Dossier documentaire

L’archivage du Web

Thomas Chaimbault

école nationale supérieure des sciences de l'information et des bibliothèques

Sommaire

Introduction

L'apparition du web aura créé un paradoxe sans précédent : jamais en effet il n'y a eu autant d'information créée, disponible rapidement et simplement pour le plus grand nombre, mais jamais également n'y a-t-il eut autant de perte. L'immense facilité de publication, aujourd'hui amplifiée par les outils de publications personnelles ne nécessitant plus de connaître les codes informatiques, se voit contrebalancé par le caractère hautement éphémère de ce nouveau média.

A l'heure où l'internet est en passe de devenir le premier média, recevant des recettes –notamment publicitaires toujours plus importantes-, où les événements publiques prennent une place de plus en plus importante sur la Toile comme l'ont montré encore récemment les élections présidentielle et législative françaises, les utilisateurs s'emparent du net et créent le contenu même des sites qu'ils visitent, le besoin de l'archivage du web n'apparaît plus temps comme une question que comme une nécessité.

Ainsi après s'être penché sur l'intérêt et les objectifs d'un archivage de la Toile, soulignant les enjeux d'un tel projet, va-t-on tenter d'exposer des considérations techniques et organisationnelles nécessaires à sa mise en place et proposer un tour d'horizon des réalisations, en terme de stratégies, initiatives individuelles ou projets collaboratifs.

Partie 1 : Pourquoi archiver le web ?

1. Le web occupe une place de plus en plus importante dans les pratiques informationnelles

1.1. Explosion du web

Le web apparaît de nos jours comme la plus importante base d'information qui ait jamais existé. Dans une étude sur les flux d'information sur internet en 2002, Peter Lyman and Hal R. Varian, professeurs à la *School of Information Management and Systems* de l'université de Berkeley en Californie ont estimé qu'en 2002, le *World Wide Web* contenait environ 170 térabytes d'information à sa surface; soit dix-sept fois le volume des collections imprimées de la bibliothèque du Congrès[1].

Au cours de la même étude, les chercheurs ont souligné l'extraordinaire croissance du web, qui ajouterait plus de sept millions de pages tous les jours tandis que parallèlement, son contenu disparaitrait : le temps de vie moyen d'une page ne serait que de quarante-quatre jours ![2] et les usagers d'internet sont devenus familiers de la fameuse erreur 404 « object not found » ou « la page demandée n'existe pas » qui désigne une erreur de localisation. Dès lors, sans politique d'archivage, le risque de perdre des données est important.

Internet se caractérise par sa masse, sa taille croissant de manière exponentielle[3]. Dès lors ; le volume des publications apparaît sans précédent et la recherche d'un archivage exhaustif semble impossible, y compris si l'on restreint le processus d'archivage à une portion de l'internet (uniquement les

---

[1] LYMAN, Peter, VARIAN Hal. How much information ? 2003 [en ligne]. In *School of information management and systems*. Consulté le 21 février 2008. Disponible sur : http://www.sims.berkeley.edu/research/projects/how-much-info

[2] *ibidem*

[3] Voir notamment :
ERTZCHEID, Olivier. Question de taille… in *Affordance.info* [en ligne]. Mars 2007 [consulté le 15 février 2008]. Disponible sur : http://affordance.typepad.com/mon_weblog/2007/03/question_de_tai.html

domaines en .fr par exemple). Il existerait ainsi quelque 128 millions de sites dans le monde, en août 2007[4], tandis que l'AFNIC a enregistré 890032 noms de domaine sous .fr[5], restriction qui ne permet d'identifier qu'une partie du web français. De fait, les premières collectes du domaine français réalisées par la BnF du 15 décembre 2004 au 30 janvier 2005 aboutit à la réception de 3 téraoctets de données, représentant un total de plus de 118 millions de fichiers identifiés par une adresse URL.

1.2. Le web comme support d'information

De plus en plus d'informations qui étaient auparavant stockées sur des supports matériels se retrouvent sur des pages web, voire uniquement sur ces supports. L'explosion des blogs en 2006, marque ainsi l'engouement des internautes pour la publication en ligne, les chiffres du e-commerce et de la publicité en ligne surtout dépassent les résultats des autres supports publicitaires (papier, télévision…). Les termes d'e-administration, d'édition en ligne, de bibliothèques numériques, d'enseignement à distance, d'arts numériques envahissent le vocabulaire soulignant que de nombreuses activités se sont ainsi déplacées vers la Toile et révélant, au delà d'une seule mutation technique, des processus sociaux inédits et innovants. En France, Médiamétrie évalue le nombre d'internaute à plus de 31 millions soit 59.4 % de la population, en décembre 2007, dont une majorité serait connectée en haut débit[6].

Dans bien des cas, la situation reste hybride et il ne s'agit encore que de la simple transposition à partir du support papier. C'est le cas par exemple des publications scientifiques dont les éditions électroniques se multiplient obligeant parfois les éditeurs à proposer plusieurs modalités tarifaires ; c'est le cas de plus en plus de la presse qui trouve par ce biais l'occasion de rencontrer de nouveaux lecteurs et de réagir plus vivement et rapidement à l'actualité. C'est encore le cas des débats politiques, ainsi qu'en attestent les campagnes

---

[4] *Web survey* [en ligne]. Netcraft LTD, août 2007 [consulté le 15 février 2008]. Disponible sur : http://news.netcraft.com/archives/2007/08/06/august_2007_web_server_survey.html

[5] *Afnic* [en ligne]. Association Française pour le Nommage Internet en Coopération (AFNIC), 2003, mise à jour le 21 mai 2007 [consulté le 15 février 2008]. Disponible sur : http://www.afnic.fr/actu/stats

[6] Médiamétrie/NetRatings. *L'audience de l'internet en France : décembre 2007* [en ligne]. Consulté le 15 février 2008. Disponible sur : http://www.mediametrie.fr/resultats.php?rubrique=net&resultat_id=504

électorales récentes, dont une part non négligeable se déroule sur le web, le gouvernement français actuel comportant ainsi nombre de politiques blogueurs.

1.3. Une modification des pratiques

Dès lors, des processus sociaux sont à l'œuvre. Les individus qui prennent ainsi possession de la Toile inventent de nouvelles proximités, redéfinissent des frontières et les modalités d'intervention dans l'espace public. Les innovations rassemblées sous le terme de web 2.0 mettant en valeur la participation de l'usager et l'importance des réseaux sociaux soulignent plus que jamais le passage d'Internet du simple statut de réservoir d'information et de service à celui de communautés d'échanges que l'on intègre pour se raconter, se rencontrer, créer des liens. On trouve dès lors dans le web un abondant contenu –texte, multimédia- qu'on ne trouve nulle part ailleurs comme on trouve des pratiques singulières qu'on ne peut ignorer quand on s'interroge sur la manière de conserver une trace des évolutions que connaissent aujourd'hui nos sociétés.

Parallèlement, le développement de ces contenus et leur circulation dans des réseaux virtuels bouleverse profondément l'économie de la mémoire. Il est ainsi devenu presque un lieu commun que de considérer le web comme un média éphémère, volatil. Non seulement 70% des pages web ont une durée de vie inférieure à quatre mois[7] –certains sites sont mis à jour très souvent, d'autres disparaissent ou changent de fournisseurs, d'hébergeurs…- mais la structure hypertextuelle des documents eux-mêmes répercute cette précarité sur l'ensemble du réseau. Une volatilité qui, cependant, semble loin d'être aussi évidente, la circulation des données numériques exigeant des procédures d'inscription et de multiplication de ces traces : métadonnées, sites miroirs, fichiers partagés, caches… Le fait qu'une page ne soit plus signalée par les moteurs de recherche ne signifie pas qu'elle n'existe plus. A ce titre, Louise Merzeau préfère parler d'internet comme d'un « espace de réverbération, où le signal ne disparaît que progressivement, par un phénomène « d'échos

[7] HOOG, Emmanuel. Internet a-t-il une mémoire. *Le Monde* [en ligne], 17 août 2002 [consulté le 15 février 2008] Disponible sur : http://www.ac-versailles.fr/pedagogi/ses/vie-ses/hodebas/hoog1.htm

successifs qui vont en s'atténuant » »[8]. Cette permanence n'invalide pas une nécessaire politique d'archivage qui offre, elle, la garantie de l'institution à une rétention jusqu'alors soumise aux seules lois de l'innovation technique et de la concurrence.

2. Préserver le patrimoine numérique et culturel

2.1. Un patrimoine particulièrement fragile

Le patrimoine numérique est un patrimoine particulièrement fragile à plusieurs points de vue. Il semble trop récent pour avoir acquis une légitimité documentaire, il est au cœur des changements techniques, il appartient à tous et personne ne semble soucieux de le préserver.

D'un point de vue culturel, il est important de savoir prendre suffisamment de recul sur ses propres pratiques et la production documentaire car si on n'y prend garde, un certain nombre de données disparaitront faute d'avoir pu ou su reconnaître à temps leur valeur historique. Avec le développement rapide des technologies, il risque d'être moins aisé de retrouver les informations contenues sur une disquette 5'1/4 à l'heure des services en ligne que de redécouvrir des archives papiers oubliées au fond d'un grenier ou d'une salle dédiée.

D'un point de vue technologique justement, il convient de faire attention à permettre une continuité de lecture des documents, c'est-à-dire préserver tant les matériels que les logiciels et faire face aux progrès technologiques. Dans cette optique, une archive de l'internet doit résoudre les problèmes techniques concernant chacun des documents à lire en plus de ses problèmes propres.

D'un point de vue économique, le maintien d'archives peut être relativement difficile. Ainsi, la question de la responsabilité des documents qui apparaissent sur le web ne semble pas vraiment résolue, et leur préservation apparaît d'autant moins importante que les documents préservés peuvent ne révéler leur

[8] MERZEAU, Louise. Web en stock [en ligne]. in *Cahiers de médiologie*, septembre 2003, N°16, [consulté le 15 février 2008]. Disponible sur : http://www.merzeau.net/txt/memoire/webenstock.html

intérêt que des dizaines d'années plus tard, pour une communauté réduite de chercheurs. Le maintien d'archives papiers rencontre le même problème ce qui tend à l'établissement d'archives spécialisées. Mais un archivage de l'internet requiert au contraire beaucoup d'investissements initiaux en terme techniques, de recherche, de développement ou de formation.

2.2. Transmettre le savoir

S'il convient en premier lieu de préserver les documents numériques, il est tout autant nécessaire de souligner l'importance de l'accès aux documents ce qui n'est pas sans poser d'autres questions tant économiques que politiques.

