ÉTUDE

*Archivage du web : quelques leçons à retenir**

Aïda Chebbi

Le web représente actuellement un espace privilégié d'expression et un milieu d'activités pour plusieurs communautés d'intérêts (individus, organismes publics et privés, gouvernements, etc.). Dans sa dimension visible ou invisible, le web constitue également un réservoir documentaire planétaire sans équivoque et un nouveau vecteur de la mémoire individuelle, organisationnelle et collective (Dennis 1998, Gharsallah 2001, O'Neill et al. 2003, McDonald 2005). En effet, cette plate-forme technologique forme un médium commun à différents genres d'information issus d'une variété d'actions communicatives orales ou écrites, synchrones ou asynchrones (forums de discussion, vidéoconférences, publications en ligne, transactions en ligne, etc.) donnant ainsi lieu à divers modèles socio-techniques de documents numériques (Pédauque 2003, Shepherd et al. 2004). Ces derniers peuvent être soit hérités de l'environnement papier, soit de nature unique et exclusivement propre à l'environnement web (Ghitalla et Boullier 2004).

Largement dépendantes de la logique du web, les ressources informationnelles produites ou déposées sur la toile mondiale se caractérisent, d'un côté, par l'hétérogénéité et l'ubiquité de leurs formats et, de l'autre, par la connectivité de leurs contenus. De même, l'autodiversité et l'autoévolution des fonctionnalités web ainsi que l'instabilité de cette technologie (surtout en absence d'une instance de contrôle et d'un cadre organisateur) contribuent, d'une manière significative et souvent chaotique, à l'accroissement du volume des contenus diffusés et à l'enrichissement de leur nature (Ghitalla et Boullier 2004). La somme de tous ces facteurs rend difficile la formation et l'identification de corpus documentaires homogènes, ou du moins stables, et n'offre donc pas les conditions nécessaires à la mise en œuvre des principes de gestion et de conservation garantissant la pérennité et l'exploitabilité des traces d'un phénomène d'ampleur mondiale et de nature à la fois technologique, économique, sociale et culturelle (Dennis 1998, Watson 1999, Brown 2003, Phillips 2003, Bachimont et al. 2005, Cunningham et Phillips 2005, Eschenfelder 2005, Kallinikos 2005).

---

* Texte produit dans le cadre du Colloque Montréal-Shanghai sur la pérennité du document numérique, 24-25 octobre 2006, Montréal.

Devant l'évolution incontrôlable de la technologie web, la complexité croissante de la nature des documents produits et la perte irréversible de certains contenus (Harries 1999, LeFugy 2001, Day 2003, Phillips 2003, Barry 2004, Pennock et Kelly 2006), plusieurs initiatives d'archivage du web sont en cours tant à l'échelle régionale et nationale qu'à l'échelle internationale. Elles sont essentiellement menées par des institutions de mémoire ou résultent d'alliances entre diverses institutions. Adoptant de plus en plus des approches distinctes, ces projets se multiplient et se diversifient, indiquant ainsi d'une part l'urgence de telles interventions et, d'autre part, la difficulté d'établir un cadre de référence unique et commun.

Dans le cadre de cette communication, nous exposons, tout d'abord, la logique du web, la nature des documents qui y circulent et les raisons pour lesquelles le web doit être archivé. Ensuite, nous décrivons les principales initiatives actuelles d'archivage du web selon les acteurs impliqués et les approches utilisées. Nous essayons de dégager les avantages et les inconvénients relatifs à chaque approche tout en relevant les tendances générales sur le plan pratique. Dans notre dernier chapitre, nous abordons les principaux défis à surmonter et les leçons à retenir des résultats obtenus par les projets réalisés jusqu'à présent.

ARCHIVAGE DU WEB : LES RAISONS ET LES FINALITÉS

Évolution de l'usage du web

Le web est le service le plus connu et le plus utilisé d'Internet. Souvent, il est considéré comme une plate-forme technologique (ensemble de serveurs) donnant accès à des documents hypermédias grâce au protocole HTTP[1] et au mode d'adressage universel URL[2].

> The World Wide Web is a wide area hypermedia information retrieval initiative aimed to give universal access to a large universe of documents. It is organized as a set of HyperText Transfer Protocol (HTTP) servers designed specially for rapid distribution of hypermedia documents. (Foo et Lim 1997, 168)

Le web utilise une approche client/serveur. Le client (poste d'utilisateur) interprète les actions de l'utilisateur pour transmettre la requête de ce dernier au serveur sous forme de commandes. Le serveur, qui distribue des services chez un fournisseur, traite les commandes du client et renvoie en échange les fichiers correspondants aux requêtes reçues. Le résultat (ensemble de pages électroniques principalement codées à l'aide d'un langage de balisage et reliées entre elles) s'affiche, enfin, sous une interface conviviale (fenêtre du navigateur) sur le poste client.

