe sous les critères déjà décrits est sujet à être collecté. Cette collecte annuelle utilise des paramètres techniques selon lesquels seule une quantité limité de données est collectée pour chaque domaine : en 2010, ce seuil était fixé à 10 000 URL (ou fichiers) par domaine. C'est suffisant pour collecter la majorité des sites dans leur intégralité, mais certains sites et plateformes importants ne sont collectés que partiellement. L'idée est de fournir une sorte d'instantané du Web français, qui, bien que limité à la fois en profondeur et en couverture temporelle, respecte l'obligation du dépôt légal de collecter le Web français au moins une fois par an (voir *supra* C.).

La seconde approche, celle des collectes ciblées, est complémentaire et implique la collecte de sites qui sont sélectionnés par des bibliothécaires chargés de collections à la BnF et, plus ponctuellement, par d'autres partenaires (comme les BDLI ou les chercheurs). Si elle participe encore du cadre législatif du dépôt légal, cette démarche peut ressembler à l'acquisition de livres ou d'autres ressources servant les missions d'une bibliothèque de recherche : des bibliothécaires choisissent des sites en se basant sur la valeur et l'intérêt du contenu, comme appartenant aux ressources conservées par la BnF dans un domaine donné ; les critères de sélection des sites doivent alors être liés à la politique globale d'acquisition des départements de collections de la Bibliothèque. Sont sélectionnés des sites qui ne peuvent être collectés dans la collecte large, ou de façon non satisfaisante : cela peut inclure des sites sous d'autres TLD que le .fr, et des sites ou parties de sites qui ne peuvent être collectés en raison de la taille du domaine (grands sites institutionnels, blogs individuels, etc.). Les collectes ciblées permettent aussi à des sites d'être collectés plus fréquemment qu'une fois par an. À partir de l'année 2011, la BnF a mis en place un système de collecte permanente, où les sites peuvent être collectés une fois par an, deux fois par an, une fois par mois, une fois par semaine ou même une fois par jour. Cela permet par exemple une collecte

---

[22]. Pour plus d'informations sur le rôle des registres de noms de domaine, voir AFNIC, Les autres registres de noms de domaine, http://www.afnic.fr/doc/autres-nic_fr (consulté le 27 juillet 2011).

quotidienne d'une sélection de sites d'actualité, pour montrer quelles sont les nouvelles qui font la une de la page d'accueil de chaque jour. Cela améliore également la qualité de la collecte de sites qui sont mis à jour fréquemment, ou de ceux qui ne gardent pas d'archives. La profondeur de la collecte et le nombre de fichiers collectés par site peuvent varier en fonction de la fréquence de moissonnage.

Enfin, la BnF a mis en place une procédure expérimentale par laquelle les producteurs peuvent proposer leur propre site en vue d'une collecte. Actuellement, ceux-ci peuvent envoyer un message sur une boîte dédiée aux propositions de sites, dont l'adresse est disponible sur les pages du site Web de la BnF dédiées au dépôt légal numérique. En fonction des résultats de cette expérience, cette approche pourra être développée une fois que le décret sera publié.

L'archivage du Web à la BnF repose principalement sur deux logiciels *open source* développés en partenariat avec d'autres institutions : le robot de collecte Heritrix[23], qui collecte les fichiers qui constituent les archives, et NetarchiveSuite[24], qui permet la planification, la programmation et la surveillance des collectes. Il est important de noter que de telles collectes, qui impliquent de faire des copies des fichiers qui forment le site Web, sont possibles uniquement parce que le code du patrimoine crée une exception dans la législation sur la propriété intellectuelle. L'article L132-4 spécifie que la reproduction du contenu sous droits est autorisée « lorsque cette reproduction est nécessaire à la collecte, à la conservation et à la consultation » dudit contenu[25]. Comme il est impossible de collecter le contenu en ligne sans en faire de reproduction, il était nécessaire d'introduire cette possibilité, et la loi DADVSI était spécifiquement destinée à résoudre de tels problèmes émergeant des incompatibilités entre les nouvelles technologies et la législation existante. D'autres aspects de cette question sont examinés dans la partie suivante (VI.B.)[26].

E. Obligations des producteurs et éditeurs

Cet article L132-4 indique aussi qu'il n'est pas nécessaire de demander de permission de collecte aux producteurs de sites Web et aux détenteurs de droits d'auteur. Des systèmes d'archivage du Web basés sur la demande de permission existent dans beaucoup d'autres pays (comme la Grande-Bretagne[27] et les États-Unis[28]), mais ils rendent impossible le moissonnage à grand échelle, car il est infaisable d'identifier et de contacter les propriétaires de millions de domaines. La fondation Internet Archive manque d'un support légal pour son travail d'archivage

---

[23]. Internet Archive. Heritrix. http://crawler.archive.org/ (consulté le 20 mai 2011)

[24]. Netarchive.dk. NetarchiveSuite. http://netarchive.dk/suite/Welcome (consulté le 20 mai 2011)

[25]. Code du patrimoine, article L132-4. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845526&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulté le 20 mai 2011)

[26]. Pour plus d'information sur l'archivage du Web à la BnF, voir Ariel BLEICHER, « A Memory of Webs past », dans *IEEE Spectrum*, mars 2011. http://spectrum.ieee.org/telecom/internet/a-memory-of-Webs-past/0 (consulté le 30 mai 2011)

[27]. Voir UK Web Archive, Legislative Status, http://www.Webarchive.org.uk/ukwa/info/about#legislative_status (consulté le 30 mai 2011) ; Department for Culture, Media and Sport. Legal Deposit. http://www.culture.gov.uk/what_we_do/libraries/3409.aspx (consulté le 30 mai 2011)

[28]. Bibliothèque du Congrès, Archivage du Web, http://www.loc.gov/Webarchiving/index.html (consulté le 30 mai 2011)

international commencé dès 1996, et elle adopte une politique de retrait sur demande (« *opt-out* »), selon laquelle l'accès au contenu est retiré de leurs archives en cas de plainte d'un producteur de site ou d'un détenteur de droits d'auteur[29]. En France, l'exception au droit de la propriété intellectuelle s'applique uniquement dans le strict cadre du dépôt légal, et en particulier en ce qui concerne les contrôles d'accès aux contenus archivés (voir VI).

En ce qui concerne le dépôt légal numérique, une autre clause introduite par le code du Patrimoine entraîne un changement dans la relation entre le producteur/éditeur et l'organisme dépositaire : contrairement au dépôt légal traditionnel, c'est la BnF qui collecte les sites Web, plutôt que de recevoir les dépôts des éditeurs. Cependant, la loi dit clairement que les producteurs ont comme responsabilité de faciliter la collecte de leur production si nécessaire :

> « [Les organismes dépositaires] peuvent procéder eux-mêmes à cette collecte selon des procédures automatiques ou en déterminer les modalités en accord avec ces personnes. La mise en oeuvre d'un code ou d'une restriction d'accès par ces personnes ne peut faire obstacle à la collecte par les organismes dépositaires précités[30]. »

Le décret doit préciser que les producteurs sont obligés de fournir tous les mots de passe et autres moyens d'accès aux documents ; cela peut s'appliquer non seulement aux parties de sites soumis à identification, mais aussi aux fichiers comme les contenus audiovisuels, les journaux payants ou les *e□books* qui peuvent être protégés par des DRM (*digital rights management*, système de gestion des droits). Les DRM peuvent limiter non seulement la collecte d'un site mais aussi sa préservation sur le long terme, et le décret devrait donc spécifier que les éditeurs doivent fournir toutes les informations et moyens d'accès nécessaires (voir IV.). Quant aux mots de passe, la définition du contenu numérique sujet au dépôt légal (citée en II.A.) spécifie qu'il doit « faire l'objet d'une communication au public », et cela exclut donc tout contenu en ligne qui pourrait être considéré comme de la correspondance privée. Ainsi ne sont pas concernés les contenus privés protégés par des mots de passe □ c'est le cas notamment des parties privées des réseaux sociaux. Cependant, les contenus publiés payants, et qui sont par conséquent accessibles par identification (mot de passe ou reconnaissance d'IP par exemple), seront probablement soumis au dépôt légal numérique comme ils le sont déjà au dépôt légal des imprimés. Les producteurs auront donc pour obligation de fournir toute l'aide nécessaire pour en garantir la collecte par la Bibliothèque.