Il s'agit de préserver l'accès aux informations afin de répondre à des exigences multiples : assurer une transmission des informations aux générations futures ; créer de nouveaux produits informationnels en accord avec les codes de propriété intellectuelle ; permettre la libre circulation des informations ; assurer un accès différent de celui des sociétés commerciales.

Dans un article intitulé « La conservation des publications électroniques et du dépôt légal », Catherine Lupovici alors directrice du département de la bibliothèque numérique, direction des services et des réseaux rappelle ainsi l'importance de la mission de conservation des documents pour la transmission du savoir :

> « *Le patrimoine des publications est le résultat des acquisitions faites par les établissements patrimoniaux sur la base d'une sélection documentaire et des collections constituées par le dépôt légal. Ce patrimoine a fait l'objet d'une conservation passive et active pour les documents identifiés comme en danger. Il faut bien comprendre que ces actions réparties sur un nombre important d'institutions avec des doublons et des recouvrements de politiques documentaires ont permis la conservation de ce patrimoine en dépit des échecs de conservation et des catastrophes qui sont survenues au fil de l'histoire. La duplication des efforts a seule permis la transmission d'une quantité importante du patrimoine mondial*

des publications. Ces leçons du passé ne doivent pas être oubliées à l'ère de la publication numérique, alors que les incertitudes sur la possibilité de la maîtrise de la conservation des nouvelles publications et l'accroissement considérable de leur quantité peut nous pousser à vouloir réserver des moyens insuffisants et que nous avons encore du mal à évaluer à une sélection draconienne initiale des objets qui seront conservés créant ainsi dès le départ des trous énormes dans notre patrimoine futur.[9]

3. Des raisons politiques et patrimoniales : élargir le champs du dépôt légal

3.1. Qu'est-ce que le dépôt légal ?

Institué en France par François 1er en 1537 (ordonnance de Montpellier), visant à l'exhaustivité, le dépôt légal a initialement pour objet de recenser tous les documents imprimés, graphiques et photographiques. Il oblige les éditeurs à déposer plusieurs copies de leurs publications auprès de bibliothèques du pays dans lequel ils publient. En France, par exemple, il s'effectue auprès de la Bibliothèque nationale de France ou des organismes dépositaires compétents en province.

En France, l'article L. 131-1 du Code du patrimoine définit ainsi le dépôt légal :

« *Le dépôt légal est organisé en vue de permettre :*

a) La collecte et la conservation des documents mentionnés à l'article L. 131-2 ;

b) La constitution et la diffusion de bibliographies nationales ;

La consultation des documents mentionnés à l'article L. 131-2, sous réserve des secrets protégés par la loi, dans des conditions conforme à la législation sur la propriété intellectuelle et compatible avec leur conservation. »

[9] LUPOVICI, Catherine. *La conservation des publications électroniques et du dépôt légal* [en ligne]. In Information for All Programme (IFAP). S.l. : UNESCO, 2007. Consulté le 26 février 2008. Disponible sur : http://portal.unesco.org/ci/en/ev.php-URL_ID=24441&URL_DO=DO_TOPIC&URL_SECTION=201.html

Le principe du dépôt légal est établi dans des conventions internationales et dans la législation de nombreux pays. Le dépôt légal en France permet notamment :

- la collecte et la conservation des documents de toute nature publiés, produits ou diffusés en France, afin de constituer une collection de référence, patrimoine irremplaçable pour la collectivité nationale dont il contribue à préserver la mémoire,
- la constitution et la diffusion de la Bibliographie nationale française,
- la consultation des documents dans les salles de la bibliothèque, sous réserve des secrets protégés par la loi, dans des conditions conformes à la législation sur la propriété intellectuelle et compatibles avec leur conservation.[10]

3.2. Dépôt légal et documents numériques

De plus en plus de documents sont publiés sous format numérique et ces documents doivent également être collectés et conservés pour assurer un recensement complet de la production culturelle d'un pays. Il s'agit donc à la fois de protéger le patrimoine numérique contre sa perte, d'en assurer la diffusion, et de protéger de la même façon le patrimoine culturel des pays.

Si les ont besoin de cadres juridiques et institutionnels appropriés pour assurer la protection de leur patrimoine numérique, la législation en matière d'archives et de dépôt légal doit être étendue. C'est ce que précise dans son article 8 le « projet de charte sur la conservation du patrimoine numérique » de l'UNESCO :

> « *Élément clé de la politique nationale de conservation, la législation en matière d'archives et de dépôt légal ou volontaire dans des bibliothèques, archives, musées et autres dépôts publics doit être étendue au patrimoine numérique.*

[10] Bibliothèque nationale de France. Qu'est-ce que le dépôt légal [en ligne]. In *Bibliothèque nationale de France*. Consulté le 21 février 2008. Disponible sur : http://www.bnf.fr/PAGES/infopro/depotleg/dl-france.htm

L'accès aux documents du patrimoine numérique en dépôt légal, doit être assuré, dans le respect de restrictions raisonnables, sans que cela nuise à leur exploitation normale.

Les cadres juridiques et pratiques protégeant l'authenticité sont indispensables pour éviter la manipulation ou l'altération volontaire du patrimoine numérique. Ils exigent que le contenu, la fonctionnalité des fichiers et la documentation soient conservés dans la mesure nécessaire pour garantir l'authenticité des documents. »[11]

Dans de nombreux pays, la législation du dépôt légal a évolué afin de prendre en compte ce nouveau type de documents. En France, la loi sur les Droit d'auteurs et droits voisins dans la société de l'information (DADVSI) du 1er août 2006, a apporté dans son titre IV quelques modifications au code du patrimoine. Le régime du dépôt des bases de données et logiciels a été revu et le dépôt est étendu aux sites Web (Code du patrimoine, art. L.131-2) :

« *Les logiciels et les bases de données sont soumis à l'obligation de dépôt légal dès lors qu'ils sont mis à disposition d'un public par la diffusion d'un support matériel, quelle que soit la nature de ce support. 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.* »

Le principe fondamental qui régit cette nouvelle législation est que l'héritage culturel et scientifique d'un pays se compose aussi bien des publications numériques que papiers.

Aujourd'hui, le dépôt légal concerne non seulement le livre et le périodique, mais aussi la gravure, la photographie, les films, la télévision, les disques audio et

[11] Organisation des nations unies pour l'éducation, la science et la culture. *Charte sur la conservation du patrimoine numérique* [en ligne].Paris : UNESCO, 2003. Consulté le 23 février 2008. Disponible sur : http://portal.unesco.org/ci/en/ev.php-URL_ID=13367&URL_DO=DO_TOPIC&URL_SECTION=201.html

vidéo, les bases de données et logiciels, la production radiodiffusée et télédiffusée et certaines pages d'internet.

En l'absence de législation, des schémas de dépôt volontaires ont été développé dans un certain nombre de pays[12]. La conférence des bibliothèques nationales européennes et la fédération européenne des éditeurs a mis en place un modèle pour soutenir les dépôts locaux de ce genre. (CENL/FEP 2005)[13].

[12] British Library. *Code of Practice for the Voluntary Deposit of Non-Print Publications* [en ligne]. Londres : British Library, 2000. Consulté le 10 mars 2008. Disponible sur : http://www.bl.uk/aboutus/stratpolprog/legaldep/voluntarydeposit/

[13] Conference of European National Librarians, Federation of European Publishers. *Statement on the Development and Establishment of Voluntary Deposit Schemes for Electronic Publications*[en ligne]. Annual Conference of European National Librarians, Luxembourg, September 28-30, 2005. Disponible sur : http://www.nlib.ee/cenl/docs/05-11CENLFEP_Draft_Statement050822_02.pdf

Partie 2 : Comment archiver le web ?

1. Archiver l'ensemble du web ?

1.1. Considérations générales

Howard Besser a identifié cinq problèmes techniques en termes de préservation numérique[14] :

1. Le problème majeur est l'installation et la maintenance d'une infrastructure et de l'expertise nécessaire pour rendre les objets numériques lisibles.
2. Le décodage de logiciels de compressions ou de services de protections des pages web.
3. La préservation du contexte qui rend aux informations leur signification, tels que les liens hypertextuels.
4. La définition de standard de bonnes pratiques et de politiques d'acquisition qui définissent les limites de la collection, sa provenance et son authenticité.
5. Des problèmes de migration à propos de la façon dont l'expérience des pages web change lorsqu'elles sont transférées sur de nouveaux appareils.

1.2. Définir les limites de la collecte

1.2.1. La page web, objet complexe

D'un point de vue technique, l'objet de la collecte n'est pas si précis qu'il paraît. En effet, ce qui détermine la construction d'une archive du web est son objet, à savoir une « page web ». D'un point de vue usager, la page web est l'ensemble des données qui surviennent lorsqu'on entre une adresse URL dans un navigateur, une approche incomplète puisqu'une archive doit également prendre

[14] BESSER, Howard. Digital Longevity. *In* SITTS, Maxine. *Handbook for Digital Projects: A Management Tool for Preservation*. Andover, Mass.: Northeast Document Conservation Center. 2000. cité dans LYMAN, Peter. Archiving the world wide web [en ligne]. In Council on library and Information resources. *Building a National Strategy for Preservation: Issues in Digital Media Archiving.* Berkeley, 2002. Consulté le 26 février 2008. Disponible sur : http://www.clir.org/pubs/reports/pub106/Web.html

en compte tout ce qui se trouve autour du document et lui permet d'être signifiant ; tout ce qui forme son contexte.

Une question qui se rapproche de la notion de dé-territorialité[15], voulant que les documents à collecter se rapportent plus à des chemins de navigation qu'à des supports d'inscription. Le document hypertextuel n'est jamais réellement fini. Si l'archivage retient pour unité intellectuelle le site, la limite ainsi définie demeure floue puisque chaque site renverra vers d'autres sites qui eux-mêmes proposeront des liens d'autres sites rendant impossible de définir le périmètre ou le volume de la collecte. Il ne présente pas même d'unité propre puisqu'il se segmente en éléments plus ou moins autonomes (boutons, images, bandeaux, textes…) que l'on doit traiter séparément.