Initialement conçu pour faciliter et accélérer les échanges informationnels, le web se transforme rapidement en un milieu d'expression ouvert et dynamique. De même, il est devenu un moyen inédit d'édition et de publication (Dennis 1998). L'appropriation et l'exploitation des fonctionnalités offertes par cette technologie favorisent, d'une part, la constitution et l'animation de nouvelles formes de sociabilité (Bachimont et al. 2005) desquelles découlent des modes particuliers de communication (forums, weblogs, e-WOM[3], etc.) et, d'autre part, l'émergence d'objets documentaires particuliers (page web, foire aux questions (FAQ), encyclopédie collaborative, bibliographie en ligne,

etc.) (Shepherd et al. 2004). Depuis récemment, le web est utilisé en tant que dispositif performant pour la création et la diffusion de corpus documentaires produits dans le cadre des différents processus et échanges organisationnels. De plus, ce média joue un rôle prépondérant dans l'amélioration de la compétitivité de divers organismes. Il contribue à maintenir la transparence de leurs activités, à offrir des services plus concurrentiels (services en ligne) et à faciliter la gestion des processus organisationnels (Frost 2001, Eschenfelder 2004, Eschenfelder 2005). Selon Statistiques Canada, 94,86% des organismes du secteur public possèdent un site web sur internet. D'après la même source, 82,48% des organismes du secteur public au Canada achètent des biens ou des services par internet et 15,60% de cette catégorie d'organismes en vendent (Statistiques Canada 2005).

Le web est considéré comme un outil fédérateur qui peut être utilisé par différents acteurs pour répondre à des besoins informationnels très variés ou pour accomplir des actions communicatives particulières. C'est une technologie en évolution continue et, surtout, incontrôlable, qui rend difficile l'anticipation de ses éventuelles conséquences sur les modes de communication ou les formes documentaires (Kallinikos 2005).

Le web et la production documentaire

Le web introduit avec lui trois nouveautés majeures qui traduisent sa logique et qui ont des incidences directes sur les caractéristiques du «document traditionnel». La première se conjugue dans la notion de connectivité entre plusieurs objets informationnels. Cette connectivité s'établit grâce à la possibilité de relier deux ou plusieurs fragments de texte (hypertexte) ou différents artéfacts technologiques : texte, image, vidéo, programme, etc. (hypermédia). Ces liens peuvent mener au même document (lien interne) ou vers d'autres documents (lien externe). De plus, des documents de nature hétérogène (contenu, format et structure) peuvent se restituer sous la même interface d'échange, voire se construire selon la volonté de l'utilisateur. La connectivité affecte profondément la structure des documents mis en circulation sur le web (diversité des formats, complexité, virtualité, instabilité, etc.), son mode de conception (dépendance par rapport à la technologie, fréquence des mises à jour, génération à la volée, personnalisation du format d'affichage, etc.), ainsi que celui de son appropriation (processus de lecture non linéaire, multimodalité, interactivité, etc.).

La deuxième nouveauté consiste en l'abolition des notions de temps et d'espace (ou de frontière). Il devient très ardu de circonscrire les contours spatio-temporels d'un document dans un environnement web (Koehler 1999). Tout d'abord, en s'appropriant les fonctionnalités du médium qui l'incorpore, le document sur le web devient de plus en plus polymorphe, dynamique, complexe, aux contours indéfinis et étanches. De plus, le document sur le web n'a pas d'existence physique. La dématérialisation de sa forme documentaire et sa connectivité avec d'autres objets informationnels favorisent sa délocalisation (duplication sur plusieurs serveurs, fusion ou liaison avec d'autres documents). Les documents qui sont produits, échangés ou déposés sur le web sont, désormais, accessibles à partir de plusieurs endroits de la planète (souvent en même temps) et à l'aide nombreux points d'entrée (adresse URL, titre, mot clé, etc.). Ensuite, comme la permanence des flux d'information représente la caractéristique principale du web, les contenus web sont constamment mis à jour. Ils se régénèrent, évoluent

et s'auto-alimentent sans cesse (flux RSS, webcams, etc.). La durée de vie moyenne d'une page ne dépasse pas deux mois (Gharsallah 2004, Bachimont et al. 2005). Enfin, plusieurs contenus sont souvent générés à la volée pour répondre à des requêtes spécifiques (pages web dynamiques) : ils n'existent que suite à l'interrogation d'une base de données en ligne et disparaissent une fois consultés. Cette permanence des flux et cette volatilité des contenus sont à l'origine, dans la plupart des cas, de la disparition de certains contenus à valeur patrimoniale, tels par exemple les premiers modèles de pages web.