F. Défis et questions ouvertes

Certaines institutions, comme la Bibliothèque royale des Pays-Bas ont d'abord commencé avec un e-dépôt basé sur des accords avec les éditeurs et n'a entrepris son programme d'archivage du Web que plusieurs années après. *A contrario*, la BnF a jusqu'ici concentré ses ressources pour le dépôt légal numérique dans l'archivage du Web, principalement pour des raisons économiques et pratiques : de grandes quantités de contenu publiquement disponible en ligne ont été collectées grâce au moissonnage automatique du Web, qui sinon auraient été perdues à jamais. Dès 2004 (lorsque la BnF a lancé sa première collecte large expérimentale, en

---

[29]. Internet Archive, Terms of Use. http://www.archive.org/about/terms.php (consulté le 20 mai 2011)

[30]. Code du patrimoine, article L132-2-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006845522&cidTexte=LEGITEXT000006074236&dateTexte=20110520 (consulté le 20 mai 2011)

partenariat avec Internet Archive), la Bibliothèque a senti l'urgence de collecter à l'échelle des changements extrêmement rapides du Web et de la disparition quotidienne de données à valeur patrimoniale. Cependant, certains contenus restent difficiles ou impossibles à collecter ainsi, soit parce que le contenu n'est pas en ligne, soit à cause de limitations d'accès techniques ou commerciales. La législation, on l'a vu, s'applique à tous les contenus publiés par voie électronique et oblige les producteurs à coopérer si nécessaire, ce qui ouvre la voie à d'autres approches, comme le dépôt de fichiers par les producteurs (voir E.).

Cette approche n'a pas encore été mise en œuvre à la BnF, notamment parce que les obligations des producteurs et celles de la BnF doivent être définies plus précisément dans le décret, et qu'il est préférable d'attendre cet appui légal. Cependant, les procédures à la fois techniques et organisationnelles qui devront être mises en place font l'objet d'un travail préparatoire. Le dépôt de fichiers a été expérimenté dans le domaine des journaux, qui posent des problèmes spécifiques en matière de stockage et de conservation. Entre 2005 et 2009, des expérimentations ont été conduites avec *Ouest France*, quotidien régional avec une distribution importante dans l'Ouest du pays. Il s'agissait d'organiser le e-dépôt des nombreuses éditions locales de ce journal (47 actuellement), afin que la Bibliothèque ne collecte et ne conserve qu'une seule édition de référence sous forme imprimée : étaient ainsi réalisés l'économie de beaucoup de manipulations à la BnF et à l'éditeur, et un gain d'espace de stockage dans les magasins. Cette expérience a été interrompue par manque de ressources propres et parce qu'il est apparu impossible, dans l'état actuel des choses, d'étendre ce mode de travail à d'autres titres de journaux. Il est probable que l'approche adoptée sera davantage axée sur la collecte des versions en ligne disponibles sur le Web, même si elle pose d'autres questions techniques et juridiques que nous verrons plus loin.

Certains de ces problèmes peuvent être abordés par le biais de l'archivage du Web, mais il y a beaucoup de questions à traiter. Il y a aujourd'hui beaucoup de formats de contenus difficiles à collecter : les médias enrichis (vidéos, *streaming*, Flash, javascript), le Web profond (les bases de données), les contenus accessibles par abonnement ou protégés par un mot de passe, etc. La nature des problèmes est généralement technique, et la BnF contribue et bénéficie des efforts internationaux visant à développer des outils et des compétences pour améliorer cette collecte. D'autres difficultés, en revanche, relèvent d'une combinaison de barrières techniques et de questions juridiques ou organisationnelles. Dans le cas de ressources protégées par un mot de passe ou nécessitant une identification d'IP, le robot Heritrix est capable de contourner les pages de connexion lorsqu'il a été programmé avec les détails d'accès ; il est bien sûr nécessaire de contacter les producteurs pour obtenir ces détails. Le code du patrimoine, soutenu par le décret à venir, oblige les producteurs à coopérer avec la BnF et à fournir toute l'information nécessaire pour les contenus soumis au dépôt légal, qu'ils soient gratuits ou payants. Comme il s'agit d'une obligation légale nouvelle, il reste à voir à quel point il sera simple d'encourager les producteurs et éditeurs à collaborer avec la BnF sur ce point ; cela demande également beaucoup de travail et de ressources à la Bibliothèque pour maintenir et suivre des contacts avec des éditeurs alors même que la Bibliothèque est confrontée à de coupes budgétaires. Dans certains cas, il sera aussi nécessaire de combiner ces contacts à d'autres développements techniques, par exemple dans le cas où le mot de passe donne accès à un lecteur Flash qui empêche Heritrix de collecter le contenu. Ce travail commencera sérieusement une fois le décret publié, et il impliquera beaucoup de changements organisationnels pour gérer la nouvelle manière de collecter les contenus. Cependant, la BnF pourra dès lors sans doute collecter plus de contenus, et

notamment des *e-books*, des journaux en ligne et d'autres ressources électroniques précieuses.

En revanche, pour certaines publications, il est probable que seul le dépôt de fichiers par les éditeurs permette à la BnF de respecter ses obligations de dépôt légal. Pour le moment, mises à part les expérimentations concernant la presse, le seul dépôt légal numérique systématique mené par les éditeurs concerne les affiches de grands formats, qui sont collectées sous forme numérique plutôt que sous format imprimé, grâce au décret de 1993 (modifié en 2006). Cependant, avec l'intérêt et la viabilité commerciale croissants des *e-books*, il reste une question ouverte sur la manière de gérer la collecte de ces publications. Au sein même de la BnF, cette question se pose aussi bien aux équipes responsables des dépôts légaux numérique et imprimé, qu'aux bibliothécaires des départements thématiques qui cherchent à acquérir des *e-books* pour leurs domaines d'action. Le flux opérationnel mis en place pour collecter ces publications aura également un impact sur d'autres domaines, tels le catalogage et la préservation de ces ouvrages. La BnF examine donc les options pour répondre au mieux à ce défi.