Si l'on considère ainsi qu'en moyenne, une page web contient une quinzaine de liens vers d'autres pages, et environ cinq objets d'origines diverses (sons, images, code, films…)[16], la description technique d'une page demeure ambiguë et floue

1.2.2. Le web profond

Définir les objets de la collecte renvoie également aux pages que l'on peut indexer. La plupart des pages web auxquelles un robot d'archivage ou un usager peut facilement accéder font partie de ce qu'on appelle le web « visible ». Par opposition, le « web invisible » (*deep web*, *hidden web*) désigne cette partie du web non accessible aux moteurs de recherche classiques. Il comprend des bases, banques de données et bibliothèques en ligne gratuites ou payantes...[17] Difficilement accessible, ce web comprend des pages protégées par un mot de passe, des pages orphelines, des documents ou des sites trop volumineux, des formats mal reconnus ou des pages générées dynamiquement, c'est-à-dire

[15] MERZEAU, Louise. Web en stock [en ligne]. in *Cahiers de médiologie*, septembre 2003, N°16, [consulté le 15 février 2008]. Disponible sur : http://www.merzeau.net/txt/memoire/webenstock.html
[16] LYMAN, Peter. Archiving the world wide web [en ligne]. In Council on library and Information resources. *Building a National Strategy for Preservation: Issues in Digital Media Archiving*. Berkeley, 2002. Consulté le 26 février 2008. Disponible sur : http://www.clir.org/pubs/reports/pub106/Web.html
[17] ASSELIN, Christophe. Web invisible, web caché, web profond [en ligne]. In *IntelligenceCenter.com*. Comsulté le 26 février 2008. Disponible sur : http://c.asselin.free.fr/french/invisible_web.htm

uniquement en réponse à une requête sur leur moteur interne. Il n'existe pas alors d'URL (adresse) statique de ces pages.[18]

En 2001, l'étude « *The Deep Web: Surfacing Hidden Value* », de Michael K. Bergman, considère que le web profond serait de 400 à 550 fois plus volumineuse que le web de surface (web visible) [19]. Si l'étude date de 2001, les proportions restent valables (en estimation basse) compte tenu des taux de croissance très élevés du volume du web profond.

L'archivage de la Toile est alors restreint à la « surface » des sites. Un problème qui se pose avec d'autant plus d'acuité dans le cas des sites de communication audiovisuelle réalisés en utilisant des technologies de diffusion élaborée (*streaming*, Flash, XUL, etc.) dont il importe de d'assurer un archivage adapté qui permettra d'appréhender chaque information dans son contexte visuel, sonore et interactif original.

1.2.3. Considérations légales et juridiques

Il ne faut pas oublier non plus que ces pages enfouies dans le web profond peuvent être protégées par le droit d'auteur en ce qui concerne des articles disponibles dans des bases de données par exemple ou par le droit privé notamment en ce qui concerne les informations personnelles contenues dans des bases commerciales ou de réseaux sociaux, eux-mêmes pas toujours très au clair sur ces questions de droit.

A propos de droit, chaque élément de la page –qu'il s'agisse de texte, d'images, de son, de graphiques, de vidéo, de code… est potentiellement soumis au code de la propriété intellectuelle, quand bien même la mention de l 'auteur est rarement présente et encore plus difficilement trouvable. Globalement, les pages web se voient protégées par le code de la propriété intellectuelle et il ne devrait pas être possible de les reproduire sans l'accord de leurs auteurs. Une question

---

[18] DIGIMIND. *Découvrir et exploiter le web invisible pour la veille stratégique* [en ligne]. Consulté le 26 février 2008. Disponible sur : http://www.digimind.fr/publications/white-papers/222-decouvrir-et-exploiter-le-web-invisible-pour-la-veille-strategique-2.htm

[19] BERGMAN, Mickael K. The 'Deep' Web: Surfacing Hidden Value[en ligne]. In *BrightPlanet*. Consulté le 26 février 2008. Disponible sur : http://www.brightplanet.com/resources/details/deepweb.html

d'autant plus délicate que le web en soi dépend de plusieurs juridictions puisque accessible depuis tous les pays.[20]

1.3. Récolter les métadonnées

Parmi les limites de la collecte des documents, il convient de ne pas oublier de récolter les informations sur les pages web, à savoir ce qu'on appelle les métadonnées des documents. Une métadonnée est littéralement une donnée sur une donnée ; plus précisément, c'est un ensemble structuré d'informations décrivant une ressource quelconque[21].

Le recueil de ces métadonnées doit pouvoir fournir des données sur le contexte technique et historique de la collecte d'une part et du document d'autre part. Ces métadonnées fournissent ainsi des renseignements sur le nom du document ; sa date de création, de mise à jour ; son environnement technique, celui nécessaire pour lire le document (standards d'encodage), leur compatibilité (les standards, les protocoles évoluant, il conviendra d'assurer des migrations régulières en termes de supports de stockage, de langages ou de formats) ; la composition de la page (texte, image, son…) ; des informations juridiques etc.

Des travaux de standardisation de ces métadonnées sont déjà en cours, notamment au sein du projet Dublin Core, maintenu par le *Dublin Core Metadata Initiative (DCMI)*, une organisation qui gère une liste de quinze éléments de bases, chacun pouvant supporter un ou plusieurs raffinements.

2. Considérations techniques

Il ne suffit pas cependant de récolter des pages sur le web pour les préserver sur des serveurs. Encore faut-il en effet préserver les possibilités de lecture de ces pages.

---

[20] KAVCIC-COLIC, Alenka. *Archiving the Web: Some legal Aspects*[en ligne]. In 68th IFLA Council and General Conference, Glasgow. Consulté le 10 mars 2008. Disponible sur : http://www.ifla.org/IV/ifla68/papers/116-163e.pdf

[21] Qu'entend-on par métadonnées ? in BASSINET, Annie, CORMENIER Marie-Noëlle, VIALA Geneviève. *Indexation de ressources métadonnées, normes et standards* [en ligne]. Educnet, mars 2007. Consulté le 26 février 2008. Disponible sur : http://www.educnet.education.fr/dossier/metadata/default.htm

Des questions qui apparaissent d'autant plus cruciales dans notre société de l'information où les objets ne sont pas produits pour durer mais doivent être rapidement mis sur le marché.

2.1. Préserver l'environnement technologique

Les technologies numériques, en effet, permettent de créer, retoucher, manipuler, diffuser, enregistrer facilement l'information ; ce qui n'est pas sans poser certains problèmes en terme de conservation.

La solution la plus simple est celle de copier les données sans aucun changement, c'est à dire sans changer le contexte logiciel et matériel. Elle requiert cependant que les progrès technologiques ne soient pas tant avancés qu'ils ne permettent plus de lire les informations.

La migration est une solution un peu plus complexe qui propose, en plus de la copie des données, de les transférer sous une nouvelle configuration logicielle et/ou matérielle lorsque, par exemple les progrès technologiques ont permis la mise au point d'une nouvelle génération d'ordinateur, une solution qui devient d'autant plus importante que les supports numériques apparaissent peu fiables sur le long terme.

L'émulation est une technique qui offre quelques bénéfices supplémentaires en préservant l'intégrité initiale des documents numériques. L'environnement technique (et donc son fonctionnement originel) est également sauvegardé et se voit simulé sur l'environnement actuel. Cette technologie permet de préserver le contexte de lecture de l'objet numérique.

Le concept de l'ordinateur virtuel universel (Universal Virtual Computer - UVC) développé par Raymond Lorie utilise des composantes des deux techniques de la migration et de l'émulation pour reconstituer la forme originelle et l'environnement initial des objets numériques. Le concept consiste en l'utilisation d'un UVC, d'un schéma logique de données avec leur description, d'un décodeur de format et d'une visionneuse.

L'encapsulation permet enfin de regrouper ensemble plusieurs documents numériques avec ce qui est nécessaire à leur lecture.

Enfin, si l'on considère que les standards et les protocoles sont importants pour permettre la préservation des données numériques, il ne faut pas oublier non plus que ces standards technologiques évoluent eux-mêmes rapidement.

On soulignera encore l'importance de la documentation comme outils d'accompagnement pour la préservation des documents numériques. Il peut s'agir d'une documentation à part, reprenant les méthodes adoptées et leurs raisons, comme il peut s'agir des données signifiées dans les métadonnées des documents eux-mêmes décrivant matériels, logiciels et autres configurations requises.

2.2. La gestion des risques

Les programmes de préservation doivent chercher à comprendre et à faire face aux menaces qui mettraient en péril l'accessibilité continue et les autres aspects de leur mission. Une méthode de gestion du risque fournit la base voulue pour déterminer les risques à surveiller et planifier une action qui abaissera le niveau du risque.[22]

2.2.1. Un double stockage

Le principe d'un double stockage avec un site miroir physiquement distant peut être appliqué et constitue un minimum pour une conservation numérique. Il s'agit en effet de se préserver des éventuelles catastrophes qui peuvent survenir.

Ceci dit, il est important d'en tenir compte pour une estimation des volumes et donc des coûts de stockage. La duplication a un impact non seulement sur le volume brut de données, mais également sur les volumes à traiter lors des

[22] Bibliothèque nationale d'Australie. *Directives pour la sauvegarde du patrimoine numérique* [en ligne]. Paris : UNESCO, 2003. Consulté le 10 mars 2008. Disponible sur : http://portal.unesco.org/ci/en/ev.php-URL_ID=13271&URL_DO=DO_TOPIC&URL_SECTION=201.html

opérations de conservation préventive qui sont la règle unique pour les ressources numériques.

En général, les fichiers conservés sur ces espaces de stockages constituent les exemplaires de conservation de haute qualité tandis que le principe d'un stockage de dimension plus restreinte peut également se voir retenu pour les copies destinées à la communication en ligne sur Internet. Cet espace plus restreint correspond au sous-ensemble des données libres de droit pouvant être communiquées publiquement sur le réseau, ainsi qu'à des copies réduites en taille et conformes à ce qui est aujourd'hui accepté comme qualité de résolution pour la consultation sur Internet. Ce sont ces copies de plus basse résolution qui sont mises à disposition d'un public et donc « publiées » au sens nouveau du support de publication Internet et qui seront sujettes à dépôt légal de publications en ligne.

2.2.2. La gestion des risques

En termes de gestion des risques, la *National library of Australia* a appliqué une méthode qui permet de distinguer les catégories suivantes de risques[23] :

- Les risques globaux (risques environnementaux, risques liés à la sécurité en général et plus particulièrement dans le cas des collections numériques la sécurité informatique, les risques liés au système informatique de gestion lui-même) ;
- Les risques organisationnel (budget adéquat régulier, risque d'erreurs humaines - passage d'un traitement de conservation a posteriori à l'unité vers un traitement de masse a priori dans lequel le respect des procédures est essentiel) ;
- Les risques technologiques (problèmes de conservation des supports physiques dont questions des formats des fichiers et d'obsolescence des environnements informatiques) ;

[23] LUPOVICI, Catherine. *La conservation des publications électroniques et du dépôt légal* [en ligne]. In Information for All Programme (IFAP). S.l. : UNESCO, 2007. Consulté le 26 février 2008. Disponible sur : http://portal.unesco.org/ci/en/ev.php-URL_ID=24441&URL_DO=DO_TOPIC&URL_SECTION=201.html

- Les risques liés à l'accès (comprendre le sens d'un contenu comme. enregistrement de toutes les informations qui seront nécessaires pour utiliser la publication, gestion des droits).

