



## Dossier de presse

### Sommaire

---

#### Les enjeux du dépôt légal de la Toile

A la veille du débat parlementaire autour du projet de loi relatif au droit d'auteur et aux droits voisins dans la société de l'information

<b>Communiqué de presse</b>	<b>2</b>
<b>Les enjeux du dépôt légal de la Toile</b>	<b>3</b>
<ul style="list-style-type: none"><li>• Un enjeu de société : l'invention collective d'un nouvel espace public</li><li>• L'extension du dépôt légal à la Toile s'inscrit dans une continuité historique</li><li>• Un défi technologique et documentaire</li></ul>	
<b>Le futur cadre juridique et la répartition des missions entre l'Ina et la BnF</b>	<b>7</b>
<ul style="list-style-type: none"><li>• Les dispositions relatives au dépôt légal de la Toile dans la loi soumise au vote du Parlement</li><li>• La répartition des missions</li></ul>	
<b>La démarche de l'Institut national de l'audiovisuel</b>	<b>10</b>
<ul style="list-style-type: none"><li>• Le domaine médias</li><li>• La mise en œuvre par l'Ina de ses nouvelles responsabilités</li></ul>	
<b>La démarche de la Bibliothèque nationale de France</b>	<b>14</b>
<ul style="list-style-type: none"><li>• La collecte des sites : le modèle intégré</li><li>• L'accès public aux archives</li><li>• L'archivage de la Toile, enjeu de coopération</li></ul>	
<b>Les annexes</b>	<b>21</b>
Ina	
<ul style="list-style-type: none"><li>• Le cas particulier du <i>streaming</i></li><li>• La chaîne de traitement expérimentale (graphique)</li><li>• Les spécificités du domaine médias (graphique)</li></ul>	
BnF	
<ul style="list-style-type: none"><li>• Le traitement documentaire</li><li>• Glossaire du dépôt légal d'Internet à la BnF</li></ul>	

---



Communiqué de presse

## La BnF et l'Ina, gardiens de la mémoire de la Toile

La Bibliothèque nationale de France et l'Institut national de l'audiovisuel sont désormais prêts à archiver la mémoire de la Toile pour assurer sa conservation à long terme et sa communication.

Très prochainement, le projet de loi « Droit d'auteur et Droits voisins dans la société de l'information » sera présenté au Parlement. Ce texte étend le champ du dépôt légal aux sites Internet français et désigne la BnF et l'Ina comme les organismes dépositaires.

Plusieurs années de tests et d'expérimentations diverses ont armé ces deux institutions, désormais aptes à procéder à une collecte automatique ou, le cas échéant, manuelle, des sites relevant du domaine correspondant à sa mission initiale, dans une logique de continuité des collections. Ainsi, l'Ina collectera-t-elle les sites relevant du domaine de la communication audiovisuelle, la BnF tous les autres.

Cette disposition permettra de maintenir, d'enrichir et d'étendre les collections patrimoniales mais aussi de garder une trace de ce nouveau média, composante de la mémoire sociale et culturelle, référence pour la recherche et source pour l'histoire.

---

### Contacts presse

BnF : Claudine Hermabessière, chargée des relations avec la presse  
Tel : 01 53 79 41 18, mél : [claudine.hermabessiere@bnf.fr](mailto:claudine.hermabessiere@bnf.fr)  
Philippa Tomasi, tel : 01 53 79 46 76, mél :  
[philippa.tomasi@bnf.fr](mailto:philippa.tomasi@bnf.fr)

Ina : Martine Tomasso  
Tel : 01 53 79 46 76, mél : [mtomasso@ina.fr](mailto:mtomasso@ina.fr)

# Les enjeux du dépôt légal de la Toile

## L'archivage de la Toile : un enjeu de société pour la mémoire collective

La révolution d'Internet se caractérise d'abord par l'essor exponentiel du réseau et de son audience. Si la « fracture numérique » n'est pas réduite, les statistiques indiquent que la Toile n'est plus l'apanage de communautés privilégiées. On compte aujourd'hui en France 25 millions d'internautes<sup>1</sup> et 12,2 millions de foyers sont équipés d'un micro-ordinateur. Il existe en 2005<sup>2</sup> plus de 62 millions de sites dans le monde. Le nombre de sites publics français est également en augmentation constante : notre domaine national, le « .fr » en recensait près de 400 000 en octobre 2005, un chiffre qui a doublé en deux ans mais qui est bien inférieur à la réalité car il faut aussi prendre en compte les nombreux sites produits en France ou concernant la France qui relèvent de domaines étrangers ou génériques tels que le « .net », le « .org » ou le « .com », ce dernier représentant à lui seul 30 millions de sites dans le monde.

« E-administration », arts numériques, radio et télévision diffusés sur Internet, information et édition en ligne, enseignement à distance, commerce et publicité, correspondances électroniques et « clavardages », expositions virtuelles et bibliothèques numériques... beaucoup d'activités tant publiques que privées se déplacent vers les écrans, enrichissent et densifient le réseau de la Toile. Souvent, les services en ligne viennent doubler ou compléter les activités du monde physique, à moins qu'ils ne s'y substituent complètement. Dans bien des cas, la situation reste hybride mais l'on assiste à une migration progressive des supports, qui varie selon les secteurs d'activité et les domaines du savoir. L'enjeu dépasse cependant celui d'une simple mutation technique.

Des processus sociaux sont à l'œuvre qui montrent que des individus, isolés ou en groupes, explorent, intègrent et digèrent les nouvelles possibilités d'édition et d'échanges. Si la création d'un site personnel ou d'un blog\* (« bloc notes ») sur Internet s'apparente juridiquement à une publication, les internautes ont inventé de nouvelles intimités, de nouvelles proximités, redéfinissant les frontières et les modalités d'intervention dans l'espace public. La Toile n'est plus seulement un immense réservoir d'informations et de services, c'est aussi une communauté d'échanges que l'on choisit d'intégrer pour se raconter, se rencontrer, créer du lien et des liens, dans tous les sens du terme. 84% des internautes français utilisent la messagerie, 18,8% s'adonnent au clavardage. On recense déjà 4 millions de blogs, qui traduisent un engouement massif pour l'auto-publication.

La généralisation de la numérisation des contenus et leur circulation dans des réseaux virtuels bouleversent profondément l'économie de la mémoire : plus les contenus sont volatils, plus il importe de les conserver. Internet est à la fois un vecteur de diffusion vers lequel convergent les supports qui nous sont familiers (la presse, la musique, le livre ou le film...) et un espace public d'un genre nouveau qui favorise l'émergence d'objets culturels et de processus sociaux inédits. Loin de se réduire à des effets de mode, ils forment sans aucun doute des éléments structurants de notre mémoire collective et témoignent des changements profonds qui affectent la société. La BnF et l'Ina se doivent d'archiver ces traces qui permettront aux chercheurs de demain de disposer des outils de lecture et de compréhension des évolutions contemporaines.

---

<sup>1</sup> En septembre 2005, selon Mediamétrie, la France comptait plus de 25 millions d'internautes, soit 48,4% de la population française (sur une population d'individus de 11 ans et plus).

<sup>2</sup> Source : « Web survey » Netcraft LTD, avril 2005.

\* cf. glossaire p.23

## L'extension du dépôt légal à la Toile s'inscrit dans une continuité historique

Le choix du législateur d'intégrer l'archivage d'Internet au dispositif juridique du dépôt légal s'inscrit dans la tradition française et prolonge la double mission du dépôt légal : assurer la continuité et la complétude des collections en intégrant chaque nouveau support, collecter les œuvres et les objets révélateurs des mutations d'une époque.