L'interactivité constitue la troisième propriété du web. Cette technologie tend à éradiquer toute distinction entre l'acte d'écrire et l'acte de lire. Les deux processus peuvent se dérouler souvent en même temps. La malléabilité et la connectivité des contenus web facilitent l'interaction entre l'usager et le document « en construction ». Un document web s'entend, se lit, se voit et se construit simultanément. Son appropriation requiert alors la mise à contribution de plusieurs sens pour le lire et le comprendre, ce qui confère au document web son caractère multimodal (Ghitalla et Boullier 2004, Leleu-Merveil 2004). Par ailleurs, l'usager peut créer son propre document en choisissant un parcours de lecture particulier. Il possède en outre de plus en plus le moyen de contrôler le contenu et la structure des documents sur le web. En effet, l'usager peut, en tout temps, interrompre une séquence vidéo, modifier la mise en page (couleur et police de caractère, etc.), activer un lien interne ou externe, sélectionner un mode de navigation personnalisé, répondre à un message, insérer un commentaire, voire modifier le contenu d'un document (à l'intérieur des sites Wiki par exemple).

Il ressort de ce qui précède que le web bouleverse et remet en question aussi bien la morphologie et les formes temporelles du document que son mode d'appropriation. Cette réalité n'est pas sans conséquence sur les pratiques et les normes qui soutiennent la gestion et la conservation des documents sur le web (Foo et Lim 1997, Pédauque 2003, Gatenby 2004). Maîtriser l'évanescence de leurs contenus devient alors un souci majeur et l'objet de plusieurs travaux et projets de recherche.

Archiver le web : pourquoi ?

Le web, en plus d'être un moyen révolutionnaire de communication et un outil d'édition, forme l'un des principaux vecteurs de la culture des sociétés et la mémoire de ses différentes communautés. Les documents web constituent une partie intégrante de son patrimoine numérique. Toutefois, les trois principales caractéristiques du web révèlent sa nature instable et évolutive : « La logique du Web n'inclut pas la persistance ni la pérennité, mais l'échange et la reprise. » (Bachimont et al. 2005, 5). La mémoire de ce média est courte et éphémère (Gharsallah 2004). Il n'est pas dédié à la conservation et à la pérennité des formes documentaires qu'il génère : « Web sites make records, but they do not keep records in ways that match up to sound recordkeeping requirements » (Barry 2004, 27). Les sites web ne disposent pas encore des fonctionnalités ou des propriétés essentielles au maintien de la permanence et de l'intégrité des documents qu'ils véhiculent. En plus de ces facteurs d'ordre technique, il ne faut pas omettre la présence d'autres facteurs organisationnels, légaux, financiers et politiques qui concourent à l'accroissement de la vulnérabilité des documents web et à leur disparition (Harries 1999, Eschenfelder 2004, Cunningham et Phillips 2005, McDonald 2005).

Pour ces raisons, une action urgente doit être entreprise dans l'objectif de préserver une richesse documentaire originale (des contenus web) dans une forme accessible et de garder une trace intègre des activités, de même que la culture des acteurs qui en sont à l'origine (Engel 1999, Watson 1999, Gharsallah 2001, Jimerson 2003, Phillips 2003, Cunningham et Phillips 2005, Pennock et Kelly 2006). C'est dans ce contexte que s'inscrivent les efforts menés par plusieurs institutions pour collecter et préserver le web.

ARCHIVAGE DU WEB : LES PRINCIPALES INITIATIVES

Acteurs, approches et modèles d'archivage

Plusieurs projets d'archivage du web sont en cours. Ils se multiplient et se diversifient (Patrimoine canadien et Shearer 2001, Brügger 2005). Pour les décrire, nous adaptons la même catégorisation retrouvée dans la littérature, qui se base essentiellement sur les critères suivants : les acteurs impliqués, l'approche d'archivage et le modèle d'archivage (Lefurgy 2001, Lyman 2002, Beagrie 2003, Day 2003, Gharsalla, Monfort et Chaussard 2003, Couture et Khouaja 2003-2004, Gharsallah 2004, Bachimont et al. 2005, McDonald 2005).