3. Considérations organisationnelles

Derrière la majeure partie des initiatives d'archivage du web se retrouvent des informaticiens d'une part, des bibliothécaires et archivistes d'autre part, chacun apportant des compétences bien précises.

3.1. Compétences des personnels

La composition de ces équipes dépend en fait du projet qui se trouve derrière : s'il s'agit surtout de moissonner les sites web, des compétences techniques seront plutôt requises. Si, en revanche, il s'agit de sélectionner l'information, un plus grand nombre de compétence peut être nécessaire, par exemple au niveau de la :

- Sélection : Mise en œuvre et développement d'une politique d'acquisition et d'un plan de développement des collections ;
- Négociation : Les propriétaires des sites sélectionnés doivent être contactés afin d'obtenir leur accord pour l'archivage, avec d'éventuelles question de restriction d'accès. C'est un travail qui peut être très preneur de temps.
- Traitement : Les droits acquis, peuvent maintenant faire l'objet d'un traitement (procédure de validation, création des métadonnées). D'un point de vue des compétences, donc, le développement du logiciel peut être confié à des informaticiens et le traitement par des spécialistes de l'information.
- Conservation : Plusieurs compétences peuvent être utilisées pour mettre en place des procedures de conservation, qu'il s'agisse de la définition de la politique et des modalités de conservation des documents ou du développement de processus de migration et d'archivage réguliers.

De fait, si les professionnels de l'information peuvent prendre en charge les processus d'archivage du web, étant entendu que le métier évolue sensiblement

au su de la prégnance de plus en plus forte d'un environnement technique. D'autres compétences, techniques cependant, et juridiques, ne sont pas inutiles.

3.2. Évolution des métiers

Dans nombres de projets, même si l'essentiel des données provient de collectes automatiques, la part de sélection humaine demeure présente voire, comme dans le cas d'une approche sélective manuelle, prépondérante. Les bibliothécaires chargés de ces acquisitions particulières doivent alors apprendre à sélectionner des ressources de l'Internet, dans une perspective de représentativité en ajustant leurs pratiques d'évaluation des contenus aux spécificités de la Toile et aux techniques d'archivage automatique.

Gildas Ilien et Valérie Game expliquent dans un article du Bulletin des Bibliothèques de France[24] qu' « *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.* » L'autre évolution majeure du métier concerne le traitement physique et intellectuel des archives qui devient de plus en plus automatisé tant au niveau de l'indexation, de la conservation ou de la consultation.

Gildas Ilien et Valérie Game poursuivent : « *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.* »

[24] ILLIEN, Gildas ; GAME, Valérie. Le dépôt légal d'Internet à la Bibliothèque nationale de France : Cadre juridique, modèle de collecte, évolutions des métiers [en ligne], in *BBF*, 2006, n° 3, p. 82-85. Consulté le 10 mars 2008. Disponible sur : http://bbf.enssib.fr

« *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.* »

3.3. Durabilité

Afin d'assurer une certaine pérennité aux entreprises d'archivage du web, il convient de favoriser les projets collaboratifs. La nécessité de mener des projets d'archivage du web semble comprise. Les bibliothèques nationales de nombreux pays ont ainsi mené des initiatives d'archivages pour préserver leur héritage culturel. Le site *Preserving acess to digital information*[25], de la *National library of Australia* recense ainsi la participation active de près d'une vingtaine de pays. Cela signifie également que de nombreuses approches ont été adoptées, approches qui devront évoluer de façon à inclure plus de contenus sémantiques ou toujours plus de données. La nature même du web le rend impossible à appréhender pour une seule et même institution. Si les bibliothèques nationales apparaissent comme devant jouer un rôle certain dans l'archivage du web « national », une collaboration efficace avec d'autres organismes (qu'il s'agisse d'autres institutions nationales ou de bibliothèques d'autres pays pour un partage d'informations et d'expériences) ne saurait nuire à la bonne mise en place des politiques de conservation, voire, cela peut aider à mettre en place un accès pérenne et unifié aux contenus de l'internet et des archives du web.

[25] Bibliothèque nationale d'Australie. Web archiving [en ligne]. In *Preserving Access to digital information*. Canberra : National library of Australia, s.d. Consulté le 26 février 2008. Disponible sur : http://www.nla.gov.au/padi/topics/92.html

Partie 3 : Initiatives et projets

L'archivage du web n'est pas un projet récent. Dès 1996, la Conférence des directeurs de bibliothèques nationales (CDNL) estimait qu'il était essentiel de répertorier et de conserver à perpétuité les publications électroniques, "*faute de quoi le passé ne laisserait plus de traces, et les recherches antérieures ne pourraient plus être retrouvées, comprises ou reproduites comme il se doit, pour entretenir le cycle continu de l'expérimentation qui fait progresser la connaissance*"[26].

De nombreuses initiatives ont ainsi émergé depuis plus d'une dizaine d'année révélant des approches différentes avec chacun ses avantages et ses inconvénients, qu'il s'agisse de collecte manuelle ou automatique, d'un archivage intégral, exhaustif ou sélectif[27]. Ces initiatives nationales ont ensuite été soutenues par la constitution de *consortia* regroupant plusieurs établissements.

1. Stratégies et modes de dépôt

Plusieurs stratégies de dépôts ont été développées par les établissements et organisations qui ont mises en places des projets d'archivage du web.

Une étude sur les avantages et les désavantages de ces différentes stratégies de collecte et de dépôt peut être trouvée en ligne sur le site du consortium anglais JISC[28].

---

[26] BRIAN Lang. *Le dépôt légal des publications électroniques* [en ligne]. Paris : UNESCO, 1996. Consulté le 21 février 2008. Disponible sur : http://unesdoc.unesco.org/images/0010/001055/105504Fo.pdf

[27] Gharsallah, Medhi. Archivage du web français et dépôt légal des publications électroniques in *Documentaliste, Sciences de l'Information*, 2004. Consulté le 27 août 2007. Disponible sur internet : http://archivesic.ccsd.cnrs.fr/sic_00001311/fr/

[28] DAY, Mickael. *Collecting and preserving the world wide web : a feasabilitty studyundertaken for the JISC and Wellcome trust* [en ligne]. JISC, 2003. Consulté le 26 février 2008. Disponible sur : http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

1.1. Différentes approches de dépôt

1.1.1. L'approche intégrale

Cette approche consiste à collecter l'ensemble du web sans aucune notion de sélection, de valeur patrimoniale ou de dépôt. Un exemple de cette approche réside dans le projet Internet Archive.

1.1.2. L'approche exhaustive

Cette approche sur le nom de domaine est caractérisée par la récolte de site internet ou de ressources en ligne correspondant à un espace national défini. Il s'agira par exemple, de récupérer tout ce qui appartient à l'internet français.
Des initiatives telles que Kulturarw3 (Cultural Heritage Cubed) en Suède ou le projet EVA en Finlande participent d'une telle approche.

1.1.3. L'approche sélective

Les approches sélectives ne tentent que d'archiver des portions définies du web, ou des ressources particulières selon des critères spécifiques. La sélection peut se fonder sur la qualité des ressources, leur thème ou en ciblant un type particulier de site web.
Plutôt que d'essayer de collecter l'ensemble du web, cette approche vise à prendre des instantanés des sites à des intervalles précis. L'exemple d'une telle approche pourrait être le projet de la *National Library of Australia* : PANDORA

1.1.4. L'approche thématique

Un aspect particulier de l'approche sélective est la constitution d'une collection de site web à l'occasion d'un événement particulier.
Cette approche a été notamment utilisée par la Bibliothèque nationale de France à l'occasion des élections présidentielles et législatives de 2002 et lors des élections régionales et européennes de 2004 notamment. Elle a également été utilisée par la Bibliothèque du Congrès qui, dans son projet Minerva, a collecté des archives à l'occasion des élections en 2002 et des Jeux Olympiques d'hiver.
L'approche.

Certains instituts de recherche ont par ailleurs également mené des projets de collecte disciplinaires. C'est le cas notamment de l'Université de Heidelberg avec son projet DACHS (*Digital Archives for Chinese Study*). C'est le cas également du projet hollandais ARCHIPOL censé récolter les sites web politiques aux Pays-Bas.

1.1.5. Des approches combinées

Aucune stratégie de dépôt ne peut être entièrement satisfaisante pour la préservation du patrimoine national sur internet, c'est pourquoi un certain nombre de pays ont décidé de combiner plusieurs de ces approches.
En France et au Danemark, par exemple, la collecte se fait sur plusieurs niveaux : une collecte exhaustive sur l'ensemble des sites relevant du domaine national et caractérisés entre autres par un nom de domaine finissant par l'indicatif du pays, couplé à une collecte sélective en fonction de critères définis et enfin la mise sur pied de campagnes de collectes thématiques à l'occasion d'événements particuliers.

1.2. Modes de dépôt

1.2.1. Le dépôt légal des sites web

Dans la plupart des pays, les modes de dépôts des sites et des ressources numériques sont appuyés par des cadres législatifs. Il s'agit la plupart du temps d'une extension des limites du dépôt légal déjà en œuvre par ailleurs pour les documents papiers (cf. supra Partie 1 : chapitre 3.3).

La définition du champ du dépôt légal varie d'un pays à l'autre et peut s'appuyer sur une définition linguistique, sur une définition géographique ou sur une définition thématique. Cette collecte n'implique pas nécessairement une sélection Le dépôt légal correspond toujours à une certaine masse la plus exhaustive possible par rapport au champ documentaire défini généralement par la loi.

1.2.2. Le dépôt volontaire

Cependant, si la plupart des pays ont en effet mis en œuvre une extension du dépôt légal aux ressources électroniques, dans certains cas, des modes de dépôts volontaires ont été enregistrés. Actuellement, le dépôt des ressources numériques par leur propriétaires n'est pas vraiment établi ni répandu bien que quelques expériences dans ce sens sont menées ici ou là.

Aux Pays-Bas, par exemple, la Bibliothèque Royale propose un service « e-depot », fruit d'un accord avec des éditeurs de périodiques électroniques –dont Elsevier- pour proposer des archives de l'ensemble des revues éditées aux Pays-Bas. Elle vient par ailleurs de passer un accord avec Portico, une organisation à but non lucratif qui offre des services aux revues qui ont besoin d'être préservées, et aux bibliothèques qui veulent s'y abonner.

1.3. Approches automatisées ou manuelles

Les stratégies de récoltent peuvent répondre à des besoins différents et utilisent des stratégies de collectes différentes. Certaines peuvent être automatisées, semi-automatisées ou manuelles.