**Le dépôt légal est l'obligation faite par la loi à tout éditeur, imprimeur, producteur, distributeur, importateur de documents d'en effectuer un dépôt auprès des organismes désignés par la loi.** Initialement promulgué pour les imprimés, il s'est progressivement étendu 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. Le dépôt légal a été institué en France par une succession de textes législatifs et réglementaires qui ont permis, au fil des siècles, de constituer une collection patrimoniale comportant tous les types de documents et de supports. C'est ainsi que les organismes dépositaires reçoivent par voie de dépôt légal depuis :

1537 Les imprimés  
1648 Les estampes, ainsi que les cartes et plans  
1793 Les partitions musicales  
1925 Les photographies, arts graphiques de toute nature  
1938 Les phonogrammes  
1941 Les affiches  
1975 Les vidéogrammes et les documents multimédias  
1992 Les documents audiovisuels de la radio télévision, l'édition électronique sur support (progiciels, bases de données et systèmes experts).<sup>4</sup>

Si la BnF assure ainsi depuis plusieurs siècles la **collecte** et la **conservation** des documents qui forment aujourd'hui le cœur de ses collections, l'Ina s'est vu confier plus récemment par le législateur<sup>5</sup> une mission spécifique de conservation patrimoniale du domaine général de la radio et de la télévision, quel que soit le procédé technique de diffusion (hertzien, câble, satellite, numérique hertzien, ADSL, Internet). Ainsi, la mise en œuvre du dépôt légal a concerné dans un premier temps la télévision nationale hertzienne, puis les chaînes du câble et du satellite. Les programmes de la TNT, de la radio et de la télévision par ADSL ou par Internet ont naturellement vocation à rejoindre ce champ de la conservation patrimoniale. Le cœur du métier de l'Ina en matière de dépôt légal se situe clairement dans le domaine de la communication audiovisuelle.

Le dépôt légal prévoit également la **constitution et la diffusion de bibliographies nationales**. Enfin, la **consultation des documents** par le public à des fins de recherche est envisagée sous réserve des secrets protégés par la loi, dans des conditions conformes à la législation sur la propriété intellectuelle et compatibles avec leur conservation.

---

<sup>4</sup> la loi de 1992 a créé le dépôt légal de la radio et de la télévision, attribué à l'Institut national de l'audiovisuel, a confié celui des films cinématographiques au Centre National de la Cinématographie et a élargi les responsabilités de la Bibliothèque nationale de France à l'édition électronique sur support chimique.

<sup>5</sup> Article L 132-3 du Code du patrimoine (ordonnance n° 2004-178 du 20/02/04 reprenant les dispositions de la loi du 20 juin 1992 relative au dépôt légal) et loi n° 86-1067 relative à la liberté de communication.

## L'archivage de la Toile : un défi technologique et documentaire

Les archives d'Internet, même réduites à leur portion « d'intérêt national », constituent un défi du fait de leur masse, de leur architecture et de leur temporalité singulière. Des contraintes particulières s'appliquent en outre aux documents audiovisuels diffusés en ligne.

La Toile se caractérise d'abord par sa masse. Ainsi, les annuaires de sites utilisés aux débuts d'Internet ont disparu au profit des moteurs de recherche car un recensement thématique des sites à l'unité n'était plus envisageable. Les chiffres parlent d'eux-mêmes : une collecte du domaine français réalisée par la BnF du 15 décembre 2004 au 30 janvier 2005 a abouti à la réception de 3 Teraoctets\* de données, représentant un total de 118 380 000 fichiers identifiés par une adresse URL\* parmi lesquels des textes, des images, des sons, des vidéos... de quoi saturer les disques durs de 150 ordinateurs personnels ! L'archivage de tels volumes d'informations pose évidemment des problèmes techniques de stockage et de conservation, mais implique également de redéfinir les modalités de sélection et de traitement documentaires par les institutions depositaires : on ne peut archiver la Toile comme on constitue les collections d'une bibliothèque. Si la sélection et le traitement manuels restent pertinents dans des cas précis que la Bibliothèque nationale de France et l'Ina ont définis chacun en fonction de leurs missions respectives, l'exhaustivité n'est plus permise et seul le recours aux captures et aux traitements automatiques permettra de conserver une partie significative de ce patrimoine immatériel et souvent éphémère.

**L'unité intellectuelle retenue pour archiver la Toile est celle du site**, qui recouvre des ensembles documentaires dont le volume et les caractéristiques techniques sont très variables. L'archivage des sites s'attache à la fois aux sites en tant qu'unités et aux liens qui tissent des relations entre les pages d'un site et entre les sites eux-mêmes. Autrement dit, les liens signalés dans un site sont considérés comme faisant partie de son contenu, ce qui représente à la fois une opportunité précieuse de situer l'archive dans son contexte (par exemple, pour connaître sa notoriété) et un casse-tête technique : la collecte de la Toile ne peut se passer d'un point de départ (souvent, une liste de sites jugés représentatifs de la production nationale à un moment donné), mais on ne connaît jamais exactement son périmètre ni son volume à l'arrivée.

**La collecte automatique à grande échelle** telle que la pratiquent nombre d'institutions engagées dans l'archivage de la Toile ne s'applique qu'à la « surface » des sites, accessible aux robots. Les archives rassemblées selon cette méthode représentent une « photographie instantanée »\* d'un ensemble de sites et offrent un panorama de la diversité et de la richesse des contenus publiés. Toutefois, ces collectes se heurtent aux pièges et aux barrières qui protègent l'accès au « web profond »\* : il en est ainsi des sites sécurisés (recours à un mot de passe), ou qui s'appuient sur des techniques, des logiciels ou des bases de données qu'un robot ne peut capturer ; nombreux sont les sites dont on ne peut archiver que la « capsule ».

---

\* cf. glossaire p 25

Ce type de problème se pose avec acuité pour les sites de communication audiovisuelle qui relèvent de la compétence de l'Ina, les sites médias en particulier, qui constituent un sous-ensemble très actif de la Toile. En effet, si le nombre des sites qui la composent est encore aujourd'hui assez restreint (environ 10 000), le volume total des contenus (flux audio/vidéo notamment) représente une part majeure du volume global d'intérêt national<sup>6</sup>. Ces sites sont situés pour moitié dans le domaine « .com », et seulement à hauteur de 30% dans le domaine « .fr ». Ils regroupent des contenus à forte valeur ajoutée, en terme de mise en forme propre au monde de la diffusion audiovisuelle. Leur réalisation suppose l'utilisation de technologies de diffusion très élaborées (*streaming*, Flash, XUL, etc.), dont il importe d'assurer un archivage adapté qui permettra d'appréhender chaque information dans son contexte visuel, sonore et interactif original. Ils sont également caractérisés par un taux de rafraîchissement très élevé – certains sites se situant délibérément du côté de la diffusion en direct partielle ou totale – ce qui implique des procédures particulières de gestion des fréquences des mises à jour.

L'archivage de la Toile constitue donc **un défi technologique** que l'Ina comme la BnF ont choisi de relever. Pour capturer, traiter, stocker, visualiser mais aussi assurer la pérennité de ce nouveau type de document – l'archive de site – plusieurs chantiers ont été ouverts, dont certains rejoignent la problématique plus générale de la conservation des documents numériques, autre projet phare de la BnF dans le cadre de la constitution de sa bibliothèque numérique.

Il s'agit aussi d'un **enjeu organisationnel et humain** considérable, car la collecte de la Toile suppose l'acquisition de compétences radicalement nouvelles, voire la réinvention de certains métiers.

---

<sup>6</sup> Une heure de vidéo composée à 2 MB/s représente environ 1000 livres imprimés

# Le futur cadre juridique et la répartition des missions entre l'Ina et la BnF

## Dispositions législatives relatives au dépôt légal de la Toile soumises au vote du Parlement

### Le dépôt légal des sites Internet

Le projet de loi étend le champ du dépôt légal aux signes, signaux, écrits, images, sons ou messages de toute nature qui font l'objet d'une communication au public par voie électronique.

L'obligation vise les personnes qui éditent et produisent les sites Internet.

- La loi habilite les organismes en charge du dépôt légal à collecter les contenus en ligne selon des procédures automatiques. Une information sur les procédures de collecte mises en œuvre sera disponible. Un amendement du député-rapporteur Christian Vanneste prévoit qu'un code ou une restriction d'accès ne doit pas empêcher cette collecte.
- La loi prévoit également le dépôt de supports ou l'envoi de fichiers. Le dépôt ou envoi n'interviendrait que dans les cas où la collecte serait techniquement impossible. Dans cette hypothèse, les modalités de transfert des fichiers seront déterminées en accord avec les éditeurs et producteurs de sites.

Les conditions de sélection et de consultation des sites seront définies par les décrets d'application.

### Une normalisation des relations entre l'exercice du dépôt légal et les règles du Code de la propriété intellectuelle

L'exercice de la mission de dépôt légal s'accompagne d'actes qui mettent en jeu le droit de la propriété intellectuelle. La nécessité de conserver les documents a toujours induit le besoin de les reproduire. La reproduction par voie numérique et la diffusion de ces documents sur des postes individuels d'écoute et de lecture dans les emprises des organismes dépositaires rendent donc nécessaire l'intervention du législateur.