Ce qui pourrait être désigné comme des dons ou donations numériques entre pareillement dans ce champ de réflexion. En France, le dépôt légal a toujours été complété avec des acquisitions ou des dons, l'addition des différents modes de collecte contribuant à construire le patrimoine et les collections de recherche dans son ensemble. Il est intéressant de voir si des combinaisons analogues peuvent être envisagées en ce qui concerne les ressources « nées numériques » : on peut ainsi penser bien sûr aux acquisitions numériques (ressources électroniques payantes), mais également à des dons numériques ou bien à des « manuscrits » (auteurs, artistes). Comment ces collections interagissent-elles juridiquement et techniquement avec le dépôt légal numérique ? Encore une fois, ces questions nécessiteront d'être examinées dans le contexte d'une solution globale pour les publications électroniques, au-delà de ce qui existe déjà pour l'archivage du Web.

Enfin, la collaboration internationale fournit une piste pour une exploration approfondie de la collecte de contenus numériques. Plusieurs initiatives sont déjà en place, particulièrement dans le contexte du consortium international pour la préservation de l'Internet (en anglais IIPC[31]). Les nouvelles technologies peuvent être un outil d'amélioration de la qualité du moissonnage du Web ; des outils *open source* sont déjà développés par la communauté internationale, comme Heritrix et Netarchivesuite (chacun ayant été à l'origine développé par une seule institution, respectivement Internet Archive et NetArchive.dk) ou d'autres logiciels populaires, comme le Web Curator Tool (soutenu par la British Library et la Bibliothèque nationale de Nouvelle Zélande). Plus récemment, la discussion a porté sur le rôle que pourrait jouer la coopération internationale dans la sélection et la collecte des contenus, et sur l'extension des limites d'une division « nationale » de l'Internet. Dans le cas d'événements de portée internationale, chaque institution pourrait ainsi collecter les contenus Web de son propre pays. Des expérimentations autour de cette idée ont déjà eu lieu, comme des projets d'IIPC relatifs aux élections européennes de 2009 et aux jeux olympiques d'hiver de 2010, en vue de collecter les prochains jeux olympiques d'été de 2012 ; et aussi des collectes *ad hoc* répondant aux situations d'urgence : les tremblements de terre en Haïti en 2010 et au Japon en 2011, ou les événements politiques en Afrique du Nord, la « Révolution de jasmin », en 2011. Internet Archive effectue des collectes basées sur des

[31]. International Internet Preservation Consortium. http://www.netpreserve.org/about/index.php (consulté le 20 mai 2011)

propositions émanant de différents organismes[32], et certaines institutions (comme la BnF) effectuent en outre leurs propres collectes. À l'avenir devraient être mise en place des procédures pour élaborer des collections dans les différents pays « communiquant » entre eux, en prenant en compte les restrictions d'accès qui existent dans les législations de nombreux pays (voir V).

IV. Conservation des contenus obtenus par le dépôt légal numérique

A. L'obligation légale de préserver les collections patrimoniales

Le but du dépôt légal est de créer une trace permanente de la production culturelle en France. L'idée de conservation est donc à son cœur : les collections créées par le biais du dépôt légal doivent être préservées sans restriction de temps. Cette responsabilité, implicite dans le décret portant création de la BnF et soulignée par le code du patrimoine, tire sa force légale du code général de la propriété des personnes publiques[33]. L'article L2112-1 de ce code spécifie qu'un exemplaire de chaque document collecté par dépôt légal (selon la liste du code du patrimoine, article L131-2) doit être considéré comme appartenant au « domaine public mobilier », c'est-à-dire aux biens appartenant au domaine public, et donc « inaliénable et imprescriptible[34]. » Cette obligation fondamentale donne une importance particulière au rôle des techniques de préservation, domaine dans lequel, comme pour la collecte des contenus, la nature du dépôt légal numérique pose à la fois des défis techniques et juridiques différents de ceux rencontrés pour d'autres médias.

B. Approches techniques de la préservation numérique : copier pour conserver

D'un point de vue technique, des systèmes sont en train d'être mis en place à la BnF pour assurer la préservation à long terme des collections de dépôt légal numérique. Des formules similaires sont actuellement développées par un nombre croissant de bibliothèques nationales, comme la bibliothèque du Congrès, les bibliothèques nationales de Nouvelle-Zélande et d'Australie. La spécificité majeure de la préservation de contenus numériques réside dans le besoin d'être capable de faire des copies, soit identiques, soit modifiées pour prendre en compte notamment les changements de formats : la préservation numérique est en effet impossible sans la capacité à copier d'un support à l'autre. La loi DADVSI de 2006 a créé une exception aux lois de la propriété intellectuelle qui autorise la reproduction de contenu sous droits en cas de nécessité pour la collecte mais aussi pour la préservation du contenu[35].

---

[32]. Internet Archive Global Events. http://www.archive-it.org/public/partner.html?id=89 (consulté le 20 mai 2011)

[33]. Code général de la propriété des personnes publiques, article L2112-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006361198&cidTexte=LEGITEXT000006070299&dateTexte=20110520 (consulté le 20 mai 2011)

[34]. Id., article L3111-1. http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006361404&cidTexte=LEGITEXT000006070299&dateTexte=20110520 (consulté le 20 mai 2011)

[35]. Code du patrimoine, article L132-4, *op. cit.*

Actuellement, les contenus Web collectés (sous la forme de fichiers ARC ou WARC[36]) sont d'abord stockés sur les serveurs utilisés pour l'accès aux collections. La prévention de la détérioration des fichiers ou des supports passe par les réalisations d'une copie de sécurité, avec des vérifications et des sauvegardes, et par des empreintes numériques (« sommes de contrôle » ou *checksums*), des copies périodiques et le remplacement des disques durs ou des cassettes. Cette copie de sécurité dédiée à la conservation du « train d'octets » (c'est-à-dire la série de 0 et de 1) est une première étape de préservation, mais pour la préservation sur le long terme, la BnF va intégrer ses archives du Web dans le système centralisé de préservation numérique appelé SPAR (système de préservation et d'archivage réparti)[37]. Ce répertoire est conforme aux principes du modèle Open Archival Information System (OAIS)[38]. Outre la vérification de l'intégrité des données, SPAR permettra une analyse des formats utilisés, ce qui autorisera par la suite des stratégies de préservation sur le long terme pour combattre l'obsolescence des formats, stratégies basées sur la migration et l'émulation[39]. L'article L132-4 du code du patrimoine autorise ce type de manipulation des données nécessaire à la conservation des collections de dépôt légal. Le décret à venir devrait aussi réaffirmer l'obligation pour les producteurs de fournir tous les détails nécessaires pour la conservation des contenus de dépôt légal ; cela s'appliquera notamment aux technologies de gestion des droits numériques (DRM) qui peuvent limiter la reproduction ou la modification des fichiers telles qu'envisagées dans SPAR.

C. Demandes pour la destruction ou la modification de contenu émanant d'un particulier ou de l'éditeur

La BnF doit également considérer la possibilité que des particuliers puissent demander qu'une information contenue dans les archives du Web soit modifiée ou détruite. C'est ce que vise particulièrement la loi de 1978 relative à l'informatique, aux fichiers et aux libertés, qui autorise les particuliers à corriger ou effacer une donnée à caractère personnel les concernant publiée sur un site Web, ou détenue par une tierce partie (art. 40[40]). Des demandes peuvent aussi être faites par l'auteur ou l'éditeur du contenu cherchant à enlever ou modifier celui-ci. Cependant, l'obligation légale de préservation des collections de dépôt légal surpasse tout autre droit ou toute demande de destruction de contenu, et la même protection devrait être appliquée aux collections numériques comme aux livres et autres collections physiques, qui ne peuvent être détruites ou modifiées. Cette sauvegarde légale est au cœur du dépôt légal et elle est essentielle pour la préservation de l'intégrité des collections patrimoniales.