1.3.1. Approche automatisées

L'approche automatisée consiste en l'utilisation d'un robot de recherche censé moissonner automatiquement l'ensemble des sites correspondants aux critères spécifiés. Il s'agit là souvent, notamment dans le cadre d'une stratégie de collecte exhaustive, d'identifier des sites correspondant à un domaine géographique donné (avec une extension en .se par exemple dans le cadre du projet Kulturarw3, ou identifié comme localisé en Suède en utilisant le WHOIS.

1.3.2. Approches semi-automatisée

La collecte semi-automatisée combine une récolte automatisée et donc l'utilisation d'un robot de moissonnage et l'application de critères de sélections plus strictes. Utilisée notamment dans des approches sélectives, elle permet de répondre à une exigence de qualité des sites collectés.

1.3.3. Le facteur humain

Si aucune approche entièrement manuelle ne saurait être mise en œuvre, il ne faut pas nier pour autant l'importance du facteur humain.

En France, l'initiative de la Bibliothèque nationale de France reconnaît la valeur humaine dans le processus de collecte quand elle affirme que les bibliothécaires sont nécessaires pour corriger les erreurs que ce soit par une action positive (forcer un crawl) ou négative (bloquer le crawl de *microsoft.com*. De plus les bibliothécaires peuvent identifier, sélectionner et collecter les sites du web profond à collecter.

2. Tour d'horizon des initiatives individuelles : quelques exemples

2.1. Internet Archive : une approche intégrale

On trouve en effet dans l'histoire de l'internet des projets d'archives intégrales de la Toile à travers notamment l'initiative *Internet Archive*[29] lancée en mars 1996 par Brewster Kahle. Cette approche consiste à collecter l'ensemble du web sans aucune notion de sélection, de valeur patrimoniale ou de dépôt.

Ce qui n'était au départ qu'un simple projet de recherche va vite devenir une société basée à San Francisco, à l'origine dès juillet 1997 d'un outil commercial appelé *Alexa*. Cet outil permet de « butiner », rapatrier et indexer un nombre important de pages et de donner des indications sur leur fréquentation, le renouvellement, le nombre de liens, … mais surtout il permet de donner accès aux versions précédentes des sites archivés par *Internet* Archive. A l'heure actuelle, elle dispose dans ses archives de près de 85 milliards de pages, consultables gratuitement. L'archive s'est également diversifiée et donne également accès à 113 000 video, 237 000 archives sonores, 43 000 concerts live, 364 000 textes numérisés.

[29] *Internet Archive* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://www.archive.org/index.php

Ce projet suit une logique très volontaire qui archive rapidement le plus de sites possibles du web sans considérations de frontières ou de nationalités, ni même de demande d'autorisation de la part des auteurs, ce qui peut porter préjudice à ces derniers comme aux sociétés souhaitant vendre leurs archives. De plus, l'indexation demeure sommaire, les sites étant uniquement indexés par date. De fait, on ne connaît pas la taille réelle de l'archive en nombre de fichiers et de documents.

Le projet Internet Archive a rejoint dès sa création un consortium pour la préservation de l'internet : l'IIPC au sein duquel il travaille avec des bibliothèques nationales sur des questions comme le moissonnage, la diffusion, la conservation sur le long terme des archives.

2.2. KulturarW^3 : un approche exhaustive automatisée

Le projet Kulturarw^3, mis en route en 1996 par la Bibliothèque royale de Suède, a opté pour une approche exhaustive, partant du principe qu'il est extrêmement difficile, aujourd'hui, de distinguer ce qui sera important dans vingt, quatre-vingts ou cents ans de ce qui ne le sera pas[30]. Une approche exhaustive veut collecter l'ensemble d'un domaine –en l'occurrence le .se pour Kulturarw^3 Mais se limite à ce domaine. Le choix d'une approche exhaustive, dans ce cas nécessairement automatisée, permet de recueillir un corpus important à moindre frais mais ne permet qu'une sélection sommaire des documents et ne propose que peu d'indexation.

Jusqu'en 2006, la Bibliothèque royale a effectué douze balayages du contenu suédois sur internet en utilisant un robot d'indexation modifié en robot d'archivage. La taille du web suédois représentait en 2000, 5 millions de pages réparties sur 31.000 sites (dont 25.000 en .se et 6.000 dans les autres

[30] Kungl. Biblioteket [bibliothèque nationale de Suède]. *Project on Digital Deposit : new technical challenges and solutions* [en ligne]. Stockholme : Kungl biblioteket, 2007. Consulté le 11 mars 2008. Disponible sur : http://www.kb.se/Dokument/Om/projekt/digital_dep_degerstedt.pdf

domaines). Avec les images et le son, la taille de la base s'élevait à 9.7 millions de fichiers soit 200 Giga bites. Le problème d'une telle base ne résidant pas tant dans la taille que dans la grande diversité des types de documents et dans la gestion des liens.

Kulturarw[3] participait à un projet de plus grande ampleur appelé NWA[31] : Projet d'accès aux archives nordiques du web, regroupant les bibliothèques nationales des cinq pays nordiques (Danemark, Finlande, Islande, Norvège et Suède) et dont l'objectif majeur était la création d'un système d'accès commun aux différents sites d'archivages des bibliothèques nationales nordiques. En 2003, les membres de ce projet ont rejoint le consortium international IIPC.

2.3. Pandora : une approche sélective semi-automatisée

En 1996, la Bibliothèque Nationale d'Australie, en collaboration avec les Archives nationales, met en place le projet PANDORA[32] (*Preserving and Accessing Networked Documentary Ressources of Australia*), une initiative censée collecter et archiver un nombre restreint de sites sélectionnés, des publications électroniques australiennes telles que des périodiques électroniques, des publications officielles, des sites web ayant un contenu pertinent d'un point de vue culturel ou de recherche.

L'approche retenue se veut sélective mais n'est que partiellement automatisée : un site candidat doit en effet subir un certain nombre de critères stricts portant sur la qualité et la pertinence de la publication mais dès 1997, une première archive d'environs 229 documents est déposée.

Le processus d'archivage est contrôlé et semi automatique. La publication est ainsi d'abord évaluée pour jauger son intérêt scientifique mais également sa structure, ses particularités, la permission de l'éditeur est demandée pour l'archivage, la publication est alors cataloguée sur la base de donnée de la BnA,

---

[31] *Nordic Web Archive* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://nwa.nb.no/

[32] Bibliothèque nationale d'Australie. *Pandora Archive* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://pandora.nla.gov.au

les fichiers récupérés sont comparés avec la source en ligne. L'archivage comprend trois copies majeures de la publication (une copie de préservation, une copie consultable et une copie pour les métadonnées). En plus de ces copies principales, une autre copie est conservée sur le serveur pour en faciliter la consultation.

Aujourd'hui, dix partenaires participent à ce projet : l'*Australian Institute of Aboriginal and Torres Strait Islander Studies,* l'*Australian War Memorial*, la *National Film and Sound Archive*, la *National Library of Australia*, la *Northern Territory Library*, la *State Library of New South Wale*s, la *State Library of Queensland*, la *State Library of South Australia*, la *State Library of Victoria*, la *State Library of Western Australia*. Chacune sélectionne des sites en fonction de ses missions. Les publications sont ainsi choisies, décrites, traitées par les bibliothèques participantes mais l'information est stockée dans un serveur à la bibliothèque nationale.

Les sites indexés sont consultables à partir de la base de données de la Bibliothèque nationale d'Australie et accessibles en ligne sur le site internet de PANDORA. Ils suivent une politique d'archivage disponible en ligne[33], qui comprend même certains sites du web profond accessibles via un outil développé pour cela : Xinq. Pour permettre l'accès à l'ensemble des titres, la bibliothèque négocie avec certains éditeurs pour acquérir les fichiers sources stables de certains formats dynamiques ; des stratégies de migration, l'utilisation d'émulateurs lorsque cela est nécessaire.

2.4. Le projet de la Bibliothèque Nationale du Québec : une approche sélective manuelle

L'approche de la bibliothèque nationale du Québec se veut au contraire des précédentes, une approche sélective. Ce choix lui permet de collecter une archive de qualité et induit une indexation fine des contenus mais se révèle coûteux en temps et en ressources humaines.

---

[33] Bibliothèque nationale d'Australie. *A digital preservation policy for the National Library of Australia* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://www.nla.gov.au/policy/digpres.html

La Bibliothèque nationale du Québec a débuté l'implantation d'un dépôt légal, applicable aux documents numériques depuis 1992, en février 2001 en ciblant prioritairement les monographies et publications en série diffusées sur Internet par le gouvernement du Québec[34]. Tous les ministères déposent désormais leurs publications diffusées sur Internet. De février 2001 à septembre 2006, près de 9 000 publications ont ainsi été recueillies.

Pour chaque publication, une licence de cession des droits d'auteur est signée avec le ministère ou l'organisme gouvernemental. Elle stipule que la Bibliothèque peut capturer une copie de la publication, archiver les fichiers recueillis sur son serveur, effectuer les opérations nécessaires à la diffusion de la publication au sein de la bibliothèque numérique et à la conservation à long terme.

L'accès est gratuit pour des publications gratuites, restreint aux locaux de la bibliothèque avec possibilité de PEB (consultation sur place) en cas de publication vendue. Une mise en accès gratuite est cependant possible si l'adresse internet de la publication n'est plus valide après un délai de trente jours. Il s'agit d'une licence non exclusive, irrévocable, sans limite de territoire ou de temps, transférable uniquement en ce qui concerne le PEB dans laquelle l'éditeur demeure le seul propriétaire de tous les droits d'auteurs.

2.5. BnF et Ina : une approche par échantillonnage semi-automatisé

La Bibliothèque nationale de France (BnF), associée à l'Institut national de l'audiovisuel (Ina), mènent depuis 1998, des études sur l'archivage de la Toile. Son approche, nourrie des initiatives des institutions précurseurs, se veut mixte et propose une collecte multiple à la fois automatique et ponctuelle.

---

[34] Bibliothèque et Archives nationales du Québec. *Publication diffusée sur Internet* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://www.banq.qc.ca/portal/dt/collections/dons_acquisitions/depot_legal/publications_assujetties/publications_internet/publications_internet.jsp

Pour ce faire, la Bibliothèque bénéficie de la loi DADVSI[35] qui, dans son titre IV, a élargi le champ du dépôt légal aux sites internet français et désigné la BnF et l'Ina comme les organismes dépositaires. Elle prévoit en effet l'extension du dépôt légal à tous « *les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique* ».