À l'ère du numérique, une normalisation des rapports entre l'exercice de la mission de dépôt légal et les règles du Code de la propriété intellectuelle est indispensable.

Aussi le texte prévoit-il une exception aux droits d'auteur, droits voisins et droit des producteurs de bases de données au profit des organismes en charge du dépôt légal.

Les organismes en charge du dépôt légal pourront licitement, sans avoir à requérir d'autorisation préalable, ni à verser de rémunération:

- reproduire sur tout support et par tout procédé les œuvres pour les besoins du dépôt légal : collecte, conservation, consultation ;
- offrir à la consultation ces œuvres dans leurs emprises, sur des postes individuels de consultation, à des chercheurs accrédités.

Cette exception ne vise pas les reproductions demandées par les lecteurs pour leurs besoins propres, ni les reproductions à des fins commerciales ; la communication à distance n'est pas autorisée.

## Répartition des missions entre organismes dépositaires

L'article 26-IV du projet de loi indique que « l'Ina participe, avec la BnF, à la collecte, au titre du dépôt légal, des signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique ».

La répartition des missions entre les organismes dépositaires n'est pas explicitement définie au niveau législatif.

C'est le décret d'application de la loi relative au dépôt légal (Loi du 20 juin 1992 introduite dans le Code du patrimoine par ordonnance du 20 février 2004) qui apporte ce niveau de précision.

**Trois grands principes** peuvent permettre de déterminer la répartition des interventions de l'Ina, de la BnF en matière de conservation patrimoniale des sites de l'Internet :

- le principe de la cohérence de l'activité par rapport à la mission initiale ;
- le principe de son inscription dans la logique des métiers et compétences de chacun (par exemple traitement du flux à l'Ina) ;
- le principe de la continuité des collections.

Ainsi s'est organisé la répartition des compétences entre l'Ina et la BnF, rappelée ci-après.

## Les compétences générales de la Bibliothèque nationale de France

La BnF a pour mission de collecter et de conserver au titre du dépôt légal :

- les documents imprimés ou graphiques de toute nature notamment les livres, périodiques, brochures, estampes, gravures, cartes postales, affiches, cartes, plans, globes et atlas géographiques, partitions musicales, chorégraphies ainsi que les documents photographiques ;
- les logiciels et bases de données ;
- les phonogrammes, vidéogrammes et documents multimédias.

Le dispositif de collecte et de conservation imaginé par la BnF dans le cadre du dépôt légal de la Toile s'inscrit dans la continuité de cette histoire et respecte cette répartition des rôles. La Bibliothèque nationale de France veillera ainsi à préserver les documents relevant de sa responsabilité lors de leur migration totale ou partielle vers Internet ainsi que les nouveaux types de documents qui apparaîtront sur la Toile à mesure de son développement.

Un modèle intégré, qui associe collectes automatiques, sélections ciblées et dépôts, a été retenu pour la mise en place de ce dispositif à la BnF.

## Les compétences générales de l'Institut national de l'audiovisuel

L'Ina exerce une mission de conservation patrimoniale du domaine général de la radio et de la télévision, quel que soit le procédé technique de diffusion (hertzien, câble, satellite, numérique hertzien, ADSL, Internet).

Ainsi, la mise en œuvre du dépôt légal a concerné dans un premier temps la télévision nationale hertzienne, puis les chaînes du câble et du satellite.

Les programmes de la TNT, de la radio et de la télévision par ADSL ou par Internet ont naturellement vocation à rejoindre ce champ de la conservation patrimoniale.

Il s'agira donc de considérer l'ensemble de l'offre de ces médias sur Internet : par exemple la captation de TF1 concernera son site et pas seulement ses web-TV.

Et si le cœur de l'activité patrimoniale de l'Ina sur la Toile concerne les sites médias radio et télévision, seront également pris en compte, dans l'esprit du dispositif actuel du dépôt légal relatif aux documents d'accompagnement, un ensemble de sites d'accompagnement et de sites périphériques, notamment :

- des sites liés aux programmes eux-mêmes, par exemple le site de *Loft Story* viendra compléter l'émission diffusée en hertzien par M6 et sa version diffusée sur un canal du satellite ;
- des sites liés à des événements relatifs aux médias ;
- des sites représentatifs d'un environnement professionnel ou institutionnel concourant à l'activité de radio et de télévision sur la Toile, par exemple seront collectés les journaux de programmes en ligne ou le site du CSA.

Enfin, sera collectée toute source en ligne d'information complémentaire permettant, dans une économie de moyens, de compléter l'information sur les contenus archivés.

# La démarche de l'Institut national de l'audiovisuel

## Le domaine médias

La conservation du web à l'Ina a pour objet la collecte suivie dans le temps, l'indexation et la mise à disposition du public des sites web relevant de la « communication audiovisuelle », un sous-ensemble de la communication au public par voie électronique, définie par les articles 1 et 2 de la loi « Pour la confiance dans l'économie numérique » du 21 juin 2004 et par l'article 2 de la loi du 30 septembre 1986 relative à la liberté de la communication.

## Un domaine hétérogène

Il est composé d'acteurs très différents regroupés en 4 grandes catégories :

- Au cœur, se trouvent **les sites de radio et de télévision proposant un contenu organisé selon une logique éditoriale** proche d'une grille de programme, dotés d'une Une, de rubriques régulières et d'une programmation qui s'apparente à celle des vecteurs de diffusion traditionnels, hertzien, câble ou satellite. Ces sites, qui représentent entre 15 et 20 % de l'ensemble, sont, dans leur grande majorité, adossés à un service de communication pré-existant, qu'il s'agisse d'une chaîne de télévision, par exemple *LCl.fr* ou *Arte-radio.com*, ou d'une radio, par exemple *les sites de Radio France* ou ceux des radios locales privées qui, très tôt, ont pratiqué la diffusion Internet.  
Quant aux pures web-tv, après des débuts chaotiques, elles bénéficient aujourd'hui de la montée en charge des réseaux haut débit et se multiplient sur des contenus thématiques, par exemple *Canal U*, la web-télé des universités, ou portées par des collectivités locales (ex : *Canal Bretagne* ou *Cité-Tv lyonnaise*).
- Le 2<sup>ème</sup> cercle, qui regroupe plus de la moitié des sites du domaine, concerne ceux qui sont liés aux **programmes diffusés sur une chaîne**. On y trouve des sites d'émissions, (par exemple *Planète Thalassa* ou le site des *Guignols de l'info*) ou de séries ( ex : site *d'Urgences*), ou encore de héros de séries, officiels et personnels, comme les sites de *Buffy*, de *Columbo* ou de *Tintin*, et enfin ceux consacrés aux personnalités des médias (artistes ou animateurs, *Planète Arthur* par exemple ou le site de *Jean-Pierre Coffe*). On y trouve également des sites événementiels liés à l'actualité, comme le site « *l'année du Brésil* » de France 5 ou celui du *festival de Cannes de Canal+*, et aujourd'hui la plupart des sites médias accueillent ou proposent des *blogs*.
- La 3<sup>ème</sup> catégorie englobe les **sites en relation directe ou indirecte avec l'activité des radios et des télévisions** : sites institutionnels (CSA), mais aussi de sociétés (Vivendi pour Canal+) ou de prestataires. Ils représentent 10 à 15% de l'ensemble.
- Dans une 4<sup>ème</sup> catégorie, se situent les **sites à vocation documentaire** qui proposent des annuaires spécialisés, des guides web, des portails de bouquets de télévisions ou des guides de fréquences radio. Citons par exemple *Loft TV*, l'annuaire des web-tv, *ComFM* pour les webradios ou encore *Series-onair.com* pour les séries télévisées. Ils sont actuellement estimés à 5 % de l'ensemble.

Ces sous-domaines forment autant de cercles concentriques autour du cœur de métier audiovisuel de l'Ina, constituant un corpus dont la cohérence thématique est contrebalancée par l'hétérogénéité des composantes éditoriales (politique de mises à jour, types de contenus, volumes, etc.). L'essentiel des besoins en terme d'archivage (collecte, stockage et indexation), se trouve ainsi concentré sur un cœur de domaine comportant un nombre réduit de sites publiant continuellement un grand nombre de contenus. Le reste du domaine comporte essentiellement des sites plus statiques dans leurs publications, mais leur nombre, leur confinement et leur volatilité imposent un travail continu de veille prospective, pour mettre en place leur archivage.