---

[36]. Internet Archive. Arc File Format. http://www.archive.org/Web/researcher/ArcFileFormat.php (consulté le 20 mai 2011)

[37]. BnF. Conserver : le projet SPAR et l'archivage numérique. http://www.bnf.fr/fr/professionnels/conserver_spar/s.conserver_SPAR_presentation.html (consulté le 16 août 2011)

[38]. ISO. Système ouvert d'archivage d'information : Modèle de référence. http://www.iso.org/iso/fr/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683 (consulté le 16 août 2011)

[39]. BERMES Emmanuelle, FAUDUET Louise, PEYRARD Sébastien, « Une approche orientée données pour la préservation du numérique : le projet SPAR », *WorldWorld Library and Information Congress* : 76th Ifla General Conference And Assembly (IFLA 76), 10-15 août 2010, Göteborg, Suède. http://www.ifla.org/files/hq/papers/ifla76/157-bermes-fr.pdf (consulté le 16 août 2011)

[40]. Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, *op. cit.*

C'est un point qui devrait être clarifié dans le futur décret. En France, l'autorité responsable des questions de vie privée et de protection des données est la Commission nationale de l'informatique et des libertés (CNIL). Celle-ci a récemment publié une recommandation concernant le décret qui reconnaît que le droit à la modification ou à la suppression des données ne doit pas s'appliquer aux collections de dépôt légal. Par conséquent, il est envisagé que de telles demandes soient traitées par la limitation ou la suppression de l'accès public au contenu en question, plutôt que par sa destruction ; les points que soulèvent cette possibilité seront développés plus loin (VI). Enfin, il devrait être noté qu'une décision de justice peut aboutir à un ordre de retirer des contenus conservés dans les archives du Web ; comme pour l'imprimé, la BnF serait obligée de respecter une telle décision, mais de tels cas sont extrêmement rares.

D. Défis et questions ouvertes

Un défi majeur pour l'avenir, déjà envisagé avec la création de SPAR, est la gestion des risques pour les collections d'archives du Web. Bien que des stratégies de préservation sur le long terme soient possibles, toute intervention dans les collections a des coûts associés en termes de temps passé à analyser et à traiter des collections, à la fois par des hommes et des machines, et des coûts d'augmentation d'espace de stockage. Par conséquent, il est possible que l'on ne puisse appliquer les stratégies de préservation de la plus haute qualité à tous les contenus des archives ; un système de gestion des risques peut alors être mis en place dans lequel il serait accordé à certains formats, qui pourraient par exemple être considérés comme présentant un risque particulier d'obsolescence, un traitement de qualité plus haute qu'à d'autres. Cette approche pourrait être comparée à celle utilisée pour les collections papier, où les ouvrages dont le support physique est considéré comme fragile sont traités différemment, et, par exemple, stockés dans des magasins spéciaux ou communiqués au public sous des conditions spécifiques. L'obligation légale prépondérante de préserver les collections de dépôt légal numérique ne change pas, mais étant donnée la masse des données collectées et à l'instar du mouvement de l'exhaustivité vers l'échantillonnage, il faut adopter une approche pragmatique dans le but de gérer d'une manière durable des ressources de préservation limitées.

V. Signalement des collections et bibliographie nationale : quelles obligations pour le catalogage des contenus obtenus par dépôt légal numérique ?

A. Les exigences du signalement

La nature des contenus électroniques, et particulièrement du Web, demandent également une nouvelle approche de l'obligation de créer une bibliographie nationale de la production française, inscrite dans le code du patrimoine comme l'un des buts du dépôt légal[41]. Cette obligation est remplie par la publication de la Bibliographie nationale française par la BnF[42], et de catalogues mis à disposition par l'INA et le CNC. Si certaines publications électroniques respectant la forme de publications traditionnelles, comme les *e-books* et les journaux en ligne, peuvent en

[41]. Code du patrimoine, article L131-1 b), *op. cit.*

[42]. BnF, *Bibliographie nationale française en ligne.* http://bibliographienationale.bnf.fr/ (consulté le 17 août 2011)

théorie continuer à être traitées de la même manière, la nature de l'information publiée sur le Web implique que les archives du Web reçoivent un traitement différent.

Comme nous l'avons vu, le Web diffère non seulement pour l'étendue des informations produites, mais aussi pour la nature et la granularité de celles-ci : il peut être vu moins comme un ensemble de publications individuelles et séparées que comme un flux constant de données liées. Par conséquent, il est impossible d'adopter une approche bibliographique traditionnelle sur une large échelle, à cause de la difficulté à identifier des « items » à cataloguer (sites Web, domaines, hôtes, répertoires de sites Web, pages, fichiers…) et de la masse pure et simple des contenus (le moissonnage 2010 du domaine français a couvert 1,6 million de domaines, et collecté environ 800 millions de fichiers).

En prenant acte de cette différence, le décret à venir devrait autoriser l'indexation automatique à prendre la place de la bibliographie traditionnelle comme moyen de garantir l'accès aux collections. Cela s'inscrit dans la lignée des mesures déjà mises en place par la BnF, qui a choisi de ne pas cataloguer manuellement les contenus collectés par le moissonnage du Web, mais de procéder à la place à une indexation. Cette approche est basée sur l'idée que les archives du Web doivent suivre la même logique que le Web vivant, où les sites sont interconnectés plutôt qu'envisagés comme des unités stables, séparées. D'autres institutions ont adopté des approches différentes, en particulier lorsque leur programme d'archivage du Web est restreint à des ensembles de données plutôt sélectifs par la loi ou par orientation ; la British Library par exemple, ou la Bibliothèque du Congrès cataloguent toutes deux des sites individuels. C'est en partie possible grâce au fait qu'elles n'effectuent pas actuellement de moissonnage à grande échelle ; cette approche ne pourra sans doute pas être appliquée à un moissonnage de domaines (voir III. E. *supra*, notes 24 et 25).

B. Les moyens techniques actuellement en place

Une interface expérimentale permettant l'accès aux collections des archives du Web de la BnF est disponible depuis 2008. Elle est basée sur la Wayback Machine, logiciel *open source* initialement développé par Internet Archive[43]. Cependant, il n'y a pour le moment pas d'indexation complète plein texte des archives du Web, mais une indexation plein texte expérimentale a été appliquée à environ 10 % de la collection ; en effet, en raison de l'accent mis sur la collecte de contenus, il n'a pas été possible de consacrer du temps de développement à la réalisation d'une indexation plein texte. Il s'agira d'un projet majeur dans un avenir proche, parce que la mise en place d'un moteur de recherche fonctionnel, avec l'ajout de la dimension temporelle, est essentielle pour exploiter pleinement la nature des archives du Web. Plusieurs institutions, comme l'INA, la British Library ou la Bibliothèque nationale et universitaire d'Islande, ont été plus heureuses dans la mise en œuvre de moteurs de recherche plein texte pour leurs archives du Web. Si ces institutions ont encore à faire avec le classement et la pertinence des résultats, elles ont démontré qu'un tel projet est parfaitement faisable même s'il requiert une quantité considérable de ressources informatiques.