Le premier critère (selon les acteurs impliqués) permet d'établir une comparaison au niveau de la portée des différentes initiatives. Dans la majorité des cas, les projets d'archivage du web sont réalisés par des organismes publics au niveau national, plus particulièrement par des institutions de mémoire (Bibliothèque nationale ou Archives nationales). Certains organismes privés assurent également l'archivage des documents qu'ils produisent ou publient (sites des maisons d'édition, journaux en ligne, etc.). D'autres projets sont réalisés grâce à la formation d'alliance entre plusieurs organismes ayant une influence au niveau régional ou international.

Le deuxième critère (selon l'approche d'archivage) permet de distinguer les projets qui ont pour objectif une collecte de l'ensemble des sites web (approche intégrale ou exhaustive) de ceux qui ne visent que le moissonnage d'un ensemble de sites selon des critères prédéfinis[4] (approche sélective) ou encore de ceux qui adoptent une approche dite par échantillonnage et qui ne retiennent que quelques sites web pour des besoins de traçabilité. Une autre approche que nous pouvons qualifier de mixte est en émergence. Elle se réalise en général en deux étapes au cours desquelles l'acteur utilise successivement l'approche exhaustive et l'approche sélective.

Le troisième critère (modèles d'archivage) indique le modèle utilisé pour aspirer, collecter ou sélectionner les sites web. Il s'agit de distinguer entre :

- Le modèle automatique qui repose sur l'usage des robots pour effectuer soit une collecte continue des sites web (selon la fréquence de leur mise à jour), soit une collecte à intervalles réguliers tel que défini par les responsables du projet, et
- Le modèle manuel qui se fonde sur deux actions principales : le dépôt légal ou l'obtention des autorisations de collecte auprès des auteurs des sites web.

Dans les approches mixtes, il y a souvent recours à ces deux modèles d'archivage : on parle dans ce cas de modèle semi-automatisé.

Dans la section suivante nous nous limiterons à l'examen de quelques initiatives représentatives des approches et modèles identifiés ci-dessus.

Quelques exemples représentatifs de différents projets d'archivage

Initiatives basées sur une approche exhaustive

Internet Archive représente la toute première initiative d'archivage et l'unique projet qui vise la collecte de l'intégralité du web mondial. Internet Archive a débuté en mars 1996 en tant que projet de recherche et s'est converti ensuite en un organisme à but non lucratif : la fondation Internet Archive (Gharsallah 2004). Cette initiative repose sur un modèle automatique de collecte des sites web sous forme d'« instantanés » (*snapshots*) collectés grâce au système Wayback Machine (voir Figure 1).

Cette méthode permet de constituer les différentes versions d'un site web[5]. L'entrepôt d'Internet Archive compte 55 milliards de pages web (dernière date de consultation : 24 octobre 2006). Le système Wayback Machine réunit plusieurs informations relatives aux taux de fréquentation du site, sa fréquence de mise à jour, le nombre de ses liens ainsi que les sites qui couvrent le même sujet. Toutefois, ce projet présente quelques points faibles. Tout d'abord, les sites web archivés sont collectés sans vérification, ce qui ne garantit pas la qualité ni la pertinence des sites aspirés. Ensuite, ces ressources informationnelles ne sont pas indexées et aucune mesure favorisant leur pérennité n'est communiquée à ce jour (Léger 2003, Gharsallah 2004, BnF 2006). En outre, le modèle d'archivage automatique ne permet que l'archivage de la partie visible du web c'est-à-dire celle qui est accessible par les robots : le web public. Les robots ne peuvent pas recueillir les sites sécurisés ou ceux qui forment une passerelle documentaire entre des bases de données et le poste client (web invisible). En ayant recours à ce modèle d'archivage, Internet Archive ne collecte pas en réalité tout le web mondial ; d'autres contenus d'importance ne sont pas capturés (Masanès 2002b). Enfin, la méthode utilisée par Internet Archive ne respecte pas les règles relatives au droit d'auteur ou à la propriété intellectuelle puisque la collecte des sites web n'a pas été autorisée par leurs auteurs (Dollar Consulting 2001, Muir 2001, Charlesworth 2003).