## La mise en œuvre par l'Ina de ses nouvelles responsabilités

### Principes généraux

*Issue des travaux de recherche et de développement menés depuis 4 ans par l'Ina, la mise en œuvre repose sur des campagnes cycliques d'aspiration, effectuées par le robot WebCollecte - un logiciel spécifique développé par l'Ina - de la totalité des sites du domaine.*

*En amont de la chaîne de collecte, une cellule de veille composée d'une équipe documentaire est chargée de prospecter le Web pour identifier et suivre l'évolution des sites du domaine.*

Chaque nouveau site identifié comme appartenant au domaine est indiqué à l'*ordonnanceur* de captation, le système automatique chargé de la planification des collectes. L'*ordonnanceur* établit le calendrier des mises à jour : il décide à quelle fréquence chaque site doit être collecté, selon son dynamisme. Il dirige le logiciel de collecte chargé d'enregistrer chaque site, et récupère le résultat, c'est-à-dire une copie du site à un instant « t ». Cette copie est traitée en séparant sa structure et ses contenus. Ce principe permet de mieux appréhender l'organisation et l'évolution des informations.

La phase d'indexation, semi-automatique, consiste à enrichir et documenter les informations pour faciliter leur recherche et leur consultation. Le résultat des collectes et de l'indexation est ensuite archivé dans un système de stockage pour permettre sa consultation par des chercheurs, en recréant l'interactivité du Web archivé.

### La collecte

La collecte est un processus automatique, mettant en œuvre des robots dirigés par un serveur central, appelé *ordonnanceur*. Ce dernier organise les collectes selon le plan de mise à jour, envoyant simultanément chaque robot sur un site différent, puis récupérant la copie complète du site une fois le travail du robot terminé. Le système indique également à la cellule de veille de nouveaux sites pouvant faire l'objet d'une future collecte.

La copie de chaque site est ensuite préparée pour une indexation séparée des contenus du site et de leur structure, c'est à dire l'URL, la date à laquelle ils ont été trouvés, leurs liens avec d'autres contenus, formant ainsi une sorte de graphe du site.

## L'indexation

L'indexation est l'étape de documentation et d'enrichissement de l'archive. La séparation de la structure et du contenu est essentielle dans cette phase : les contenus sont tout d'abord traités de manière totalement déconnectée de la structure qui était la leur sur le site. On traitera ainsi indépendamment les différentes images qui composent une page, la page elle-même, les films, les musiques, *etc.* Ces contenus sont accompagnés d'une clé identifiante, sorte de signature, calculée par le robot au moment de la collecte. La question de savoir si un contenu a déjà été collecté revient ainsi à vérifier si sa signature est déjà connue. Ainsi, lors de deux collectes successives d'un même site, seuls les contenus réellement nouveaux feront l'objet d'une indexation et d'un stockage. Cette identification permet, en outre, de suivre l'évolution des contenus à travers le site, ou de constater la présence d'un même contenu sur différents sites (logs, photos d'actualités, *etc.*).

Pour les contenus nouveaux, la phase suivante est l'indexation automatique qui consiste en l'extraction des données textuelles permettant de classer ces contenus. Le but est de pouvoir, par la suite, procéder à une recherche sémantique dans l'archive. Certains formats sont plus difficilement indexables que d'autres, notamment les extraits vidéo pour lesquels les techniques d'extraction de texte dans l'image ou de transcription de la bande sonore sont encore peu fiables et trop coûteux en termes de ressources machines. Dans tous les cas, l'indexation du contenu peut être enrichie automatiquement par le contexte fourni par les données de structure. Une image peut ainsi être indexée automatiquement grâce au texte qui l'entoure au sein de la page dans laquelle elle a été trouvée. Un extrait vidéo peut également être indexé selon le texte des pages qui la référencent, ou bien selon des informations de sous-titrage associés. Un même contenu trouvé ou référencé sur plusieurs pages, à plusieurs époques, ou sur plusieurs sites, verra ainsi son indexation s'enrichir d'autant.

Le fait qu'un même contenu fasse l'objet de nombreuses références permettra également de les mettre en avant pour une éventuelle documentation manuelle. L'essentiel de la documentation manuelle se situera cependant à l'échelle du site et sera effectuée par une cellule de veille.

## La consultation de l'archive

La consultation est bien évidemment le dessein final de cette archive. Il y a deux étapes à bien distinguer lors de la consultation : la reconstitution de l'interactivité d'un site à une date donnée (sorte de machine à remonter le temps), et l'accès à toutes les données d'enrichissement de l'archive (indexation, versions, cartes, analyse, *etc.*). Le but de l'archive n'est en effet pas uniquement de permettre la consultation des sites tels qu'ils étaient à une époque donnée, mais également de fournir les outils pour rechercher, classer, visualiser et analyser ces informations.

# La démarche de la Bibliothèque nationale de France

## La collecte des sites : le modèle intégré

Afin d'apporter une réponse pragmatique mais complète aux difficultés techniques comme aux enjeux documentaires et patrimoniaux du dépôt légal de la Toile, la Bibliothèque nationale de France a choisi une approche qui conjugue trois modes de collecte :

- des captures massives et automatiques du domaine français réalisées au moyen de robots ;
- des collectes thématiques et événementielles qui se fondent sur l'expertise de bibliothécaires travaillant dans les départements de collection et de dépôt légal ;
- la mise en place d'un circuit de dépôts à l'unité pour un nombre limité de sites qu'on ne peut archiver autrement.

La plus grande part des archives conservées proviendra des captures automatiques : seules ces dernières sont capables de fournir des collections dont le volume est à la mesure de la Toile. Le recours aux collectes thématiques et aux dépôts évite d'abord de laisser échapper certains sites essentiels. Il permet également de sensibiliser et de former les bibliothécaires au repérage et à la sélection d'un type de document qui prendra demain une place décisive dans l'enrichissement des collections et l'organisation des services proposés au public.

### 1. Les collectes automatiques

La BnF réalise actuellement ses collectes automatiques en partenariat avec Internet Archive, un organisme américain à but non lucratif, pionnier (dès 1996) dans l'archivage de la Toile au niveau mondial. Les données ainsi collectées constituent non seulement une amorce importante aux collections d'archives de la BnF pour l'avenir mais aussi un matériau indispensable à la poursuite du travail d'analyse qui permet aujourd'hui de préfigurer les méthodes et l'organisation fonctionnelle à l'œuvre après le passage d'un stade expérimental à un stade opérationnel. De plus, les robots jouent un rôle exploratoire car ils révèlent souvent l'existence de sites pertinents mais inconnus des bibliothécaires.

Les données collectées à ce jour représentent un volume de plusieurs dizaines de Teraoctets. Elles sont stockées sur une tour de deux mètres de haut intégrant les serveurs des disques durs comportant les données, leur index et un logiciel d'accès. Parce que les unités de mesure ne cessent de changer d'échelle pour prendre en compte la croissance des volumes, on commence à parler en « Petaoctets »\* ( $2^{50}$  octets) !

La BnF travaille à la **définition de scénarios alternatifs à l'actuel dispositif de fourniture d'archives par Internet Archive**. En s'appuyant sur des estimations financières et techniques complétées par une étude approfondie des modèles mis en place dans d'autres bibliothèques nationales, au Danemark notamment, elle entend évaluer précisément les ressources et l'organisation nécessaires à la réalisation de sa propre infrastructure de collecte. Les coûts étudiés concernent aussi bien l'infrastructure technique (bande passante additionnelle, machines, stockage) que les moyens humains. Cette étude devrait permettre de proposer soit d'autres modalités d'externalisation des services de collecte à grande échelle, soit une internalisation au sein de l'établissement.

