Le dépôt légal d'Internet à la Bibliothèque nationale de France

Cadre juridique, modèle de collecte, évolutions des métiers

Avec 25 millions d'internautes, 12 millions de foyers équipés d'un micro-ordinateur et un nombre croissant de sites publics (plus de 445 000 pour le « .fr » fin 2005), Internet a désormais atteint toutes les sphères de la société française : e-administration, arts numériques, édition en ligne, enseignement à distance, commerce et publicité, expositions virtuelles, bibliothèques numériques, blogs… Beaucoup d'activités se sont déplacées vers les écrans ou sont apparues avec eux. L'enjeu dépasse celui d'une simple mutation technique. Des processus sociaux sont à l'œuvre, qui montrent que des communautés explorent, intègrent et digèrent de multiples possibilités d'édition et d'échanges.

Gildas Illien

Bibliothèque nationale de France
gildas.illien@bnf.fr

Valérie Game

Bibliothèque nationale de France
valerie.game@bnf.fr

Les internautes réinventent les modalités d'intervention dans l'espace public en même temps qu'ils bouleversent les formes, les frontières et les cartes documentaires. La Toile n'est pas seulement un immense réservoir d'informations et de services, c'est aussi une société que l'on s'approprie pour se raconter, se rencontrer, créer du lien et des liens, dans tous les sens du terme. Les institutions de mémoire comme la Bibliothèque nationale de France doivent prendre en compte ce phénomène et, à la faveur de l'extension du cadre juridique du dépôt légal à la Toile, s'atteler à la préservation de cette expression nouvelle de notre patrimoine.

Un nouveau modèle documentaire est à concevoir en même temps que les architectures techniques adaptées, sans oublier la dimension humaine et organisationnelle qui sous-tend toute évolution de cette ampleur. Depuis ses premières expérimentations en 1998, la BnF a parcouru un long chemin qui lui permet d'envisager le passage du mode expérimental à une exploitation courante du dépôt légal d'Internet à l'horizon 2008. Nous nous concentrerons ici sur les aspects juridiques, le modèle de collecte et les évolutions professionnelles qui dessinent le cadre général de sa démarche.

Un cadre juridique qui inscrit le dépôt légal d'Internet dans la continuité de son histoire

Initialement promulguée pour les imprimés en 1537, l'obligation de dépôt légal pour les éditeurs, imprimeurs, producteurs, distributeurs

Diplômé de l'Institut d'études politiques de Paris et titulaire d'un master en communication de l'université McGill (Montréal), ***Gildas Illien*** *est chef du projet de dépôt légal d'Internet au Département de la bibliothèque numérique de la Bibliothèque nationale de France. Il a exercé auparavant à la bibliothèque universitaire de Paris VIII, a fait partie de l'équipe de préfiguration de la bibliothèque de l'Institut national d'histoire de l'art et a dirigé les médiathèques de l'Institut français de Vienne et du centre culturel français d'Oslo. Il est l'auteur d'un essai,* La place des arts et la révolution tranquille : les fonctions politiques d'un centre culturel *(Presses de l'université Laval, 1999).*

Titulaire d'un doctorat en droit international public, ***Valérie Game*** *est directrice du Département des affaires juridiques et de la commande publique de la Bibliothèque nationale de France. Elle était précédemment chef du service juridique de l'Institut national des appellations d'origine.*

et importateurs de documents s'est progressivement étendue à tous les types d'expression et de création, en intégrant à son champ d'application les nouvelles techniques, au fur et à mesure de leur apparition. Après les livres, les estampes, les partitions, les photographies, les affiches, les documents audiovisuels et multimédias, les sites Internet devraient bientôt s'ajouter au périmètre de production éditoriale à conserver par la Bibliothèque nationale de France.

Le titre IV de l'actuel projet de loi « Droit d'auteur et droits voisins dans la société de l'information » (Dadvsi), en cours d'examen au Parlement, prévoit en effet l'extension du dépôt légal à tous « *les signes, signaux, écrits, sons ou messages de toute nature qui font l'objet d'une communication au public par voie électronique* ». Cette obligation deviendra effective dès la promulgation de la loi. Les sanctions pénales pour non-respect de cette obligation n'entreront toutefois pas en vigueur avant un délai de trois ans. Un décret d'application viendra ultérieurement préciser les conditions de sélection et de consultation des informations collectées.

Ce projet de loi a été préparé depuis plus de cinq ans. Le calendrier de son examen législatif a pris du retard. Adopté le 21 mars 2006 par l'Assemblée nationale, le texte sera examiné par le Sénat début mai et devrait être adopté avant l'été. La BnF a pris toutes dispositions afin d'être prête à remplir cette nouvelle mission le moment venu.