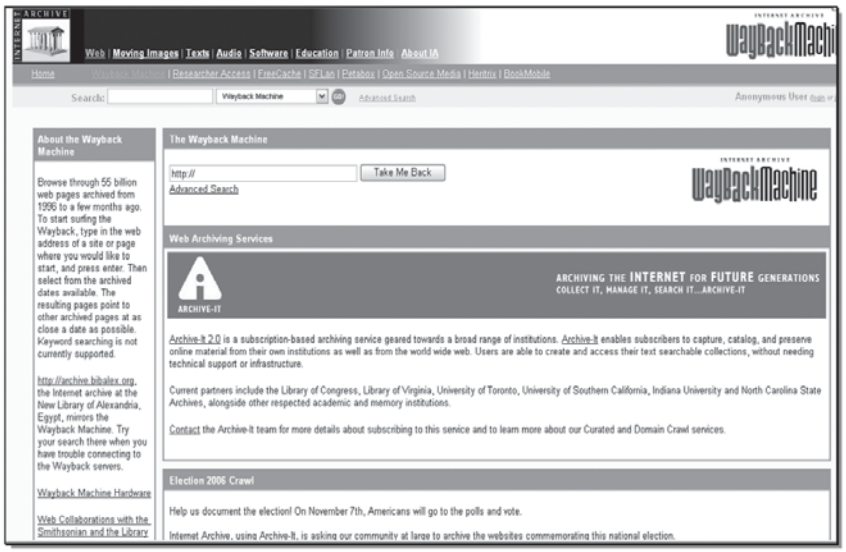

Figure 1. Capture d'écran de la page d'accueil d'Internet Archive (octobre 2006)

Malgré les limites de son approche, Internet Archive, en adoptant une démarche d'archivage proactive, apporte cependant son aide à plusieurs institutions de mémoire telles que la Bibliothèque nationale de France, la Library of Congress, afin de reconstituer les tout premiers corpus de sites web actuellement inaccessibles ou disparus.

L'approche intégrale est également adoptée par la Bibliothèque Royale de Suède depuis 1997 pour assurer une collecte à intervalles réguliers de l'ensemble des sites web du domaine suédois dans le cadre du Kulturarw3 Project (Arvidson et al. 2000)[6].

Initiatives basées sur une approche sélective

Deux variantes découlant de l'approche sélective sont présentement utilisées : l'approche sélective semi-automatisée et l'approche manuelle. La section suivante présente quelques exemples de projets faisant appel à ces deux types d'approche sélective.

Approche sélective semi-automatisée

La Bibliothèque nationale d'Australie et les Archives nationales d'Australie collaborent pour relever les défis inhérents à la préservation du patrimoine documentaire numérique australien. Les deux institutions ont lancé plusieurs initiatives tels que les projets PADI et PANDORA. PADI (Preserving Access to Digital Information) a pour objectif l'élaboration de lignes directrices pour la préservation de l'accès à l'information numérique. PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) vise le maintien de l'accessibilité des ressources documentaires australiennes en réseau (NAA and Commonwealth of Australia 2001, NLA 2002, Cunningham et Phillips 2005, McDonald 2005).

Les Australiens choisissent une approche sélective semi-automatisée qui vise la collecte, grâce au logiciel PANDAS, des publications en ligne et des sites web australiens les plus significatifs, ainsi que leur organisation au sein de PANDORA : « The National Library of Australia's PANDORA project has implemented a selective collection policy – Recommending Officers selected Web sites that have particular interest in Australia and decide the frequency of the collection separately for each site » (Library of Congress 2004). La sélection des sites web repose aussi bien sur leur valeur informative que sur leur portée culturelle. Certains sites sont sélectionnés par sujet (Cathro et al. 2001, NAA 2005).

PANDORA désigne à la fois le nom du réseau d'archives distribuées au sein duquel opèrent plusieurs bibliothèques et la base des « archives internet » australiennes collectées depuis 1996 (voir Figure 2). La collecte des sites web s'inscrit donc dans le cadre des activités de développement des collections (Beagrie 2003, Gatenby 2004, NLA and Partners 2006a). Plusieurs autres bibliothèques australiennes y participent et doivent déposer dans PANDORA les publications en ligne et les sites web dont elles se partagent l'acquisition, la description et le traitement. Elles peuvent toutefois conserver une copie des ressources collectées au niveau local (c'est un mode d'archivage à la fois centralisé et décentralisé[7]). Selon les statistiques du mois d'octobre 2006, la collection comprend environ 33 616 027 titres et réassemblages (ou extraits) de sites internet (NLA and Partners 2006b). Les titres sont collectés avec la permission de leurs éditeurs ou déposés volontairement par ces derniers puisque la loi sur le dépôt légal

du Commomnwealth australien ne s'applique pas sur les publications électroniques : «Because of the lack of legal deposit provisions covering online publications both at the national and State level, all PANDORA partners seek permission from publishers prior to copying publications and web sites into the Archive» (NLA and Partners 2006a).