---

\* cf. glossaire p 25

Le second projet de la BnF concerne le **développement d'un robot plus performant** que celui aujourd'hui utilisé. L'apport principal de cette nouvelle technique consiste à rendre possible la gestion des priorités de collecte : on pourrait ainsi demander au robot de passer plus de temps sur la collecte de certains sites ou de les visiter dans un ordre hiérarchisé. En affinant la capture d'Internet, on limiterait ainsi le risque de perdre certaines ressources essentielles, mais également celui d'augmenter inutilement la masse d'archives peu pertinentes que le dispositif actuel ne permet ni d'extraire ni de filtrer. Ce robot intelligent a pour objectif de restreindre la collecte des archives à un volume un peu moins important par sa taille mais plus pertinent par son contenu. Outre son intérêt documentaire, cette réalisation autoriserait des économies de stockage. Ce projet d'envergure est conduit en partenariat avec d'autres bibliothèques nationales membres du Consortium IIPC<sup>9</sup>.

## 2. Les collectes thématiques et événementielles

**Les collectes thématiques visent à pallier les insuffisances des robots en s'appuyant sur une prospection documentaire ciblée. Il s'agit de signaler manuellement au robot l'adresse des sites à conserver et de lui indiquer selon quelle fréquence et à quelle « profondeur » il doit les capturer.**

Les deux principaux critères de sélection retenus pour ce type de traitement seront la continuité des collections et le signalement de nouveaux objets documentaires particulièrement représentatifs des formes nouvelles de l'édition.

Il apparaît d'abord indispensable de capturer les sites qui prolongent ou remplacent des collections qui ont engagé, voire achevé, leur migration vers Internet : on doit notamment pouvoir suivre l'évolution des publications en série dont la BnF conserve les collections sur format imprimé souvent depuis leurs origines. C'est, par exemple, le cas des revues scientifiques ou des publications officielles, de la musique et bientôt du cinéma. S'agissant des nouvelles formes de publications qui émergent sur la Toile, il appartiendra aux experts de chaque domaine documentaire de repérer celles qui présentent un intérêt particulier dans leur champ éditorial : les nouvelles formes de création numériques en ligne pourront par exemple intéresser les départements chargés des arts visuels et du spectacle ; les blogs offrent un autre exemple de sources potentielles pour l'histoire sociale et politique.

La BnF vient de réaliser sa première collecte thématique sur ce modèle. Ce travail a mobilisé pendant plusieurs mois un réseau d'une trentaine de bibliothécaires, qui ont repéré environ 4000 sites méritant un traitement approfondi. L'évaluation qualitative des résultats de cette collecte permettra de préciser dès 2006 les critères de sélection et les moyens humains et techniques nécessaires à l'organisation de prochaines campagnes thématiques.

---

<sup>9</sup> Consortium International pour la Préservation d'Internet (IIPC) : voir la fiche « coopération » concernant la Bibliothèque nationale de France.

Des campagnes de collecte peuvent être organisées autour d'un événement d'intérêt national. D'autres pays ont, par exemple, choisi de couvrir des événements comme les attentats du 11 septembre 2001 ou, plus récemment, le *tsunami* en Asie. En effet, les collectes automatiques ou ciblées organisées une à plusieurs fois par an ne sont pas adaptées au dépistage de sites événementiels qui surgissent et disparaissent très rapidement : il importe alors de faire preuve d'une réactivité forte, mais aussi d'avoir déterminé au niveau de l'établissement une approche cohérente pour définir et qualifier les événements afin d'éviter la dispersion.

La Bibliothèque nationale de France a, pour le moment, choisi de se concentrer sur l'archivage thématique des sites électoraux en France. Environ 1900 sites ont ainsi été archivés lors des élections présidentielle et législatives de 2002, et près de 1700 lors des élections régionales et européennes de 2004. L'expérience sera reconduite pour les élections de 2007 et l'organisation du travail de repérage et de collecte des sites, dont le pilotage sera confié aux spécialistes de la BnF dans le domaine des sciences politiques, débutera dès 2006.

### 3. Le dépôt de sites

**Le dépôt de sites est une démarche complémentaire de la collecte automatique (qu'elle soit large ou ciblée). Elle consiste à traiter manuellement, de manière unitaire et hors contexte (c'est-à-dire sans archiver les contenus pointés par les liens sortants), un nombre limité de sites, voire des portions de sites, échappant à la capture automatique pour des raisons techniques ou parce que l'accès en est réservé.** Cette démarche implique des échanges et un suivi régulier entre les services de la BnF et les producteurs, qui s'apparente à l'organisation traditionnelle du dépôt légal pour les autres supports. Le circuit des dépôts implique en effet les étapes suivantes : décision de dépôt, contact avec le producteur, instruction technique, choix d'un dispositif et d'une périodicité de collecte, transfert des contenus, validation, archivage, signalement, mise à disposition du public et suivi régulier avec le producteur jusqu'à extinction du site.

Les critères de sélection seront identiques à ceux appliqués aux collectes thématiques. Dans l'immédiat, la priorité est d'assurer la continuité des collections dans des cas de migration de support concernant des publications importantes. Les dépôts, qu'ils soient proposés par les producteurs ou requis par la Bibliothèque dans le cadre du dépôt légal, devront faire l'objet d'une demande motivée et d'une instruction technique préalable pour s'assurer de leur pertinence et de leur faisabilité.

En 2001 et en 2002, la BnF a réalisé une première série de dépôts à titre expérimental. En 2004, elle a entrepris le dépôt de publications officielles qui constituent une priorité compte tenu des ses missions. Ce travail a fait l'objet d'une concertation avec la Direction des Archives de France et la Direction des Journaux officiels. Le dépôt légal du Journal Officiel de la République Française électronique est effectif depuis le 1<sup>er</sup> juin 2005 : chaque nuit, un robot va recueillir les données de la dernière édition sur le serveur du J.O et les rapatrier dans les archives de la BnF. Ce travail sera poursuivi en 2006 en lien étroit avec les producteurs dans le cadre d'une coopération institutionnelle et scientifique.

## L'accès public aux archives

La collecte des archives constitue un élément essentiel de l'engagement de la Bibliothèque nationale de France dans sa nouvelle mission. Celle-ci suppose que soient garanties au public les meilleures conditions de consultation.

### Développer des outils d'accès et imaginer les usages futurs

Le vote de la loi de transposition de la directive européenne conditionne depuis sa préparation en 2003 la consultation par le public des archives collectées par les établissements dépositaires du dépôt légal de la Toile. Cette perspective se précisant, la Bibliothèque nationale de France doit engager un chantier qui revêt d'autres aspects que techniques : organiser les accès à cette documentation en salle de lecture et réunir les outils, les services et les conditions de consultation capables de répondre aux futures attentes des chercheurs confrontés à un matériau d'un genre nouveau. Les archives de la Toile constituent, en effet, un défi du fait de leur masse et de leur architecture.

Pour les chercheurs, l'intérêt scientifique et patrimonial de ce nouvel ensemble documentaire est à la mesure des difficultés qu'il suscite : une archive de site est souvent incomplète (des images, des bases de données ou des pans entiers du site peuvent manquer) et, à l'exception des sites faisant l'objet de dépôts systématiques, la complétude des collections au sens où on l'entend aujourd'hui ne pourra plus être assurée. Des possibilités nouvelles d'exploration et d'analyse se dévoilent : l'étude d'un réseau ou d'un tissu d'objets dans leur contexte d'ensemble, et non plus de manière isolée, présente un intérêt scientifique majeur. L'architecture de la Toile et de ses archives évoque un grand chaos organisé, qui n'est ni une juxtaposition ni une accumulation : ce que les liens qui relient les sites entre eux cimentent est bien une *collection* qui trouvera naturellement sa place dans la Bibliothèque. L'accès simultané à des strates d'existence successives d'un site à partir d'une même interface constituera un mode d'investigation inédit pour l'historien. Enfin, les archives de la Toile donneront accès à des types de documents jamais réunis jusque là. Par exemple, la capture effectuée en novembre 2005 par la BnF de plus d'un million de blogs ne constitue-t-elle pas un formidable matériau pour les chercheurs qui s'intéresseront demain à l'espace public de la Toile ?

Parce que les **outils de visualisation et de consultation des archives de la Toile sont destinés avant tout aux chercheurs**, la Bibliothèque nationale de France souhaite **les associer dès à présent à leur conception**. Ainsi la BnF a-t-elle proposé à la Fondation nationale des Sciences politiques de s'associer dès 2006 à ce projet qui rassemblera un échantillon d'utilisateurs autour d'un corpus de sites (les élections française de 2002 et 2004) afin d'observer leurs usages et d'écouter leurs réactions.