Qui est concerné par ce nouveau dispositif ? Du côté des institutions de mémoire, l'Institut national de l'Audiovisuel (Ina) collectera les sites relevant du domaine de la communication audiovisuelle (en particulier ceux de la radio et de la télévision) et la BnF tous les autres. Du côté des opérateurs, l'obligation de dépôt légal pèsera sur les personnes qui éditent et produisent des sites Internet. Contrairement à ce qui est pratiqué pour les autres supports, elle n'impliquera pas de démarche particulière de leur part, car la collecte sera principalement effectuée par le biais de collectes automatiques réalisées par des robots que piloteront les institutions dépositaires.

La seule obligation qui incombera aux producteurs sera de fournir les codes et les informations techniques susceptibles de faciliter l'archivage de leurs sites en cas de difficulté. Une procédure de dépôt pourra en outre être mise en œuvre dans les cas où l'architecture d'un site sélectionné ou les formats utilisés rendraient impossible la collecte automatique.

Les archives de la Toile seront consultables sur place dans les salles de recherche de la BnF comme les autres collections issues du dépôt légal. La loi prévoit, à cette occasion, une exception au droit d'auteur et aux droits voisins au profit des organismes dépositaires, qui leur permettra de reproduire sur tout support et par tout procédé les œuvres pour les besoins du dépôt légal : collecte, conservation et consultation, et de communiquer ces œuvres dans leurs enceintes sur des postes individuels à des chercheurs dûment accrédités.

Le modèle intégré : une réponse pragmatique aux défis de l'archivage de la Toile

Le volume des publications sur Internet est sans précédent : on ne peut archiver la Toile comme on constitue les collections d'une bibliothèque – et encore moins envisager de cataloguer des millions de sites à l'unité. L'exhaustivité ne peut plus être de mise et le recours aux captures et aux traitements automatiques est la seule issue pour conserver une partie significative de cette masse éphémère. L'opération est plus complexe qu'il n'y paraît, car cette nouvelle forme d'archivage s'attache à la fois aux sites Internet en tant qu'unités et aux liens qui tissent des relations entre les pages d'un site et entre les sites eux-mêmes : ce sont des « tissus » de documents entremêlés qu'on capture.

De plus, la collecte automatique ne s'applique souvent qu'à la surface des sites et se heurte aux pièges et aux barrières qui protègent l'accès au « web profond » : qu'ils soient sécurisés, ou qu'ils s'appuient sur des techniques ou des bases de données qu'un robot ne peut capturer, nombreux sont les sites dont on ne peut archiver que la « capsule ». Les archives d'Internet, même réduites à leur portion « d'intérêt national » constituent ainsi un défi du fait de leur volume, de leur architecture et de leur temporalité singulière, qui bat en brèche toute notion de complétude.

Afin d'apporter une réponse pragmatique à ces difficultés techniques comme aux enjeux documentaires, la BnF a choisi une approche qui conjugue trois modes de collecte : des captures massives et automatiques du domaine français ; des collectes ciblées qui s'appuient sur l'expertise de bibliothécaires ; des dépôts à l'unité pour un nombre limité de sites qu'on ne peut archiver autrement.

Des collectes automatiques

La collecte automatique à grande échelle d'instantanés du domaine français est effectuée au moyen du robot Heritrix, première réalisation du consortium international pour la préservation d'Internet (IIPC), piloté par la BnF depuis 2003. Dans le cadre d'un partenariat de recherche avec l'organisme américain Internet Archive, la BnF a réalisé deux instantanés du domaine « .fr », fin 2004 et fin 2005.

Chaque collecte représente 118 à 140 millions de fichiers pour un volume total de 7 téraoctets. Des copies d'instantanés des domaines génériques et français (collections historiques de 1996 à 2004) ont également été acquises : elles représentent plus de 6 milliards de fichiers pour un volume de 60 téraoctets. Outre la sauvegarde d'un certain nombre « d'incunables » de l'Internet, ce passage à l'échelle a facilité la prise en compte par l'établissement d'enjeux techniques et documentaires stratégiques pour la mise en place de son organisation.

Grâce à ces premières livraisons, la BnF est aujourd'hui capable d'appréhender en grandeur réelle la problématique de sélection des sites au regard de la masse, de tester des procédures de contrôle qualité et bientôt des dispositifs d'indexation automatique et de conservation pérenne sur de grands volumes.

Des collectes ciblées

Des collectes thématiques et événementielles viennent compléter ce dispositif par des sélections plus fines. Les critères de sélection de ces collectes ciblées sont en cours de définition.

Deux principes président au choix : il est d'abord indispensable de capturer les sites qui prolongent ou remplacent des collections qui ont engagé, voire achevé, leur migration vers Internet, par exemple les publications en série dont la BnF conserve les collections imprimées, souvent depuis leurs origines.