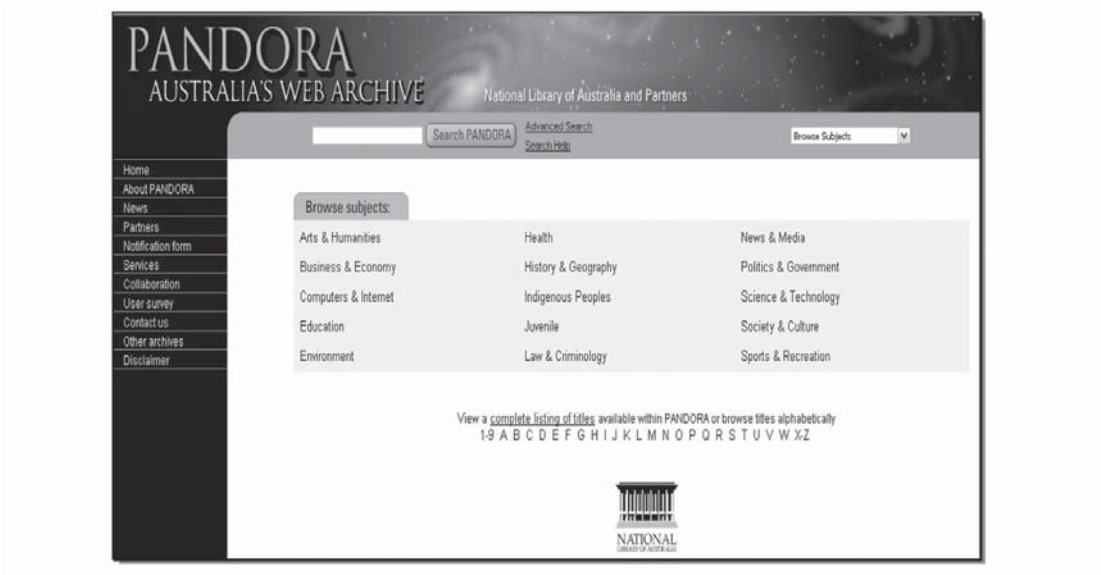

Figure 2. Capture d'écran de la page d'accueil du PANDORA (NLA octobre 2006)

En 2003, le programme *e-permanence* pour la préservation des documents d'archives numériques est lancé par les Archives nationales d'Australie. Ce programme vient consolider les efforts menés pour la conception et la mise en œuvre de systèmes de conservation des documents sous la garde des ministères et des organismes gouvernementaux du Commonwealth d'Australie (NLA 2002, NLA 2004, McDonald 2005, NLA and Partners 2006a). Le modèle choisi est décrit dans le rapport *An Approach to the Preservation of Digital Records* (NLA 2002) et repose essentiellement sur la normalisation du format de création des documents d'archives : «The National Archives' approach is based on the use of XML» (NLA 2002, 17).

Plusieurs autres ressources sont disponibles sur le site web des Archives nationales d'Australie, qui traitent spécifiquement de la gestion et de l'archivage des documents découlant des activités gouvernementales en ligne, dont *Archiving Web Resources : Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government*, un rapport publié en 2001 (NAA 2001). Ce dernier comprend un ensemble de principes directeurs et de conseils techniques qui guident les institutions gouvernementales vers les meilleurs choix lors de l'élaboration des stratégies de préservation des documents d'archives en provenance du web (NLA 2001).

Approche sélective manuelle

L'extension du champ d'application du dépôt légal au Québec pour inclure les documents électroniques mène la Bibliothèque nationale du Québec[8], en 2001, vers la mise en œuvre d'un programme de dépôt légal des publications gouvernementales en ligne. Dans une première phase, 750 titres existant dans la Banque des publications