Dès lors, ces deux institutions doivent se répartir la tâche en fonction du type de contenu du site ainsi collecté, chacune s'intéressant aux documents relevant du domaine correspondant à sa mission initiale dans une logique de continuité de ses collections. Ainsi l'Ina sera-t-elle en charge de la conservation des sites axés sur la télévision et la radio tandis que la BnF collectera tous les autres. Du côté des opérateurs, l'obligation de dépôt légal pèsera sur les personnes qui éditent et produisent des sites internet mais n'impliquera pas de démarche particulière de leur part, la collecte étant principalement le fait de robots.

Les archives de la Toiles seront alors consultables sur place dans les salles de recherches de la BnF. A cette occasion, la loi prévoit 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, sans avoir à requérir d'autorisation préalable ni à verser de rémunération, 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[36].

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 :

---

[35] loi n° 2006-961 du 1er août 2006 *relative au droit d'auteur et aux droits voisins dans la société de l'information* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://www.legifrance.gouv.fr/WAspad/UnTexteDeJorf?numjo=MCCX0300082L

[36] ILLIEN, Gildas ; GAME, Valérie. Le dépôt légal d'Internet à la Bibliothèque nationale de France : Cadre juridique, modèle de collecte, évolutions des métiers [en ligne], in BBF, 2006, n° 3, p. 82-85. Consulté le 10 mars 2008. Disponible sur : http://bbf.enssib.fr

- des captures massives et automatiques du domaine français réalisées au moyen de robots. Dans le cadre d'un partenariat de recherche avec l'organisme américain Internet Archive, la BnF a ainsi réalisé deux instantanés du domaine « .fr », fin 2004 et fin 2005. Des copies d'instantanés des domaines génériques et français (collections historiques de 1996 à 2005) ont également été acquises : elles représentent plus de 6 milliards de fichiers pour un volume de 60 téraoctets. Entre 2004 et 2006, des instantanés du domaine.fr et des collectes ciblées intégrées ont été réalisées apportant quelque 3 à 4 To pour chaque instantané.[37]

- des collectes thématiques et événementielles qui se fondent sur l'expertise de bibliothécaires travaillant dans les départements de collection et de dépôt légal. Elles viennent pallier l'insuffisance des robots en s'appuyant sur une prospection documentaire ciblée. Les principaux critères de sélection retenus seront pour la continuité des collections et le signalement de nouveaux objets documentaires particulièrement représentatifs des formes nouvelles de l'édition, comme les blogs. Des campagnes de collectes peuvent encore être organisées autour d'un événement d'intérêt national, comme ce fut le cas lors des élections présidentielles et législatives de 2002 et lors des élections régionales et européennes de 2004. Le volume de ces collectes représente pour les élections en 2002 et 2004 : 3500 sites ; 12617 captures ; 23 millions de fichiers ; 535 Go et pour les élections en 2007 : 2700 sites, 35 millions de fichiers, 2,3 To soit au total quelque 10 milliards de fichiers et 130 To de données !

- la mise en place d'un circuit de dépôts à l'unité pour un nombre limité de sites qu'on ne peut archiver autrement. Il s'agit de 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. 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

---

[37] ILLIEN, Gildas. *Le dépôt légal de l'internet : principes et réalisations* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://vds.cnes.fr/pin/presentations/2007/bnf_depot_legal.pdf

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 1er juin 2005

Le traitement documentaire 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.

L’Ina collecte quant à elle également les sites en utilisant des robots sous l’autorité d’une cellule de veille. La copie de chaque site est ensuite préparée pour une indexation séparée des contenus du site et de leur structure, les différentes images qui composent une page, la page elle-même, les films, les musiques, etc. étant traités indépendamment.

3. Projets collaboratifs

Il y a pour les programmes de préservation de bonnes raisons technologiques, économiques et politiques de coopérer. Les décisions relatives à une collaboration pouvant se fonder sur une évaluation des bénéfices attendus et des dépenses à engager. De fait, un consortium a vu le jour entre plusieurs initiatives qui essaie de fédérer les travaux menés sur le sujet : le consortium international pour la conservation de l’internet.

Les avantages qui incitent à entreprendre un effort de coopération sont multiples qui permettent l'accès à un éventail plus large de compétences (en terme de développement, ou de gestion…) et des coûts moindres puisque partagés. Par ailleurs, cela autorise une couverture accrue des matériaux préservés. En termes d’influence enfin, un projet collaboratif permettra un encouragement pour d'autres parties prenantes influentes à prendre la préservation au sérieux ainsi qu’une influence certaine sur les accords avec les producteurs ou sur la recherche et le développement des normes et des pratiques assurant une

interopérabilité entre les systèmes et donc de meilleures conditions de préservations et d'accès sur le long terme.[38]

3.1. International Internet Preservation Consortium

Le consortium international pour la conservation de l'internet (*International Internet Preservation Consortium* - *IIPC*) est créé en juillet 2003. Il regroupe alors les bibliothèques nationales d'Australie, du Canada, du Danemark, de Finlande, de France, d'Islande, d'Italie, de Norvège, de Suède, la *British Library*, la Bibliothèque du Congrès et le projet *Internet Archive*.

Ces douze institutions se sont regroupées pour mettre en œuvre divers projets et groupes de travail devant répondre aux objectifs du consortium :

- Permettre la collecte et la conservation à long-terme d'une part importante des contenus de l'internet
- Encourager le développement et l'utilisation d'outils communs, de techniques et des normes pour la création d'archives internationales.
- Plaider pour la mise en œuvre d'initiatives et de législation en faveur de la collecte, de la conservation et d'un accès au contenu d'Internet.
- Encourager et soutenir les établissements (bibliothèques, archives, musées et autres établissements culturels) à s'engager dans la collecte et la conservation du contenu d'Internet[39]

En 2008, le consortium regroupe quelques trente-six institutions venues de toutes les régions du monde (Bibliothèques du Canada, d'Israël, de Croatie, Singapour, Nouvelle Zélande, Autriche…) dont vingt institutions publiques nationales (archives ou bibliothèques nationales), deux bibliothèques publiques, une bibliothèque universitaire, et trois société d'offres de service. Il a notamment accueilli d'autres consortia comme le *Nordic Web Archive*[40], en 2003, un projet

---

[38] Bibliothèque nationale d'Australie. Directives pour la sauvegarde du patrimoine numérique [en ligne]. Paris : UNESCO, 2003. Consulté le 10 mars 2008. Disponible sur : http://portal.unesco.org/ci/en/ev.php-URL_ID=13271&URL_DO=DO_TOPIC&URL_SECTION=201.html

[39] *International Internet Preservation Consortium* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://netpreserve.org/about/index.php

[40] *Nordic Web Archive* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://nwa.nb.no/

d'archivage collaboratif réunissant les bibliothèques nationales du Danemark, de Finlande, d'Islande, de Norvège et de Suède

Entre 2003 et 2006, durée du premier accord regroupant exclusivement les premiers signataires, six groupes de travail ont été mis en place afin d'étudier les questions techniques, les besoins pour la recherche, les outils d'accès, la gestion des contenus, le web profond et les éléments d'évaluation du web. En 2007, quatre autres groupes sont mis en place qui travaillent sur les normes et les standards, le moissonnage du web, les modalités d'accès aux archives et la conservation des données[41].

3.2. NEDLIB

Réalisé par la Commission européenne, le projet NEDLIB[42] (*Networked European Deposit Library*) s'est déroulé entre 1998 and 2001 avec pour objectif de développer l'infrastructure de base sur laquelle un entrepôt de dépôt collaboratif européen pour les publications numériques pourrait être constitué.
Le projet regroupait onze établissements issus de huit pays européens (Allemagne, Finlande, France, Italie, Norvège, Pays-Bas, Portugal, Suisse) et trois éditeurs de publications numériques (*Kluwer Academic*, *Elsevier Science BV*, *Springer-Verlag*). Nedlib fut l'un des premiers logiciels développés uniquement dans une optique d'archivage des données.

NEDLIB fournit aux institutions un ensemble d'outils pour mettre en œuvre des systèmes de dépôts numériques : un modèle de dépôt proposant capture, stockage, accès et conservation à long-terme de publications numériques ; un guide de bonnes pratiques, de normes et de solutions techniques, de procédures pour une implémentation pratiques…

En janvier 2000, le moissonneur de Nedlib fut testé par les institutions partie prenantes du projet et les bibliothèques nationales d'Estonie et d'Islande. D'autres versions mises à jour furent testées sur le web islandais en 2001 et

[41] LUPOVICI, Christian. *The international Internet Preservation Consortium : General strategy and state of work* [en ligne]. In Conference of European National Librarians, Helsinki,, 27 septembre 2007. Consulté le 12 mars 2008. Disponible sur : http://netpreserve.org/events/IIPC%20CENL%20070927.ppt
[42] Bibliothèque nationale des Pays-Bas. *Nedlib* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://nedlib.kb.nl/

finnois en 2002. Le moissonneur aujourd'hui a cependant été remplacé par le moteur Heritrix[43], un moteur open-source développé par l'*Internet Archive* en collaboration avec les bibliothèques nationales des pays scandinaves.

3.3. UK Web Archiving Consortium

Créé en Octobre 2003, le *UK Web Archiving Consortium (UKWAC)*[44] regroupe six institutions majeures qui se sont réparties de manière originale la collecte de sites internet clefs pour le Royaume-Uni, dans les domaines médical, scientifique et culturel, en accord avec les propriétaires des droits.

- Ainsi, la *British Library* archive des sites possédant un intérêt certain du point de vue culturel, reflétant des événements d'importance historique et des sites proposant un caractère novateur.
- Les Archives nationales (*National Archives*) s'intéressent aux documents des six ministères importants que sont : La défense et la politique étrangère ; la justice et les affaires intérieures ; l'éducation, la santé et la culture ; l'économie et les finances ; l'administration.
- *JISC* conservera les sites web de projets innovants issus de la recherche et de l'enseignement supérieur.
- La bibliothèque nationale d'Écosse (*National Library of Scotland*) doit collecter tous les documents s'intéressant à l'histoire, la culture, le patrimoine de l'Écosse et des écossais.
- La bibliothèque nationale du Pays de Galle (*National Library of Wales*) doit collecter tout site évoquant le patrimoine et la vie au Pays de Galle
- Le *Wellcome Trust* doit s'occuper de collecter les sites web médicaux et de santé anglo-saxons.

L'accès aux ressources a ouvert au public dès 2005. Il est gratuit pour les chercheurs.

---

[43] InternetArchive. *Heritrix* [en ligne]. Consulté le 10 mars 2007. Disponible sur : http://crawler.archive.org/
[44] *UK Web Archiving Consortium* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://info.webarchive.org.uk

Le consortium prend des instantanés à peu près tous les six mois des sites web qu'il visite. Les membres font la démarche d'obtenir au préalable l'accord des propriétaires des sites avant de les archiver. Ensuite, ils travaillent à utiliser des outils compatibles et en accord avec les normes proposées par l'IIPC. Par ailleurs, la British Library, les National Archives et la National library of Scotland sont toutes trois également adhérentes au consortium international.