À la BnF, le principal moyen d'accès actuel est donc une recherche par URL, semblable en principe aux fonctions de recherches disponibles sur le site Web d'Internet Archive (www.archive.org). Ce niveau d'indexation permet l'accessibilité

[43]. Internet Archive. Wayback. http://archive-access.sourceforge.net/projects/wayback/ (consulté le 20 mai 2011)

de la collection entière des archives du Web, comme l'exige la législation. Cependant, ce mode d'accès est clairement limité : il est nécessaire de connaître l'URL du site que l'on est en train de chercher, et il n'y a pas de continuité dans les archives pour un site qui change d'URL avec le temps. Par exemple, l'UMP a changé l'adresse de son site Web, de http://www.u-m-p.org/ à http://www.lemouvementpopulaire.fr/, et le passage d'une version à l'autre n'est pas explicite lors d'une consultation des archives du Web. Il est également impossible d'effectuer des recherches thématiques, sans compter des analyses plus sophistiquées d'exploration de données (*data mining*) (voire partie VI. E. *infra*). Cela signifie aussi, au moins pour le moment, que les archives du Web ne sont pas accessibles par le biais du catalogue de la BnF, mais seulement en utilisant leur propre interface dédiée, isolée des autres applications fédérées de la Bibliothèque. D'autres institutions, comme la British Library ou la Bibliothèque nationale de Singapour, ont fait un travail très intéressant en intégrant mieux leurs archives du Web au reste de leurs collections numériques, et il s'agit évidemment d'un autre domaine clé pour l'avenir, dans le même but de démontrer et d'améliorer la valeur des archives du Web pour l'utilisateur final.

C. Défis et questions ouvertes

Ces limitations peuvent être vues comme restreignant l'accès aux archives. Même si l'obligation de produire une bibliographie peut être remplacée, selon le cadre légal, par l'indexation (incluant l'indexation des URL), il est clair qu'il sera nécessaire de mettre en œuvre de nouvelles solutions technologiques correspondant aux attentes des chercheurs et aux pratiques émergentes d'exploration de données et de liens sur le Web vivant. Cela permettra de répondre à l'esprit de cette obligation de signalement et d'aider l'analyse par les chercheurs de la production culturelle en ligne en France.

Encore une fois intervient la question de la coopération internationale. La mise en place de systèmes permettant aux utilisateurs de chercher dans les collections de nombreux pays pourrait être possible, là où des collections ont été créées dans le contexte d'une coopération internationale même là où un accès en ligne est impossible pour des raisons légales, comme en France , afin au moins de voir quel contenu est conservé par quelles institutions. C'est le sujet d'une discussion au sein du consortium IIPC ; du point de vue de la BnF, les développements ouvrant les archives du Web à la recherche depuis le catalogue pourraient être une avancée technique dans cette direction.

Enfin, la nature des ressources électroniques, et particulièrement des archives du Web, ouvre des possibilités de création d'autres moyens d'accès, au-delà de la « description » de ressources telle que traditionnellement pratiquée. L'indexation plein texte sera un premier pas en ce sens, mais d'autres outils pourraient être mis en place, permettant un traitement bien plus détaillé des contenus électroniques par des moyens comme l'exploration de données. En Europe, le projet LAWA (Longitudinal Analytics of Web Archive Data[44]), soutenu par la Commission européenne, explore des cas pratiques intéressants et des prototypes pour avancer dans ce domaine. La mise en œuvre de tels outils, qui requiert la manipulation de grands réservoirs de données dans des environnements logiciels particuliers, est déjà largement liée à la situation légale relative à l'accès à ces ressources (voir VI.).

---

[44]. Site Web de LAWA : http://www.lawa-project.eu/ (consulté le 24 août 2011). Présentation du projet sur le site de la fondation Internet Memory : http://internetmemory.org/fr/index.php/projects/lawa1 (consulté le 17 juillet 2011)

VI. Conditions d'accès aux contenus obtenus par dépôt légal numérique

A. Donner accès aux collections *vs* les protéger

De manière générale, il y a trois modèles d'accès aux archives du Web : une archive noire, où les contenus sont collectés mais ne sont pas rendus accessibles (ou du moins pas avant l'achèvement d'un délai de restriction de communication) ; un archive blanche ou « ouverte », qui est entièrement ouverte au public (via l'Internet) ; et une archive « grise », qui offre un accès contrôlé sous certaines conditions. Certaines institutions, comme la Bibliothèque nationale de Norvège, sont actuellement obligées d'avoir des archives noires, généralement à cause du droit d'auteur et surtout de la loi de protection des données en application dans leur pays ; dans ces pays, les conditions d'accès aux contenus Web archivés restent alors à définir. D'autres pays, comme le Royaume-Uni[45], ont mis en place des archives ouvertes, souvent grâce à un système de collecte basé sur la permission, où la demande de rendre public le contenu est incluse dans celle de permission de la collecte. Internet Archive[46] maintient également une archive ouverte, mais avec une politique de retrait en cas de plainte. La Bibliothèque nationale et universitaire d'Islande est à notre connaissance la seule institution patrimoniale publique qui a choisit une approche similaire.

La législation française permet pour le contenu collecté par dépôt légal numérique une mise à disposition en archive grise, avec de strictes restrictions d'accès. Cela crée une tension entre le besoin évident de fournir un accès aux archives, dont la constitution n'a de sens que si elles ont des lecteurs, et les restrictions légales.

Une des missions de la BnF, énoncée dans le décret de 1994, est :

> « D'assurer l'accès du plus grand nombre aux collections, sous réserve des secrets protégés par la loi, dans des conditions conformes à la législation sur la propriété intellectuelle et compatibles avec la conservation de ces collections[47]. »

Cela montre bien la tension entre d'un côté les besoins de l'accès et de la conservation, et de l'autre celui de respecter les droits de la propriété intellectuelle et de la protection des données personnelles. Cependant, au contraire du papier et des autres supports physiques, l'utilisation des ressources électroniques ne risque pas de les endommager, puisque la loi autorise la reproduction spécifiquement à titre de conservation (voir IV.B.), et que le système mis en place à la BnF séparera à terme les copies utilisées pour l'accès et celles utilisées pour la préservation. Par conséquent, c'est l'aspect des droits de la propriété intellectuelle en particulier qui entre en jeu. Le code du patrimoine spécifie que les contenus numériques collectés par dépôt légal peuvent être consultés « sur place par des chercheurs dûment accrédités (…) sur des postes individuels de consultation dont l'usage est exclusivement réservé à ces chercheurs[48]. » Le décret doit maintenir cette limitation, même s'il pourrait ouvrir l'accès à un nombre limité d'autres bibliothèques partageant avec la BnF la responsabilité de dépôt légal. Dans ce cas, l'accès depuis

---

[45]. UK Web Archive. http://www.Webarchive.org.uk/ukwa/ (consulté le 30 mai 2011)

[46]. Internet Archive. http://www.archive.org/Web/Web.php (consulté le 30 mai 2011)

[47]. Décret n° 94-3 du 3 janvier 1994 portant création de la Bibliothèque nationale de France, article 2, *op. cit.*

[48]. Code du patrimoine, article L132-4, *op. cit.*

ces bibliothèques régionales impliquerait toujours de maintenir l'exigence de postes individuels et d'accréditation pour les chercheurs. Ces possibilités sont discutées plus loin (E.).