L'étude devrait d'abord permettre de tester auprès de chercheurs les outils de consultation et d'analyse actuellement disponibles à la BnF : outils de localisation et de visualisation, moteur de recherche, moteur d'analyse linguistique. L'évaluation de ces outils doit contribuer à leur perfectionnement (ergonomie, modules fonctionnels) avant leur manipulation par un plus large public.

Elle s'appliquera ensuite à observer plus largement les pratiques des chercheurs confrontés aux archives de la Toile afin de repérer les enjeux de formation et de médiation propres à cette documentation.

La Bibliothèque nationale de France et Sciences-Po réaliseront ces tests auprès de deux types de population ; des chercheurs de premier niveau (étudiants de Sciences Po et lecteurs de la BnF) et des bibliothécaires et des chercheurs confirmés. Ces derniers interviendront à la fois en tant que tuteurs et observateurs du dispositif, apportant d'une part, le regard de l'expert du point de vue de l'intérêt scientifique des contenus et, d'autre part, une vision plus prospective et plus large des situations de recherche rencontrées par les utilisateurs. La BnF s'efforcera d'étendre cette politique de consultation de chercheurs à d'autres champs disciplinaires et de rechercher des partenaires institutionnels susceptibles de s'y associer.

Ce projet, programmé sur deux années et qui vise au développement d'outils sur le moyen et le long termes, n'obère en rien la nécessité de mettre en place un dispositif provisoire de consultation des archives en salle de lecture dès que la loi l'autorisera.

## **L'archivage de la Toile, enjeu de coopération**

La responsabilité confiée à la Bibliothèque nationale de France pour collecter et sauvegarder la mémoire de la Toile n'exclut pas le travail en réseau : bien au contraire, le défi est si considérable qu'il suppose déjà un partage des tâches au niveau national et international, ouvert à d'autres partenaires dans le futur.

### La coopération nationale

La coopération nationale et le partage des domaines ont déjà été prévus par le législateur comme l'indique clairement le texte du projet de loi et comme le prouve cette communication organisée aujourd'hui conjointement par la BnF et l'Ina.

A l'intérieur des frontières de ses compétences, la BnF envisage de recourir à des experts documentaires et à des « veilleurs » de la Toile rattachés à des établissements avec lesquels elle poursuit une collaboration scientifique et institutionnelle : c'est déjà le cas, par exemple, avec la Direction des Archives de France qui s'est fortement mobilisée autour des enjeux de l'archivage électronique et qui contribue par son autorité et son expertise à la collecte des publications officielles.

La méthode mise en place permet à la BnF de proposer une coopération à des organismes spécialisés pour réaliser des collectes ciblées et thématiques autour d'événements dont les manifestations en ligne sont difficiles à identifier, et qui pourront constituer des « zooms » sur des portions de la Toile en complément des collectes automatiques et régulières conduites à grande échelle. C'est un nouveau domaine d'activité potentiel pour le réseau régional des bibliothèques pôles associés.

Il conviendra préalablement de stabiliser au sein de la BnF le modèle fonctionnel, les outils et l'infrastructure informatiques indispensables au bon fonctionnement du dépôt légal de la Toile. Toutefois, la voie est déjà ouverte aux partenariats qui pourront contribuer à améliorer et à consolider le dispositif en cours d'édification.

### La coopération internationale

Convaincues de la pertinence de la collaboration internationale pour servir, à hauteur universelle, la préservation des contenus de la Toile au profit des générations futures, onze bibliothèques ont créé, en juillet 2003, le Consortium International pour la Préservation d'Internet (IIPC). Piloté et coordonné par la Bibliothèque nationale de France, le consortium comprend également la Bibliothèque du Congrès, la British Library et les bibliothèques nationales d'Australie, du Canada, du Danemark, de la Finlande, d'Italie, de Norvège, de Suède ainsi qu'Internet Archive.

Le Consortium s'est fixé plusieurs objectifs. Il s'agit pour l'essentiel :

- d'approfondir une collaboration technique centrée sur l'identification, l'élaboration et la mise en œuvre d'instruments et de procédés propres à identifier, collecter, préserver et rendre accessibles les contenus de la Toile;
- d'établir progressivement un inventaire des collections des contenus de l'Internet, dans le respect des législations de chaque nation et en fidélité aux politiques de sélection propres à chacune ;
- de plaider partout en faveur d'initiatives et de décisions gouvernementales qui favorisent cette belle ambition ;
- d'apporter un appui chaleureux aux pays qui souhaiteront s'engager dans cette voie.

Alors qu'il entre dans sa troisième année d'existence, le Consortium a dressé un premier bilan de son activité lors de la réunion de son comité de pilotage à Washington en octobre dernier. Ce bilan est globalement très positif si l'on confronte les objectifs initiaux aux réalisations d'ores et déjà acquises.

Le Consortium a d'abord conçu une architecture globale pour l'archivage de la Toile qui définit et articule entre elles les fonctions de versement, de stockage, d'indexation, de recherche, d'accès et de gestion.

Cette architecture repose sur le développement, en mode logiciel libre par les membres du Consortium, d'un ensemble d'outils de haute qualité et simples à utiliser. Les outils réalisés en deux ans sont les suivants :

- le robot de collecte à grande échelle Heritrix ;
- un outil d'extraction et de transformation en XML des bases de données pour l'archivage des passerelles documentaires du *web* profond;
- un outil de manipulation des fichiers ARC;
- des outils d'indexation, de recherche et d'accès.

Ces outils sont aujourd'hui disponibles sur le site du Consortium : <http://netpreserve.org/software/downloads.php>

L'action du consortium a également porté sur les questions de standardisation et de normalisation. Les travaux en cours concernent des APIs (*Application Performing Interfaces* : modules qui facilitent l'interopérabilité entre modules fonctionnels), des métadonnées de préservation et le format WARC.

Les travaux du consortium ont enfin permis d'amorcer une réflexion collective sur les enjeux documentaires et les modalités de sélection des collectes thématiques : celles-ci varient d'un pays à l'autre et seront probablement approfondies lorsque les outils de capture seront devenus plus performants.

Le consortium IIPC a démontré sa capacité à réunir autour du projet mondial d'archivage de la Toile les compétences d'un groupe restreint d'institutions qui avaient déjà réalisé des avancées significatives dans la mise en place de modèles fonctionnels. En 2006, ses membres envisageront l'élargissement de leur groupe de travail à d'autres institutions.

# Annexes Ina

## Le cas particulier du *streaming*

Le *streaming* est un procédé technique permettant la diffusion immédiate de vidéo ou de son sur le Web. Contrairement à un téléchargement classique, l'utilisateur n'est pas obligé d'attendre que l'intégralité du contenu ait été téléchargé pour pouvoir le consulter. Cette technologie, qui prend la forme de plusieurs formats et protocoles techniques différents, a permis l'émergence de radios et de télévisions diffusant sur le Web, mais également multiplié les extraits d'émissions sur les sites des chaînes.

Il convient de différencier deux types de *streaming* qui, bien qu'employant les mêmes technologies, présentent des enjeux tout à fait différents dans l'approche de leur archivage : les extraits *streamés* de type vidéo à la demande, et le *streaming* direct de type radio ou télévision sur le Web.

### - Les extraits *streamés*

Les extraits vidéos ou sonores accessibles en *streaming* (bandes-annonces de films, extraits d'émissions, archives de journaux télévisés, etc.) sur la plupart de sites médias sont, en fait, relativement proches des autres contenus Web, accessibles en téléchargement :

- Ils ont une durée finie et déterminée, et sont donc intégralement copiables en un temps fini ;
- Ils sont identiques à chaque consultation, et produiront donc le même contenu à chaque collecte ;
- Ils génèrent un flux unique du serveur vers le client.

On peut donc assimiler ce type de *streaming* à un téléchargement classique. Ces extraits peuvent être pris en charge par les robots de collecte et s'inscrivent dans la même logique d'indexation et de stockage que les autres contenus Web.