Les membres du consortium anglo-saxon utilisent le système d'archivage PANDAS de la bibliothèque nationale d'Australie. Les coûts de tels développement sont supportés à part égales entre les membres.

Bibliographie

1. PRINCIPES POUR ARCHIVER LE WEB

Afnic [en ligne]. Association Française pour le Nommage Internet en Coopération (AFNIC), 2003, mise à jour le 21 mai 2007 [consulté le 15 février 2008]. Disponible sur : http://www.afnic.fr/actu/stats

Web survey [en ligne]. Netcraft LTD, août 2007 [consulté le 15 février 2008]. Disponible sur : http://news.netcraft.com/archives/2007/08/06/august_2007_web_server_survey.html

Bibliothèque nationale d'Australie. *Directives pour la sauvegarde du patrimoine numérique* [en ligne]. Paris : UNESCO, 2003. Consulté le 10 mars 2008. Disponible sur : http://portal.unesco.org/ci/en/ev.php-URL_ID=13271&URL_DO=DO_TOPIC&URL_SECTION=201.html

Bibliothèque et Archives du Canada. *Vers une stratégie canadienne sur l'information numérique* [en ligne]. Disponible sur internet : http://www.collectionscanada.ca/scin/012033-801-f.html (consulté le 2 septembre 2007)

ERTZCHEID, Olivier. Question de taille… in *Affordance.info* [en ligne]. Mars 2007 [consulté le 15 février 2008]. Disponible sur : http://affordance.typepad.com/mon_weblog/2007/03/question_de_tai.html

HOOG, Emmanuel. Internet a-t-il une mémoire. *Le Monde* [en ligne], 17 août 2002 [consulté le 15 février 2008] Disponible sur : http://www.ac-versailles.fr/pedagogi/ses/vie-ses/hodebas/hoog1.htm

LANG, Brian (dir.). *Le dépôt légal des publications électroniques*. Paris : UNESCO, 1996.41 p. ; 30 cm. - (CII-96/WS/lO). consulté le 29 août 2007. Disponible sur : http://unesdoc.unesco.org/images/0010/001055/105504Fo.pdf

LAWRENCE G. W., KEHOE W. R., RIEGER O. Y., WALTERS W. H., KENNEY A. R., *Risk Management of Digital Information: A File Format Investigation*, in Council on Library and Information Resources, Washington, D.C., 2003. Consulté le 10 mars 2008. Disponible sur : http://www.clir.org/pubs/reports/pub93/contents.html

LUPOVICI, Catherine. « La mise en œuvre du dépôt légal électronique » [en ligne] *in Journées internet pour le droit.* Paris : sl, 3-5 novembre 2004. Disponible en ligne : http://www.frlii.org/spip.php?article48 (consulté le 29 août 2007).

LUPOVICI, Catherine. *La conservation des publications électroniques et du dépôt légal* [en ligne]. In Information for All Programme (IFAP). S.l. : UNESCO, 2007. Consulté le 26 février 2008. Disponible sur : http://portal.unesco.org/ci/en/ev.php-URL_ID=24441&URL_DO=DO_TOPIC&URL_SECTION=201.html

LYMAN, Peter, VARIAN Hal. How much information ? [en ligne]. In *School of information management and systems*, 2003. Consulté le 21 février 2008. Disponible sur : http://www.sims.berkeley.edu/research/projects/how-much-info

Médiamétrie/NetRatings. *L'audience de l'internet en France : décembre 2007* [en ligne]. Consulté le 15 février 2008. Disponible sur : http://www.mediametrie.fr/resultats.php?rubrique=net&resultat_id=504

O'NEILL Edward, LAVOIE Brian, BENNETT Rick. Trends in the Evolution of the Public Web 1998 – 2002. IN *D-Lib Magazine*, April 2003, Volume 9, 4. ISSN 1082-9873 Consulté le 15 septembre 2007. Disponible sur : http://www.oclc.org/research/projects/archive/wcp/

2. DÉPÔT ÉGAL DES RESSOURCES NUMÉRIQUES

loi n° 2006-961 du 1er août 2006 *relative au droit d'auteur et aux droits voisins dans la société de l'information* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://www.legifrance.gouv.fr/WAspad/UnTexteDeJorf?numjo=MCCX0300082L

Conference of European National Librarians, Federation of European Publishers. *Statement on the Development and Establishment of Voluntary Deposit Schemes for Electronic Publications*[en ligne]. Annual Conference of European National Librarians, Luxembourg, September 28-30, 2005. Disponible sur : http://www.nlib.ee/cenl/docs/05-11CENLFEP_Draft_Statement050822_02.pdf

Bibliothèque nationale de France. Qu'est-ce que le dépôt légal [en ligne]. In *Bibliothèque nationale de France*. Consulté le 21 février 2008. Disponible sur : http://www.bnf.fr/PAGES/infopro/depotleg/dl-france.htm

BRIAN Lang. *Le dépôt légal des publications électroniques* [en ligne]. Paris : UNESCO, 1996. Consulté le 21 février 2008. Disponible sur : http://www.unesco.org/webworld/memory/depot1.htm

British Library. *Code of Practice for the Voluntary Deposit of Non-Print Publications* [en ligne]. Londres : British Library, 2000. Consulté le 10 mars 2008. Disponible sur : http://www.bl.uk/aboutus/stratpolprog/legaldep/voluntarydeposit/

ILIEN, Gildas. *Repenser la politique documentaire à l'échelle de la Toile : L'expérience du dépôt légal d'Internet à la à la Bibliothèque nationale de France*. In Congrès de l'ADBU, 21 septembre 2007. Consulté le 10 mars 2008. Disponible sur : http://www.adbu.fr/IMG/pdf/Gildas_1.pdf

Organisation des nations unies pour l'éducation, la science et la culture. *Charte sur la conservation du patrimoine numérique* [en ligne].Paris : UNESCO, 2003. Consulté le 23 février 2008. Disponible sur : http://portal.unesco.org/ci/en/ev.php-URL_ID=13367&URL_DO=DO_TOPIC&URL_SECTION=201.html

3. CONSIDERATIONS TECHNIQUES ET ORGANISATIONNELLES

ASSELIN, Christophe. Web invisible, web caché, web profond [en ligne]. In *IntelligenceCenter.com*. Comsulté le 26 février 2008. Disponible sur : http://c.asselin.free.fr/french/invisible_web.htm

BASSINET, Annie, CORMENIER Marie-Noëlle, VIALA Geneviève. Qu'entend-on par métadonnées ? in *Indexation de ressources métadonnées, normes et standards* [en ligne]. Educnet, mars 2007. Consulté le 26 février 2008. Disponible sur : http://www.educnet.education.fr/dossier/metadata/default.htm

Bibliothèque nationale d'Australie. *A digital preservation policy for the National Library of Australia* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://www.nla.gov.au/policy/digpres.html

Bibliothèque nationale d'Australie. *PADI (Preserving Access to Digital Information)* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://www.nla.gov.au/padi/

Bibliothèque nationale d'Australie. Web archiving [en ligne]. In *Preserving Access to digital information*. Canberra : National library of Australia, s.d. Consulté le 26 février 2008. Disponible sur : http://www.nla.gov.au/padi/topics/92.html

BERGMAN, Mickael K. The 'Deep' Web: Surfacing Hidden Value [en ligne]. In *BrightPlanet*. Consulté le 26 février 2008. Disponible sur : http://www.brightplanet.com/resources/details/deepweb.html

BESSER, Howard. Digital Longevity. *In* SITTS, Maxine. *Handbook for Digital Projects: A Management Tool for Preservation*. Andover, Mass.: Northeast Document Conservation Center. 2000. cité dans LYMAN, Peter. Archiving the world wide web [en

ligne]. In Council on library and Information resources. *Building a National Strategy for Preservation: Issues in Digital Media Archiving*. Berkeley, 2002. Consulté le 26 février 2008. Disponible sur : http://www.clir.org/pubs/reports/pub106/Web.html

DIGIMIND. *Découvrir et exploiter le web invisible pour la veille stratégique* [en ligne]. Consulté le 26 février 2008. Disponible sur : http://www.digimind.fr/publications/white-papers/222-decouvrir-et-exploiter-le-web-invisible-pour-la-veille-strategique-2.htm

Gharsallah, Medhi. Archivage du web français et dépôt légal des publications électroniques in *Documentaliste, Sciences de l'Information*, 2004. Consulté le 27 août 2007. Disponible sur internet : http://archivesic.ccsd.cnrs.fr/sic_00001311/fr/

KAVCIC-COLIC, Alenka. *Archiving the Web: Some legal Aspects* [en ligne]. In 68th IFLA Council and General Conference, Glasgow. Consulté le 10 mars 2008. Disponible sur : http://www.ifla.org/IV/ifla68/papers/116-163e.pdf

LUPOVICI, Christian. The international Internet Preservation Consortium : General strategy and state of work [en ligne]. In *Conference of European National Librarians*, Helsinki,, 27 septembre 2007. Consulté le 12 mars 2008. Disponible sur : http://netpreserve.org/events/IIPC CENL 070927.ppt

LYMAN, Peter. Archiving the world wide web [en ligne]. In Council on library and Information resources. *Building a National Strategy for Preservation: Issues in Digital Media Archiving*. Berkeley, 2002. Consulté le 26 février 2008. Disponible sur : http://www.clir.org/pubs/reports/pub106/Web.html

MERZEAU, Louise. Web en stock [en ligne]. in *Cahiers de médiologie*, septembre 2003, N°16, [consulté le 15 février 2008]. Disponible sur : http://www.merzeau.net/txt/memoire/webenstock.html

RUSSEL, Kelly. *Why Can't We Preserve Everything? Selection Issues for the Preservation of Digital Materials*. St Pancras : The British Library, 1999. Consulté le 10 mars 2008. Disponible sur : http://www.leeds.ac.uk/cedars/documents/ABS01.htm

Serda. *Panorama mondial de l'archivage Web* [en ligne]. Bruxelles : Serda, 2003. Consulté le 11 mars 2008. Disponible sur : http://easi.wallonie.be/servlet/Repository/panorama_mondial_du_web.pdf?IDR=8605

4. RETOUR D'EXPÉRIENCES

International Web Archiving Workshop [en ligne]. Consulté le 12 mars 2008. Disponible sur : http://www.iwaw.net/

ARVIDSON Allan, MANNERHEIM Johan, PERSSON Krister. The Kulturarw3 Project - The Royal Swedish Web Archiw3e - An example of "complete" collection of web pages [en ligne]. In *66th IFLA Council and General Conference*, Jerusalem, Israel, 13-18 August 2004. Consulté le 10 mars 2008. Disponible en ligne : http://www.ifla.org/IV/ifla66/papers/154-157e.htm

ASCHENBRENNER Andreas. *Long-Term Preservation of Digital Material : Building an Archive to Preserve Digital Cultural Heritage from the Internet* [en ligne]. Consulté le 12 mars 2008. Disponible sur : http://www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/

Bibliothèque nationale de France. *Les enjeux du dépôt légal de la Toile* [en ligne]. Consulté le 10 mars 2008. Disponible sur: http://www.bnf.fr/pages/presse/dossiers/toile.pdf.