L'accès aux archives du Web est disponible dans les salles de recherche des différents sites de la BnF. Ces salles de lecture sont réservées à des chercheurs qui ont un besoin démontré d'utiliser ces collections. Cette limitation existe pour protéger les collections de dépôt légal physique, principalement pour les besoins de la préservation. Dans le cas des archives du Web, d'autres raisons justifient cette forme d'accès.

B. Le droit d'auteur et la propriété intellectuelle

Il est important de noter encore une fois que la loi de 2006 qui a créé le dépôt légal numérique visait à traiter spécifiquement des questions de droit d'auteur et de propriété intellectuelle, dans le paysage changeant de la « société de l'information ». Cela explique la situation, qui peut apparaître paradoxale, selon laquelle la BnF collecte actuellement du contenu qui est librement accessible sur le Web, et n'en fournit ensuite l'accès que sous de strictes conditions. À un premier niveau, une sorte de « compétition » est ainsi évitée entre les versions archivées d'un site et le Web vivant : si les archives étaient disponibles directement sur l'Internet, et indexées par des moteurs de recherche, une version périmée d'un site Web pourrait théoriquement être placée plus haut que la version courante dans les résultats de recherche. Même sans cela, les auteurs et les éditeurs de sites Web préféreront que les internautes utilisent le site « vivant », car cela leur permet de tracer les usages, de générer des revenus publicitaires, etc. De plus, si la BnF commence à collecter des sites payants, cet accès contrôlé deviendra encore plus important, puisque la Bibliothèque ne peut bien sûr fournir en ligne et gratuitement des ressources pour lesquelles les éditeurs demandent une souscription.

C. Respect de la vie privée, protection des données et contenu sensible

Une autre raison pratique aux limitations d'accès vient des questions de respect de la vie privée et de protection des données (voir IV). La loi de 1978 sur l'informatique, les technologies et les libertés permet à des particuliers de corriger ou d'effacer des informations les concernant publiées sur un site Web, et son application est contrôlée par la CNIL. L'existence d'un dépôt légal numérique signifie que d'anciennes versions dudit site Web, contenant les informations erronées ou non désirées, peuvent encore être accessibles dans les archives de la BnF. La CNIL a récemment reconnu que l'on ne pouvait pas demander à la BnF de modifier ou de détruire un contenu tombant sous la loi de 1978 (voir IV.C) ; dans de tels cas, la Bibliothèque peut cependant décider de mettre en place des restrictions additionnelles d'accès.

Dans le même ordre d'idée, on peut également penser aux contenus déclarés diffamatoires par un juge, ou aux contenus illégaux. Par ailleurs, une partie du contenu des archives peut être légale mais potentiellement choquante, la pornographie par exemple, et peut donc nécessiter un contrôle spécial de l'accès (dans ce dernier cas, il faut noter que l'accès au niveau recherche, où les archives sont disponibles, est déjà limité aux personnes de plus de 18 ans).

D. Les implications pratiques pour la BnF

Dans ces cas, les limitations d'accès imposées par la loi fournissent une première réponse, puisque le contenu potentiellement sensible est protégé *de facto* de l'accès au grand public, et ne peut être consulté que dans le contexte de la recherche ou d'autres utilisations justifiées. La CNIL, dans ses recommandations pour le décret proposé, a jugé suffisantes les mesures d'accréditation des chercheurs en place à la BnF. Cependant, comme cela pourrait ne pas suffire pour faire face à toutes les situations sensibles qui apparaîtraient, d'autres solutions peuvent être imaginées.

Le système envisagé par la BnF est celui d'un « accès restreint » à certains contenus : ceux-ci pourraient être tenus à l'écart des archives accessibles au niveau recherche et placés dans une « archive noire » inaccessible, ou bien ils pourraient nécessiter une demande spéciale requérant une justification additionnelle. Un système semblable existe déjà pour les imprimés et autres supports, et les mêmes principes pourraient probablement être appliqués aux archives du Web. Ainsi, la BnF est obligée de restreindre l'accès où une décision de justice l'a spécifiquement demandé ; elle est également obligée par le code du patrimoine de restreindre l'accès aux contenus définis comme « secrets protégés par la loi[49] », ce qui inclut par exemple les secrets militaires. De plus, dans de très rares cas, les contenus sensibles ou illégaux (comme la pornographie ou les publications incitant à la haine raciale) ne sont pas accessibles sur simple demande mais seulement par le biais d'une procédure spécifique. Il en est de même pour le cas où un livre a été jugé diffamatoire par la justice et a donc été retiré de la circulation ; l'exemplaire entré à la BnF par dépôt légal ne peut être détruit, de par le statut des collections de dépôt légal (voir IV.), mais dans ce cas, le système d'accès contrôlé protège la personne qui se trouve diffamée d'une diffusion du contenu. Dans tous ces cas, l'application d'un tel accès restreint est jugée au cas par cas afin de prendre en compte les droits des particuliers et les demandes concernant leur vie privée tout en respectant la mission patrimoniale de la Bibliothèque.

La Wayback Machine permet de créer un tel système, déjà fonctionnel dans d'autres institutions comme Bibliothèque et Archives Canada, mais qui reste à être mis en œuvre à la BnF. En outre, il faut aussi fixer les bases sur lesquelles une telle retenue peut être mise en place : à part les cas d'obligation légale décrits ci-dessus, chaque demande devra pouvoir être jugée sur ses mérites. De même, les questions de savoir si les restrictions seront mises en place de façon permanente ou pour une période définie, et avant quel délai le contenu pourra être rendu à un accès normal, devront être examinées au cas par cas, avec peut-être une réévaluation régulière pour juger si l'accès restreint est encore justifié.

Toujours dans le domaine de la consultation, la reproduction du contenu des archives pose question. Elle aussi est contrôlée par la loi sur la propriété intellectuelle, à l'instar de celle des autres collections de dépôt légal, et toute reproduction est donc strictement limitée la facilité de faire des copies identiques des ressources électroniques rend les problèmes potentiels d'autant plus importants. Actuellement, les copies sont limitées à des impressions d'écran ; les copies d'écran ne sont pas permises, encore moins la copie de fichiers (images, PDF ou autres documents, code HTML…). Le code du patrimoine ne permet la reproduction que dans la limite du nécessaire pour la collecte, la conservation et la consultation du contenu ; par conséquent, il autorise la reproduction de fichiers entre les serveurs de stockage et ceux utilisés pour la consultation, mais il

[49]. Code du Patrimoine, article L131-1, *op. cit*.

n'autorise pas de copies permanentes (tout comme pour la collecte, l'idée de « copie » devient problématique puisque que toute utilisation de ressource électronique implique la copie de fichiers, même si c'est temporairement). Le décret à venir devrait aussi donner plus de précision dans ce domaine, même s'il est probable qu'il sera ajouté que l'accès ne pourra être fourni qu'en utilisant « les interfaces d'accès, de recherche et de traitement mises à disposition par la BnF, l'INA et les bibliothèques et organismes habilités ». La restriction sur le traitement des données dans les archives est importante, et introduit l'un des défis concernant les utilisations possibles par les chercheurs.