### - Le *streaming* direct

Le *streaming* direct présente des caractéristiques tout à fait différentes : il s'agit de flux infinis dont on ne peut connaître par avance la durée (et donc la taille). Leur collecte et leur indexation ne peuvent être prises en charge par le modèle de chaîne technique présenté ici, puisqu'il est impossible de délimiter ces contenus. Ce type de contenu Web s'apparente en fait à une diffusion de type *broadcast* radio ou télévision (bien que techniquement il s'agisse en fait de diffusion *multicast*, voire *unicast*), chaque utilisateur « connecté » à ce flux recevant les mêmes informations au même moment. Des outils de collectes spécifiques ont été développés et s'inscriront dans une chaîne de traitement adaptée, fortement inspirée de celle de la captation radio/télévision déjà pratiquée à l'Ina. La collecte du Web est ainsi partagée en deux processus techniques : l'un, adapté au Web interactif et aux extraits *streamés*, l'autre adapté aux flux radios ou télévisions diffusés à travers le Web.

## La chaîne de traitement expérimentale

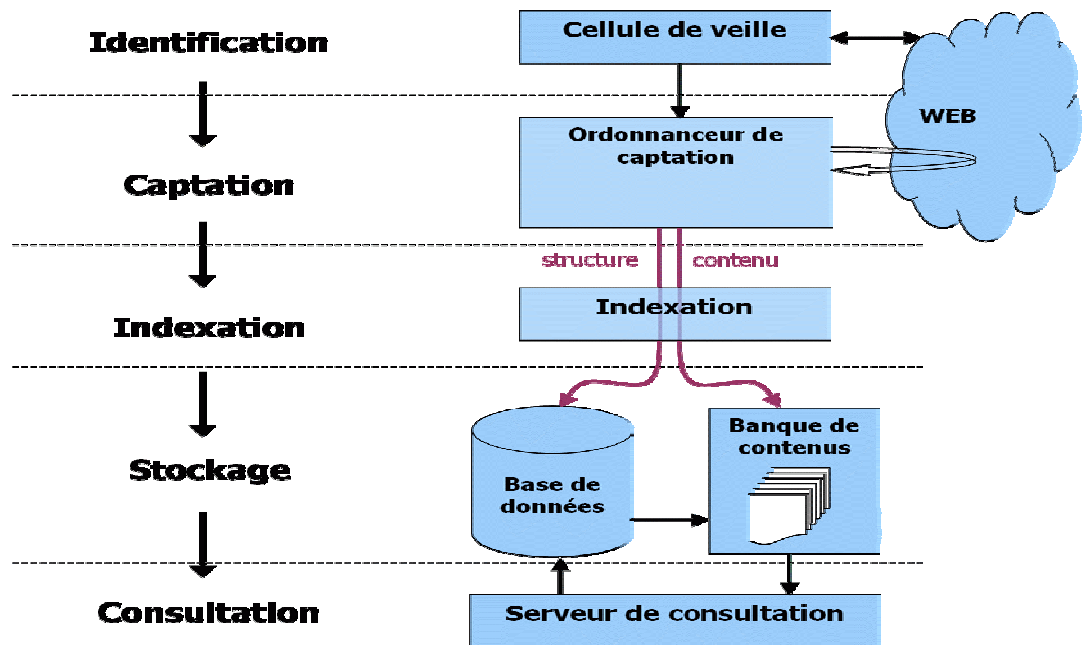


Figure 1 : Chaîne de traitement expérimentale

## Les spécificités du domaine médias

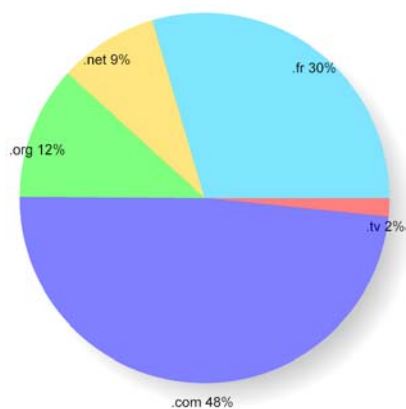


Figure 1 : Répartition des TLD<sup>7</sup> dans le domaine des sites médias<sup>8</sup>

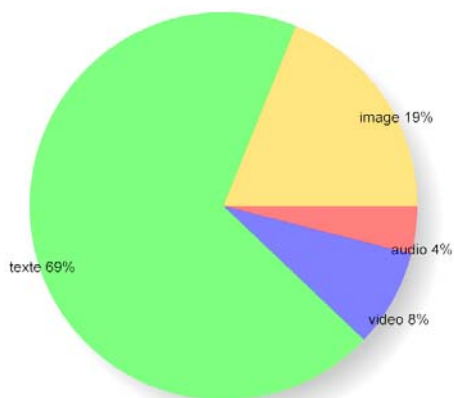


Figure 2 : Répartition (hors flux diffusé en streaming) du nombre de contenus pour un site Tv typique du domaine

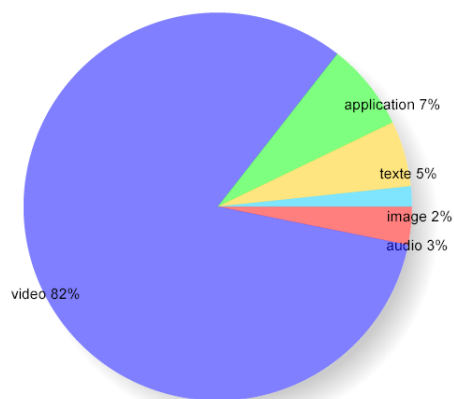


Figure 3 : Répartition du poids des contenus (en terme de stockage) pour un site Tv typique du domaine

<sup>7</sup> TLD: Top Level Domains, noms de domaine au sommet de la hiérarchie.

<sup>8</sup> Ces mesures ont été réalisées lors de crawls prospectifs du domaine.

# Annexes BnF

## Le traitement documentaire des archives

Le traitement des données issues des collectes de la Toile impose la mise en place d'une architecture et d'une chaîne de traitement technique et intellectuelle qui dépassent la seule gestion des archives de sites, rejoignant ici la problématique plus générale du traitement et de la conservation des documents numériques à la Bibliothèque nationale de France. C'est donc dans le cadre de la constitution d'un système global de stockage et de préservation pérenne des documents numériques que des solutions techniques adaptées aux collections de la Toile sont définies.

Les volumes sont si importants que, là encore, seule une approche largement automatisée, n'excluant cependant pas des traitements unitaires ponctuels, peut être retenue. Cette approche induit notamment l'utilisation de métadonnées qui seront pour la plupart générées automatiquement pendant et à l'issue de la collecte. Ces métadonnées contiendront des informations de plusieurs types :

- **métadonnées descriptives** (qui apportent des informations sur le contenu, utiles à l'indexation, à la recherche et à la localisation) ;
- **métadonnées de préservation** (données techniques sur les formats et sur les règles et les durées de conservation) ;
- **métadonnées de gestion des droits** (pour l'organisation des accès et plus largement des conditions d'utilisation).

Le traitement des sites s'organise autour des principales étapes suivantes :

- **la réception et la vérification** (de l'intégrité et de la conformité des contenus, de leurs volumes, de leurs formats) pendant et à l'issue des collectes ;
- **leur validation** (contrôle qualité) ;
- **leur versement dans le système de stockage de la Bibliothèque.**

Le traitement pourra varier selon qu'il s'agit de collectes larges, ciblées ou de dépôts : on peut en effet imaginer que certaines collections et certains dépôts thématiques fassent l'objet d'un effort de valorisation et de signalement particulier ou de précautions renforcées en termes de conservation. En particulier, la BnF veillera à ce que les notices bibliographiques des ressources déjà conservées par elle sous une forme analogique et ayant changé de support à l'issue d'une migration vers Internet fassent mention de ce changement et donnent les indications utiles pour retrouver la suite de la collection sur ce nouveau support.

Ainsi le traitement des archives de la Toile ne devrait pas profondément différer de celui d'autres ressources du domaine numérique. Les spécificités techniques des archives de sites et les fonctionnalités requises pour leur collecte sont néanmoins prises en compte, comme l'illustre la définition d'un nouveau format en cours de normalisation.

Le Consortium IIPC a en effet adopté le format ARC\* (ARChive file format) comme format de stockage et d'échange des archives de la Toile. Le format ARC facilite la collecte, le stockage et la gestion d'un très grand nombre de petits fichiers (la manipulation de gros volumes de fichiers est un processus excessivement long et difficile pour un ordinateur), l'extraction d'un fichier lors de la consultation, l'échange de fichiers entre institutions de mémoire et leur préservation sur le long terme.

Le Consortium IIPC travaille à l'extension du format ARC pour le mettre en mesure de répondre à des besoins plus nombreux et, une fois normalisé par l'ISO (Organisation internationale de Normalisation), de devenir le format universel des archives de la Toile. La BnF participe aux travaux du Consortium depuis sa création en juillet 2003.