BRYGFJELD, Svein Arne, BARYLA, Christiane (trad.). « Accès aux archives du web : le projet d'accès aux archives nordiques du web » [en ligne] in *68th IFLA Council and General Conference*. Glasgow : 2002. Consulté le 10 mars 2008. Disponible sur : http://www.ifla.org/IV/ifla68/papers/090-163f.pdf

Council on library and Information resources. *The State of Digital Preservation: An International Perspective*. Washington : CLIR, July 2002. Consulté le 10 mars 2008. Disponible sur : http://www.clir.org/pubs/abstract/pub107abst.html

DAY, Mickael. *Collecting and preserving the world wide web : a feasabilitty studyundertaken for the JISC and Wellcome trust* [en ligne]. JISC, 2003. Consulté le 26 février 2008. Disponible sur : http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

ILLIEN, Gildas. *Le dépôt légal de l'internet : principes et réalisations* [en ligne]. Consulté le 10 mars 2008. Disponible sur : http://vds.cnes.fr/pin/presentations/2007/bnf_depot_legal.pdf

ILLIEN, Gildas ; GAME, Valérie. Le dépôt légal d'Internet à la Bibliothèque nationale de France : Cadre juridique, modèle de collecte, évolutions des métiers [en ligne], in *BBF*, 2006, n° 3, p. 82-85. Consulté le 10 mars 2008. Disponible sur : http://bbf.enssib.fr

Kongelige bibliotek. Publications [en ligne]. In *NetArchive.dk*. Consulté le 12 mars 2008. Disponible sur : http://netarchive.dk/publikationer/index-en.php

Library of Congress. Web capture [en ligne]. Consulté le 12 mai 2008. Disponible sur : http://www.loc.gov/webcapture/

UK Web Archiving Consortium. *Project overview* [en ligne]. Consulté le 12 mai 2008. Disponible sur : http://info.webarchive.org.uk/

VERHEUL, Ingeborg. *Networking for Digital Preservation: Current Practices in 15 National Libraries*. Munich : Saur, 2006. Consulté le 12 mars 2008. Disponible sur : www.ifla.org/VI/7/pub/IFLAPublication-No119.pdf

Table des Annexes

Annexe 1 : Quelques chiffres

1. Le Web

1.1. français :

- Médiamétrie évalue le nombre d'internaute à plus de 31 millions soit 59.4 % de la population, en décembre 2007, dont une majorité serait connectée en haut débit
- L'AFNIC[45], centre d'information et de gestion des noms de domaine internet .fr (France) et .re (Île de la Réunion), a enregistré 1081458 domaines *.fr* et *.re* en mars 2008. Ces statistiques prennent en compte la totalité des domaines *.fr* et *.re* gérés par l'AFNIC, c'est-à-dire tous les domaines et sous-domaines enregistrés pour chacune de ces deux extensions.

1.2. En général

- 70% des pages web ont une durée de vie inférieure à quatre mois
- Selon Netcraft[46], il y aurait quelques 155 millions de sites dans le monde en janvier 2008.

[45] http://www.afnic.fr/
[46] http://news.netcraft.com/archives/2008/01/index.html

2. Projets d'archivage

- Les premiers projets d'archivage du web ont ouvert en 1996 : *Internet Archive* (entreprise privée américaine), *Pandora* (Australie), *Kulturarw3* (Suède).
- Internet Archive proposerait la consultation gratuite de 85 milliards de pages.
- L'archive s'est également diversifiée et donne également accès à 113 000 video, 237 000 archives sonores, 43 000 concerts live, 364 000 textes numérisé

- Les premières collectes du domaine français réalisées par la BnF du 15 décembre 2004 au 30 janvier 2005 ont abouti à la réception de 3 téraoctets de données, représentant un total de plus de 118 millions de fichiers identifiés par une adresse URL.
- Entre 2004 et 2006, des instantanés du domaine.fr et des collectes ciblées intégrées ont été réalisées apportant quelque 3 à 4 To pour chaque instantané
- Le volume des collectes représente pour les élections en 2002 et 2004 : 3500 sites ; 12617 captures ; 23 millions de fichiers ; 535 Go et pour les élections en 2007 : 2700 sites, 35 millions de fichiers, 2,3 To soit au total quelque 10 milliards de fichiers et 130 To de données
- Le dépôt légal en France :
    - les imprimés depuis 1537
    - Les estampes, cartes et plans depuis 1648
    - Les partitions musicales depuis 1793
    - Les photographies et les phonogrammes depuis 1925
    - Les vidéogrammes et les documents sur plusieurs supports depuis 1975
    - L'édition électronique sur support depuis 1992
    - L'internet français depuis 2006

- EN 2005, KulturarW3 avait collecté 305,85 millions de fichiers répartis en 9895 Gigabites et 347642 sites web.

- Le consortium le plus important pour la préservation de l'internet est l'IIPC. Il a été initié et mené par la France de 2003 à 2006. il comptait alors 12 membres. En 2008, le consortium compte 36 membres :
  - **Asia**
    - *Jewish National and University Library* (Israel)
    - *National Library Board, Singapore*
  - **Europe**
    - *Biblioteca de Catalunya* (Library of Catalonia)
    - *Biblioteca Nazionale Centrale di Firenze* (National Library of Italy, Florence)
    - *Biblioteka Narodowa* (National Library of Poland)
    - Bibliotheque nationale de France (National Library of France)
    - *British Library* (U.K.)
    - *Deutsche Nationalbibliothek* (German National Library)
    - *European Archive Foundation*
    - *Hanzo Archives Ltd*. (U.K.)
    - *Kansalliskirjasto* (National Library of Finland)
    - *Koninklijke Bibliotheek* (National Library of the Netherlands)
    - *Kungl. biblioteket* (National Library of Sweden)
    - *Landsbokasafn Islands – Haskolabokasafn* (National and University Library of Iceland)
    - *Latvijas Nacionālā bibliotēka* (National Library of Latvia)
    - *Nacionalna i sveučilišna knjižnica u Zagrebu* (National and University Library in Zagreb, Croatia)
    - *Narodna in univerzitetna knjižnica* (National and University Library, Slovenia)
    - *Národní knihovna České republiky* (National Library of the Czech Republic)
    - *Nasjonalbiblioteket* (National Library of Norway)
    - *National Archives* (U.K.)
    - *National Library of Scotland*

- *Netarchive.dk* (Royal Library and the State and University Library, Aarhus)
- *Österreichische Nationalbibliothek* (Austrian National Library)
- *Schweizerische Nationalbibliothek* (Swiss National Library)
- *Virtual Knowledge Studio – Royal Netherlands Academy for Arts and Sciences*

- **North America**
  - Bibliothèque et Archives Nationales du Québec (BAnQ)
  - *California Digital Library* (U.S.)
  - *Centre for Global eHealth Innovation, WebCite® Internet Citations Archiving Project* (Canada)
  - *Internet Archive* (U.S.)
  - *Library and Archives Canada*
  - *Library of Congress* (U.S.)
  - *Library of Virginia* (U.S.)
  - *United States Government Printing Office*
  - *University of North Texas Librarie*s (U.S.)

- **Oceania**
  - *National Library of Australia*
  - *National Library of New Zealand*

Annexe 2 : Liste des principales initiatives d'archivage du web dans le monde

CONSORTIA ET INITIATIVES COLLABORATIVES

IIPC :

- Consortium international pour la préservation d'Internet (*International internet preservation consortium*) http://netpreserve.org

Nedlib :

- *Networked European Deposit Library* http://nedlib.kb.nl/

NWA :

- Nordic Web Archive . http://nwa.nb.no/

UK Web Archiving Consortium

- http://www.webarchive.org.uk

PROJETS NATIONAUX EN LIGNE

Allemagne :

- Archive Server DEPOSIT.D-NB.DE http://deposit.ddb.de/index_e.htm
- DACHS (Digital Archive for Chinese Studies), Institut d'études chinoises - Université d'Heidelberg http://www.sino.uni-heidelberg.de/dachs/

- NESTOR - Network of Expertise in Long-Term Storage of Digital Resources www.langzeitarchivierung.de/index.php?newlang=eng

Australie :

- Pandora http://pandora.nla.gov.au/index.html
- PADI Preserving Access to Digital Information. http://www.nla.gov.au/padi/

Autriche :

- AOLA (Austrian Online Archive) http://www.ifs.tuwien.ac.at/~aola/

Canada :

- Archives du Web du gouvernement du Canada http://collectioncanada.ca/archivesweb/index-f.html

Chine

- China Web InfoMall http://www.infomall.cn/index-eng.htm

Croatie

- [article] Projet DAMP (Université de Zagreb) http://widwisawn.cdlr.strath.ac.uk/Issues/Vol3/issue3_3_1.html

Danemark:

- Netarchive.dk Netarchive.dk

Etats-Unis :

- MINERVA (Mapping the INternet Electronic Resources Virtual Archive) http://www.loc.gov/minerva/
- Internet Archive http://www.archive.org/index.php

Grèce

- [article] *Archiving the greek web* http://www.iwaw.net/04/Lampos.pdf

Japon

- WARP (Web ARchiving Project) http://warp.ndl.go.jp/

Nouvelle-Zélande :

- National Digital Heritage Archive http://www.natlib.govt.nz/about-us/current-initiatives/ndha

Pays-Bas

- Archive nationales des Pays-Bas (*Digitale Duurzaamheid*) www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=286&categorie=6
- Dépôt des resources numériques à la *Koninklijke Bibliotheek* www.kb.nl/hrd/dd/dd_onderzoek/reports/3-preservation.pdf

République Tchèque :

- Webarchiv http://en.webarchiv.cz/

Singapour

- Web Archive Singapore http://www.nlb.gov.sg/CPMS.portal?_nfpb=true&_pageLabel=CPMS_page_eResources_eArchive

Suède

- Kulturarw3 - http://www.kb.se/kw3/ENG/

Taïwan :

- National Digital Archives programme (NDAP) http://www.ndap.org.tw/index_en.php

Wallonie :

- Netarchiv http://archivesweb.wallonie.be/apps/spip/