Pour ce qui est des nouvelles formes de publications qui émergent sur la Toile, il appartient aux « veilleurs » de chaque domaine documentaire de repérer ce qui présente un intérêt particulier dans son champ éditorial et de déterminer à quelle fréquence et à quelle profondeur il faut le capturer. Les nouvelles formes de création numérique en ligne pourront, par exemple, intéresser les bibliothécaires chargés des arts visuels et du spectacle ; les blogs et les sites et forums d'opinion sont un autre exemple de sources potentielles pour l'histoire sociale et politique.

C'est ainsi qu'ont été archivés 3 500 sites des campagnes électorales de 2002 et de 2004 (23 millions de fichiers), tandis que se prépare l'archivage de la campagne de 2007. D'autres collectes thématiques sont prévues. La dernière en date, fin 2005, a permis de capturer 4 500 sites, soit 40 millions de fichiers, dont environ un tiers de blogs.

De manière plus exceptionnelle compte tenu des coûts de traitement, des dépôts spécifiques pourront être effectués par les producteurs de sites à la demande de la BnF. Celle-ci réalise par exemple, depuis juin 2005, l'archivage quotidien de la version électronique du *Journal officiel*.

Les enjeux d'évolution pour le métier

Quantitativement, l'essentiel des données proviendra des collectes automatiques. Pour des raisons économiques évidentes, la part de la sélection humaine et « manuelle » doit se limiter à certaines traces de la Toile.

Les bibliothécaires chargés de ces acquisitions ciblées doivent apprendre à sélectionner des ressources de l'Internet, dans une perspective de représentativité (et plus uniquement de qualité, car il s'agit de dépôt légal, donc de conserver le meilleur

comme le pire), en ajustant leurs pratiques d'évaluation des contenus aux spécificités de la Toile et aux techniques d'archivage automatique.

Il s'agit d'apprendre à analyser et à archiver un site Internet tant au niveau documentaire et logique que physique, et de le situer dans un ensemble qui, par sa structure (hypertextuelle) et sa masse (exponentielle), oblige à reconsidérer totalement les pratiques d'enrichissement des collections.

Un des objectifs « métier » de l'établissement est d'accompagner ces nouvelles pratiques. La BnF a mis en place en 2005 un réseau de 35 correspondants du dépôt légal d'Internet qui apprennent à archiver les sites, évaluent le résultat des collectes et formalisent par la même occasion la première ébauche d'une politique de sélection pour ces collectes ciblées. La veille et la sélection assurées aujourd'hui par la BnF n'excluent pas la participation d'autres partenaires scientifiques et documentaires : cette activité pourrait ainsi devenir un nouveau pan de sa coopération avec les bibliothèques pôles associés lorsque l'établissement aura consolidé son dispositif d'exploitation.

L'autre évolution majeure du métier concerne le traitement physique et intellectuel des archives. L'automatisation de pratiquement tous les processus aujourd'hui assurés par des humains est une nécessité compte tenu du passage à grande échelle mais aussi une opportunité nouvelle liée au support – numérique – de ces nouveaux documents. Dans le cadre de la réalisation de son futur magasin numérique (« la cinquième tour »), la BnF instruit ainsi les procédures d'indexation automatique, de stockage, de conservation et de consultation de ses archives.

Ces évolutions impliquent la définition de nouvelles compétences et de nouveaux profils de postes : par exemple, des « opérateurs numériques » capables d'exploiter au quotidien les processus automatisés de collecte et de traitement, mais aussi des experts en mesure de superviser l'indexation à grande échelle des contenus et de gérer les risques propres à la préservation pérenne des documents numériques alors que les formats et les dispositifs de consultation évoluent et disparaissent très vite.

Un dernier pan, essentiel, de cette évolution métier concerne l'accès public aux archives en salle de lecture. De ce point de vue, le rôle des bibliothécaires concerne d'abord la mise au point d'outils de consultation et de services de médiation adaptés aux besoins des utilisateurs. La préparation des futures conditions de consultation des archives s'appuie sur une collaboration étroite avec des représentants des usagers, notamment des communautés de chercheurs, spécialistes d'Internet et de la sociologie des médias mais aussi de l'histoire sociale.

Une première étude d'usage montée en partenariat avec la bibliothèque pôle associé de Sciences Po et impliquant un panel d'étudiants, de chercheurs et de lecteurs de la BnF sera organisée à l'automne 2006. Cette étude sera l'occasion d'observer pour la première fois la rencontre du public avec un corpus d'actualité : les archives des campagnes électorales de 2002 et de 2004.

Mars 2006

EN SAVOIR PLUS

– Le projet de loi « Droit d'auteur et droits voisins dans la société de l'information » sur le site de l'Assemblée nationale : http://www.assemblee-nationale.fr/12/projets/pl1206.asp

– Le dossier de presse *Les enjeux du dépôt légal de la Toile* sur le site de la BnF : http://www.bnf.fr/pages/dossiers/toile.pdf

– Le site du Consortium international pour la préservation d'Internet : http://www.netpreserve.org