E. Défis et questions ouvertes

Il y a actuellement une double restriction de l'accès aux archives qui limite leur utilisation : la restriction légale demandant un accès contrôlé, et une restriction technique découlant de l'absence de recherche plein texte et d'autres outils d'accès et de traitement des archives (voir V.B. et VI.A.). Actuellement, l'accès est limité aux salles de lecture du niveau recherche sur les différents sites de la BnF. Pour améliorer le service offert aux usagers, le décret à venir devrait autoriser un accès contrôlé dans un nombre limité de bibliothèques régionales (les BDLI), déjà impliquées dans la sélection des sites (voir III.B.). Il existe une BDLI dans chaque région française, et la publication du décret signalant que le dépôt légal numérique appartient désormais fermement au paysage législatif français, permettra aux moyens d'accès actuels de dépasser le stade expérimental ; il devrait par conséquent permettre aussi à la BnF de promouvoir plus largement ce nouveau service. Pour profiter pleinement d'une telle ouverture, de nouveaux outils et de nouvelles manières d'utiliser les archives seront nécessaires ; cependant, la création d'outils supplémentaires ne peut être imaginée que dans le contexte des limites imposées par la législation. Des cas d'usage possibles se sont déjà posés qui nous permettent d'imaginer au moins trois domaines pour de possibles développements : une utilisation plus approfondie par des chercheurs, des demandes de copies de sites par les producteurs eux-mêmes, et des demandes de copies pour une utilisation judiciaire.

Pour les chercheurs, l'utilisation limitée aux copies d'écran restreint les manières dont ils pourraient utiliser les archives dans leur travail. Il pourrait être possible d'imaginer des exceptions pour un usage académique de petites parties d'un site archivé, même s'il est difficile de voir comment cela pourrait s'accommoder avec la législation, qui semble exclure toute reproduction ou utilisation d'un contenu du dépôt légal numérique hors de la BnF ou des autres institutions responsables.

Une question plus sérieuse est posée par l'analyse par les chercheurs des archives elles-mêmes. Comme nous l'avons vu, le décret précise que tout accès aux archives de la BnF doit se faire par le biais des interfaces proposées par la bibliothèque. Les chercheurs travaillant avec le Web, et donc avec les archives du Web, ont besoin d'outils qui ne leur permettent pas seulement de chercher dans les archives (comme le permettra la recherche plein texte), mais aussi de techniques d'exploration de données permettant une utilisation plus créative des collections. Cela pourrait inclure des outils pour tracer l'utilisation de certains termes, de tendances ou de noms au cours du temps et d'un site Web à l'autre, ou l'analyse de l'évolution des liens entre les sites. Différents centres de recherche sont déjà en train de travailler avec de tels outils, mais les restrictions légales les empêchent de faire des reproductions des données conservées par la BnF à des fins de traitement, ou d'installer eux-mêmes des logiciels à la Bibliothèque pour analyser les archives. La BnF doit donc envisager des manières de rendre de tels outils accessibles aux

chercheurs, et en particulier de permettre l'installation de logiciels autorisés. Indubitablement, des partenariats avec des universités et des centres de recherche seront incontournables dans ce domaine, et la BnF vient de lancer un projet collaboratif avec le Médialab de Sciences Po pour explorer ces questions.

Un autre cas a déjà été rencontré, et va sans aucun doute devenir de plus en plus courant à mesure que les archives du Web deviendront plus connues et que davantage de contenus disparaîtront du Web vivant : un producteur ou un auteur de site Web cherche à retrouver du contenu qui n'est plus en ligne et dont il n'a pas gardé de copie. Il peut s'agir d'un producteur de site Web dont le site était hébergé par une tierce partie qui n'existe plus, d'un particulier qui souhaite garder une copie d'un blog entretenu par le passé et qui a disparu de la plateforme en ligne, ou bien d'un journaliste qui a produit du contenu pour une source en ligne qui n'est plus disponible. À strictement parler, la législation empêche toute reproduction de contenu collecté par dépôt légal numérique, excepté à fins de collecte, de préservation et de consultation en ligne[50]. Cependant, il s'agit à l'origine de protéger les détenteurs des droits de propriété intellectuelle, et dans le cas où il s'agit de l'ayant droit qui demande le contenu, il devrait être possible de faire une exception. Les termes exacts selon lesquels une demande peut être faire sont encore à définir, comme de savoir qui est autorisé à faire une demande (les descendants de quelqu'un qui a écrit un blog cent ans plus tôt ?). Il sera en particulier nécessaire dans chaque cas de prouver les droits de la personne sur le contenu demandé, puisqu'il ne serait possible de fournir des reproductions que dans les cas où la demande viendrait du détenteur des droits sur tout le contenu en question ; par exemple, les droits sur la musique hébergée sur un blog peuvent appartenir à quelqu'un d'autre. Il est important également de noter que les procédures techniques d'export de tous les fichiers relatifs à un site Web ou à une partie de site Web, capturés à un moment donné, ne sont pas encore en place. Ce point devrait s'améliorer grâce à la mise en place de SPAR, mais il sera nécessaire de mettre en œuvre d'autres mesures techniques. À titre de comparaison, la reproduction d'imprimés et autres contenus est possible contre rémunération, mais uniquement pour les contenus hors droits.

Enfin, plusieurs demandes de copies d'archives du Web en lien avec des cas judiciaires ont déjà été soumises à la BnF, afin de prouver la présence de contenu en ligne à une date donnée ; cela peut être pertinent dans des affaires de propriété intellectuelle, ou de litiges concernant des conditions de vente en ligne. La BnF, avec son statut d'institution nationale, devrait pouvoir jouer un rôle de tierce partie de confiance pour de tels contenus, et les moyens de collecte en place, qui associent des métadonnées à chaque fichier collecté, rendent possible la preuve de la présence en ligne d'un fichier au moment exact où il a été collecté. Cependant, cela présente divers problèmes qui demandent un examen approfondi ; il faut en particulier définir les conditions exactes justifiant une exception à l'interdiction de reproduction des contenus entrés par dépôt légal. Si une telle exception était jugée possible, la BnF aurait également besoin de mettre en place les moyens techniques nécessaires pour produire une copie « authentifiée » des fichiers en question, avec la date de la collecte et la provenance clairement précisées. Cette utilisation des archives du Web nécessite encore un examen convenable devant la justice ; cependant, s'il y a jurisprudence, on peut imaginer que ce type de requête deviendra de plus en plus commun dans le futur. Cette dernière illustration fournit une démonstration supplémentaire de l'intérêt de maintenir un idéal de dépôt légal étendu du Web français, à l'exemple de la collecte large de la BnF ; il est impossible

[50]. Code du Patrimoine, article L132-4, *op. cit*.

de prédire quels sites, dont des sites commerciaux apparemment inintéressants, pourraient devenir importants ou « utiles » dans les années à venir. Il est donc nécessaire de collecter aussi largement que possible, puisque la valeur des collections de dépôt légal numérique ne se révélera que dans le futur.

VII. Conclusion

Le dépôt légal numérique en France n'est âgé que de quelques années, mais s'est déjà établi comme une des missions essentielles de la BnF. Dès les premières expérimentations en 2002, la BnF a mis en place un système d'archivage du Web impliquant des solutions techniques, à la fois matérielles et logicielles, mais aussi des éléments organisationnels, puisque cette mission demande l'expertise de conservateurs chargés de collections numériques, de spécialistes des technologies de l'information, de bibliothécaires spécialisés et d'experts juridiques, tout aussi bien qu'un solide soutien de la direction. Aujourd'hui, la BnF est dotée d'un flux opérationnel robuste, flexible et efficace pour l'archivage du Web, qui assure que le contenu publié sur le Web français trouve sa place dans les collections patrimoniales générées par le dépôt légal.