gouvernementales diffusées sur le site internet du Ministère des Relations avec les Citoyens et de l'Immigration (MRCI) sont acquis. Lors de la deuxième phase (2002), la collection comprend 1250 titres. Entre 2003 et 2004, dans une démarche d'acquisition rétrospective et régulière, l'ensemble des titres diffusés par les ministères et organismes gouvernementaux au Québec est déposé (environ 6360 titres). Les fichiers à collecter doivent être signalés par l'organisme producteur et soumis au dépôt légal au début de leur vie «active» (Clapperton 2004). Pour ce faire, l'éditeur du contenu du site web doit remplir un formulaire particulier (formulaires de dépôt d'une monographie, d'un numéro de périodique, d'un périodique) (voir Figure 3). La publication est ensuite capturée directement sur le site web de l'éditeur une fois que ce dernier autorise cette action. L'équipe d'archivage assure la collecte des publications grâce aux logiciels MemoWeb et HTTrack. La consultation en ligne de ces ressources est possible à partir du Catalogue IRIS (Léger 2003, Clapperton 2004, Gharsallah 2004, Lupovici 2005, BAnQ 2005, BAnQ 2006, BnF 2006).

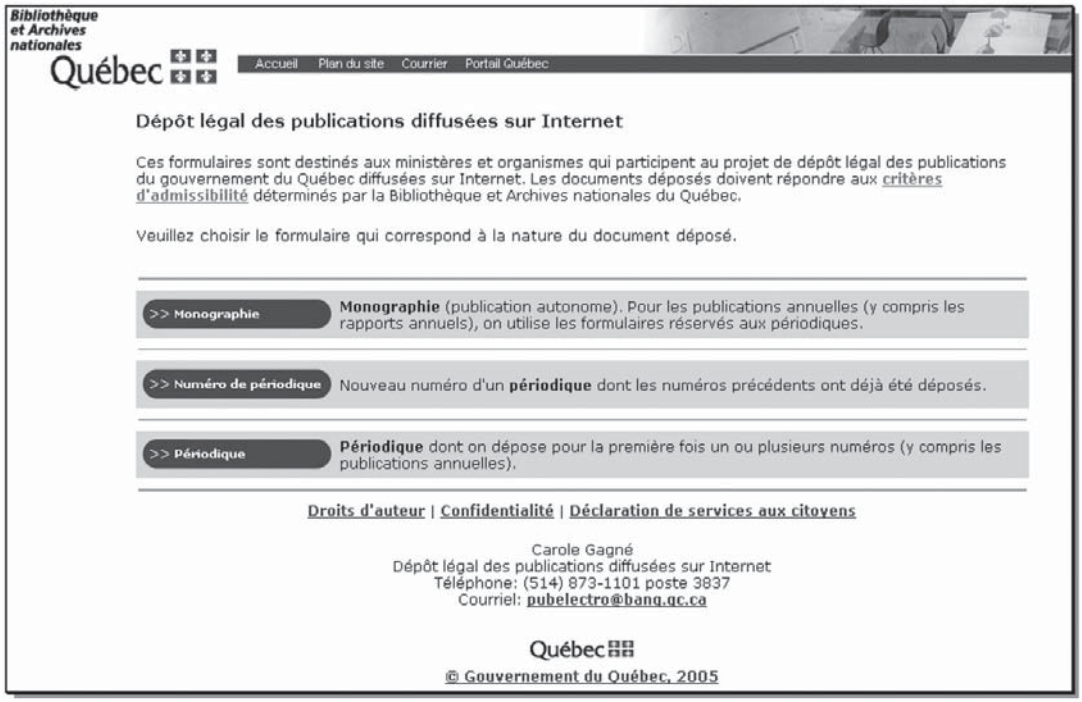

Figure 3. Capture d'écran des différents formulaires de dépôt légal des publications diffusées sur internet (BAnQ octobre 2006)

Entre 1994 et 1995, la Bibliothèque nationale du Canada[9] réalise dans le cadre d'un projet pilote sur les publications électroniques (PPPE) une collecte limitée d'une variété de ressources documentaires canadiennes diffusées sur internet. À partir de 1997, ces publications sont cataloguées dans AMICUS et figurent également dans Canadiana, la Bibliographie nationale. Une année après la création de la Section de l'acquisition des publications électroniques[10], la Bibliothèque nationale du Canada élabore ses propres *Politiques et directives relatives aux publications électroniques diffusées en réseau* (1998). De même, la capture des ressources documentaires est régie par une liste de critères

de sélection (voir Figure 4). La collection électronique de Bibliothèque et Archives Canada (BAC) est cotée en fonction de trois niveaux :

1. Archivé (publication intégrée dans la collection de la bibliothèque),
2. Versé (publication destinée à être intégrée dans la collection de la bibliothèque, mais faisant encore objet de négociation avec des éditeurs ou autres institutions d'archivage), ou
3. Lié (publication conservée sur un autre serveur).