---

\* cf. glossaire p.25

## Glossaire du dépôt légal de la Toile à la BnF

Cette liste apporte quelques précisions et définitions sur des termes techniques fréquemment utilisés à propos de l'archivage de la Toile, ainsi que l'explication et la traduction de termes issus de la terminologie anglo-saxonne. La plupart de ces définitions sont extraites du glossaire établi par AFCEE/EDIFRANCE, Observatoire du commerce et des échanges électronique (accessible en ligne depuis le site du Forum des droits sur Internet : <http://www.foruminternet.org/glossaire/>) et des glossaires établis par la Délégation Générale à la Langue Française (disponibles à l'adresse : <http://www.culture.gouv.fr/culture/dgjf> ).

**ARC** : format de stockage des archives d'Internet. Il permet d'enregistrer le résultat d'une collecte de manière agrégée, c'est-à-dire sous la forme d'un fichier dont la taille varie de 100 à 600 méga-octets. Un fichier ARC se compose d'enregistrements. Chaque enregistrement commence par une en-tête contenant des informations relatives au contexte technique de la collecte et qui sont issues du protocole d'échange entre le robot de collecte et le serveur hébergeant le fichier à collecter qui est identifié par un URI (*Uniform Resource Identifier*). A la suite de cette en-tête se trouve le contenu même du fichier collecté (un fichier HTML, une image, un enregistrement sonore ou vidéo, une feuille de style, une animation, un document issu d'un logiciel de traitement de texte...un élément composant une page web). Tous les types de fichiers (qu'ils soient sous forme de texte ou sous forme binaire, lisible ou illisible à l'œil nu) composant des pages trouvées au fur et à mesure du parcours du robot de collecte sont ainsi intégrés dans différents enregistrements ARC.

**Bit** : unité élémentaire d'information codée sous la forme de 0 ou 1. Il faut huit bits pour former un octet. Un octet permet de représenter un caractère. Voici les unités de mesure les plus courantes :

- 1 kilo-octet (ko ou Ko) =  $2^{10}$  octets = 1 024 octets,
- 1 méga-octet (Mo) =  $2^{20}$  octets = 1 024 ko = 1 048 576 octets,
- 1 giga-octet (Go) =  $2^{30}$  octets = 1 024 Mo = 1 073 741 824 octets,
- 1 téra-octet (To) =  $2^{40}$  octets = 1 024 Go = 1 099 511 627 776 octets,
- 1 péta-octet (Po) =  $2^{50}$  octets = 1 024 To = 1 125 899 906 842 624 octets

**Bloc-notes (*weblog, web blog, blog*)** : site sur la Toile, souvent personnel, présentant en ordre chronologique de courts articles ou notes, généralement accompagnés de liens vers d'autres sites.

La publication de ces notes est généralement facilitée par l'emploi d'un logiciel spécialisé qui met en forme le texte et les illustrations, construit des archives, offre des moyens de recherche et accueille les commentaires d'autres internautes (définition du Journal Officiel du 20 mai 2005).

**Capture (*crawl*)** : collecte automatisée de sites au moyen d'un robot. On distingue les *crawls* larges des *crawls* ciblés (*focused crawls*) qui permettent une capture en profondeur des sites à collecter.

**Crawler** : robot de capture chargé de la collecte automatisée de sites en vue de les archiver.

**Domaine (nom de)** : ensemble d'adresses faisant l'objet d'une gestion commune. Système de nommage à l'intérieur duquel est garantie l'unicité des noms. Ce système est hiérarchique et permet la définition de sous-domaine(s) d'un domaine existant. Le nom de domaine est composé d'un label et d'un suffixe. Il existe deux catégories de noms de domaines :

- les domaines génériques (gTLD, *Global Top Level Domains*) gérés par l'ICANN (*Internet Corporation for Assigned Names and Numbers*) par l'intermédiaire de centres d'enregistrements (*registrars*) ex : .com, .org, .net, .info, etc.
- les domaines géographiques (ccTLD, *Country Code Top Level Domains*) ex : .fr, .uk, .de, etc.

**DNS (Domain Name Server)** : service dont le rôle est de convertir les noms de domaines lisibles par l'homme par les adresses numériques (adresses IP) auxquelles ils correspondent permettant aux machines de communiquer entre elles.  
Exemple : www.culture.gouv.fr = 143.126.211.220.

**Forum (newsgroup)** : service permettant discussions et échanges sur un thème donné : chaque utilisateur peut lire à tout moment les interventions de tous les autres et apporter sa propre contribution. Par extension, on désigne également par ce terme les systèmes de discussions télématiques, qui offrent généralement un service de téléchargement (connus en anglais sous le nom de BBS, *Bulletin Board System*).

**Graines (seeds)** : URL de départ données au robot pour effectuer une collecte.

**Hypertexte** : mode d'organisation des documents numériques caractérisé par l'existence de liens dynamiques entre ses différentes sections. Sur la Toile, des mots ou des groupes de mots soulignés ou des images indiquent les liens hypertextes sur lesquels on clique à l'aide de la souris pour accéder à une autre partie d'un document ou à un autre fichier.

**IP, Adresse IP (Internet Protocol)** : identifiant standard d'un ordinateur connecté à Internet. Une adresse IP v4 est constituée de quatre nombres séparés par des points, et chaque nombre est inférieur à 256, par exemple, 192.200.44.69. Une adresse IP peut être convertie en un nom de domaine, forme plus facilement lisible et mémorisable.

**Instantané (snapshot)** : une photographie instantanée de la surface de la toile à un instant donné.

**Lien hypertexte (link)** : mot, groupe de mots ou image permettant de passer d'une page à l'autre à l'intérieur d'un même site (lien interne) ou sur un autre site (lien externe ou lien sortant).

**Métadonnées** : ensemble d'informations permettant de définir ou de décrire une ressource analogique ou numérique pour en assurer la production, la gestion, la diffusion et/ou la préservation.

**Moteur de recherche (search engine)** : service qui collecte des informations sur les ressources disponibles sur la Toile en vue de les rendre accessibles à partir de mots clés.

**Portail (portal)** : site répertoriant de nombreux autres sites dans un domaine précis et destiné à servir de point d'entrée du parcours d'un internaute sur la Toile.

**Site (web site)** : ensemble de documents et d'applications constituant un ensemble navigable et placés sous une même autorité et accessibles par la Toile à partir d'une même adresse URL.

**URL (*Universal Resource Locator*)** : chaîne de caractères désignant la localisation d'une ressource. **Sur la Toile**, une URL est composée :

- du langage permettant d'accéder à la ressource (c'est le protocole HTTP, *HyperText Transfer Protocol* qui permet à un logiciel de navigation d'accéder à une page web),
- de la machine hébergeant la ressource,
- et du chemin d'accès à la ressource.

Exemple : <http://www.culture.gouv.fr/culture/min/index-min.htm>

**WARC (*Web Archive file format*)** : évolution du format ARC, en cours de normalisation.

**Web invisible ou profond (*Invisible web, deep web, hidden web*)** : Le "web invisible" désigne la partie de la Toile non accessible aux moteurs de recherche.

Chris Sherman et Gary Price, spécialistes américains des outils de recherche, distinguent quatre catégories de "web invisible" :

- le web opaque : les ressources qui pourraient être indexées par les moteurs de recherche mais qui ne le sont pas à cause, notamment, de la limitation du nombre de pages d'un site indexées, de la fréquence d'indexation, des liens absents vers certaines pages qui ne permettent pas au moteur de les trouver ;
- le web privé : les ressources rendues volontairement inaccessibles par les administrateurs des sites ;
- le web propriétaire : les ressources accessibles uniquement aux personnes qui s'identifient ;
- le web vraiment invisible : les ressources qui ne peuvent pas être indexées pour des raisons techniques (ex : parce que leur format n'est pas reconnu par le moteur, des pages générées seulement dynamiquement, lors d'une requête, etc.).

**Wiki** : site web dynamique permettant aux internautes d'en modifier les pages de façon simple et rapide sans système de contrôle ou de validation. Les wikis sont utilisés par des communautés, professionnelles ou non, pour gérer des projets collectifs. L'exemple le plus connu est l'encyclopédie collaborative Wikipédia.