Comme les différentes parties de cet article l'ont démontré, des questions et des défis pour le futur subsistent à chaque maillon de la chaîne du dépôt légal numérique (la collecte, la conservation, la description et l'accès). Se combinent à la fois des questions juridiques et le besoin de solutions techniques et organisationnelles pour répondre pleinement aux obligations du dépôt légal numérique. À court terme, la publication du décret sera un pas important : il établira clairement les bases légales du dépôt légal numérique, et autorisera la BnF à poursuivre d'importants projets, comme l'amélioration de la collecte des publications payantes et l'étude de l'ouverture des possibilités d'accès aux chercheurs dans les bibliothèques régionales.

Ces deux projets, ainsi que la collecte des médias enrichis, celle des *e-books* ou d'autres contenus qu'on ne peut collecter par moissonnage, l'indexation plein texte, l'accès depuis le catalogue, la recherche fédérée, les stratégies de préservation différenciées, la fouille de données et autres outils destinés à l'exploitation des archives tout cela demandera la mise en place d'un travail et de ressources techniques significatifs et la BnF devra établir ses priorités pour les années à venir.

La vision que les utilisateurs, notamment les chercheurs, auront de ces collections, représentera un facteur important de maturation du dépôt légal numérique. Plusieurs usages possibles peuvent être imaginés et ont déjà commencé à apparaître, mais, pour nombre de ces cas, il reste encore à établir fermement quelle situation légale doit s'appliquer. Cependant, les différents rôles possibles des archives du Web détermineront dans une large mesure les priorités : des ressources devront être dédiées aux domaines où il y a un clair besoin de nouveaux outils, usages ou services.

En tant que membre fondateur du consortium IIPC depuis 2003 et membre de long terme de son comité de pilotage, la BnF est depuis longtemps active au sein de la communauté internationale, sur ces questions et bien d'autres. La coopération internationale, à la fois sur le plan technique et sur le plan organisationnel, jouera encore à l'avenir un rôle central pour le développement du dépôt légal numérique. La constitution d'une collection internationale fédérée, les discussions d'experts, la constitution « d'états de l'art », la normalisation, tout ceci obtenu grâce à des échanges de compétences, d'équipes et de meilleures pratiques ont représenté

d'inestimables atouts et ont créé des opportunités uniques pour mettre en œuvre le programme d'archivage du Web de la BnF pendant les dix dernières années. Archiver le Web étant par nature une tâche mondiale, la coopération internationale restera l'une des clefs de la découverte de solutions aux nombreux défis auxquels nous allons encore être confrontés.

VIII. Bibliographie

A. Textes législatifs gouvernant le dépôt légal numérique en France

Code du patrimoine : Titre 3, dépôt légal. http://www.legifrance.gouv.fr/affichCode.do?idArticle=LEGIARTI000006845515&idSectionTA=LEGISCTA000006159934&cidTexte=LEGITEXT000006074236&dateTexte=20110517 (consulté le 17 mai 2011)

Loi n° 2006-961 du 1 août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information. http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006054152&dateTexte=20110520 (consulté le 20 mai 2011)

Code de la propriété intellectuelle. http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006069414&dateTexte=20110520 (consulté le 20 mai 2011)

Code général de la propriété des personnes publiques. http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006070299&dateTexte=20110520 (consulté le 20 mai 2011)

Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068624&dateTexte=20110520 (consulté le 20 mai 2011)

Décret n° 94-3 du 3 janvier 1994 portant création de la Bibliothèque nationale de France. http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006082797&dateTexte=20110520 (consulté le 20 mai 2011)

Décret n° 93-1429 du 31 décembre 1993 relatif au dépôt légal. http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006082758&dateTexte=20110520 (consulté le 20 mai 2011)

B. Sélection de textes juridiques d'autres pays

1. United States

Copyright Law of the United States of America and Related Laws Contained in Title 17 of the United States Code; Chapter 4: Copyright Notice, Deposit, and Registration. http://www.copyright.gov/title17/92chap4.html (consulté le 20 mai 2011)

2. United Kingdom

Legal Deposit Libraries Act 2003. http://www.legislation.gov.uk/ukpga/2003/28/contents (consulté le 30 mai 2011)

C. Institutions responsables du dépôt légal en France

1. BnF

http://www.bnf.fr/fr/professionnels/depot_legal/a.dl_sites_Web_mod.html (consulté le 19 août 2011)

2. INA

http://www.institut-national-audiovisuel.fr/nous-connaitre/entreprise/statut.html (consulté le 26 juillet 2011)

D. Autres institutions non dépositaires en France

Fondation Internet Memory. http://internetmemory.org/fr (consulté le 19 août 2011)

Sciences Po, Médialab. http://www.medialab.sciences-po.fr/index.php?page=accueil (consulté le 19 août 2011)

E. Autres sources

AFNIC, Observatoire 2010 du marché des noms de domaine en France, p. 20-22. http://www.afnic.fr/data/actu/public/2010/afnic-observatoire-domaines-france-2010.pdf (consulté le 27 juillet 2011)

BERMES Emmanuelle, FAUDUET Louise, PEYRARD Sébastien, « Une approche orientée données pour la préservation du numérique : le projet SPAR », *WorldWorld Library and Information Congress* : 76th Ifla General Conference And Assembly (IFLA 76), 10-15 août 2010, Göteborg, Suède. http://www.ifla.org/files/hq/papers/ifla76/157-bermes-fr.pdf (consulté le 16 août 2011)

BLEICHER Ariel, « A Memory of Webs past », dans *IEEE Spectrum*, mars 2011. http://spectrum.ieee.org/telecom/internet/a-memory-of-Webs-past/0 (consulté le 30 mai 2011)

BnF. Conserver : le projet SPAR et l'archivage numérique. http://www.bnf.fr/fr/professionnels/conserver_spar/s.conserver_SPAR_presentation.html (consulté le 19 août 2011)

Department for Culture, Media and Sport, « Legal Deposit ». http://www.culture.gov.uk/what_we_do/libraries/3409.aspx (consulté le 30 mai 2011)

HUCHET Bernard, ILLIEN Gildas, OURY Clément, « Le Temps des moissons. Le dépôt légal du Web : vers la construction d'un patrimoine coopératif », *Revue de l'Association des Bibliothécaires de France*, 2010, n° 52, p. 28-31.

International Internet Preservation Consortium. http://www.netpreserve.org/about/index.php (consulté le 20 mai 2011)

Internet Archive. http://www.archive.org/ (consulté le 20 mai 2011)

Internet Archive. Arc File Format. http://www.archive.org/Web/researcher/ArcFileFormat.php (consulté le 20 mai 2011)

Internet Archive. Heritrix. http://crawler.archive.org/ (consulté le 20 mai 2011)

ISO. Système ouvert d'archivage d'information : Modèle de référence (ISO 14721:2003).

http://www.iso.org/iso/fr/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683 (consulté le 16 août 2011)

Library of Congress. Web Archiving. http://www.loc.gov/Webarchiving/index.html (consulté le 30 mai 2011)

Netarchive.dk. NetarchiveSuite. http://netarchive.dk/suite/Welcome (consulté le 20 mai 2011)

UK Web Archive. http://www.Webarchive.org.uk/ukwa/ (consulté le 30 mai 2011)

UK Web Archive, « Legislative Status ». http://www.Webarchive.org.uk/ukwa/info/about#legislative_status (consulté le 30 mai 2011)