Cette approche permet à Bibliothèque et Archives Canada (section bibliothèque) d'établir des priorités et des plans d'action dans la gestion et la préservation de sa collection électronique.

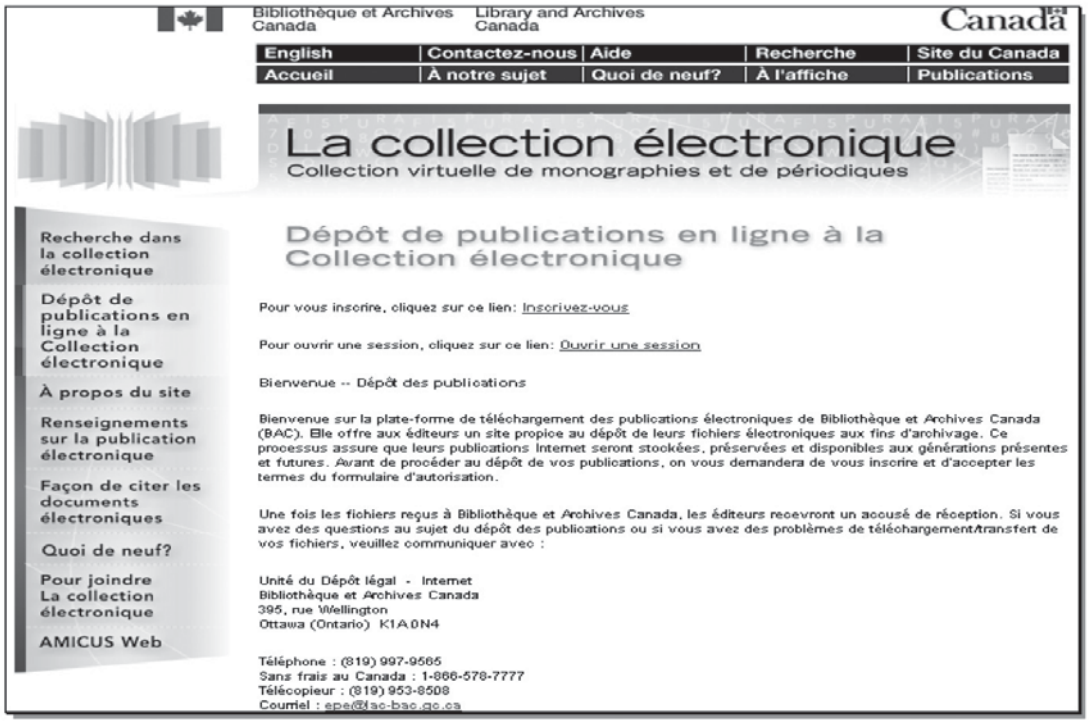

Figure 4. Capture d'écran de la page web relative au dépôt des publications en ligne (BAC octobre 2006)

L'archivage des publications en ligne devient une activité routinière de développement des collections chez Bibliothèque et Archives Canada ainsi que chez Bibliothèque et Archives nationales du Québec. Les deux institutions sont conscientes des limites de leur choix. Des travaux sont en cours qui visent l'extension de l'archivage des ressources documentaires sur le web en plus des publications en série et des publications gouvernementales (Clapperton 2004, BAnQ 2005, BAC 2006).

Initiatives basées sur une approche par échantillonnage

Archives des Premiers ministres des gouvernements français

Diverses versions des sites des gouvernements français[11] (voir figure 5) sont accessibles à partir de la page d'accueil du Portail du Gouvernement du Premier Ministre.

En effet, nous avons accès à une version du site du gouvernement du ministre Alain Juppé (1996-1997), à trois versions du site du gouvernement de Lionel Jospin (1997-2002) et, enfin, à deux versions du site du gouvernement de Jean-Pierre Raffarin (2002-2005). La capture des sites des Premiers ministres s'effectue une fois chaque deux ans.

Ces sites sont archivés en tant que copies statiques uniquement. Les instantanés sont réalisés en cas d'une refonte majeure du site ou lors de l'usage de nouvelles techniques ou technologies pour sa mise à jour. Ce choix est fait même si les versions datant de 2000 sont de nature dynamique, et ce, dans l'objectif de standardiser les moyens de conservation et d'accès aux différentes versions des sites web capturés (Gharsallah, Monfort et Chaussard 2003, Gharsallah 2004). Le site «www.archives.premier-ministre.gouv.fr» représente l'un des rares projets d'archivage qui visent l'acquisition et la conservation des traces d'activités d'un organisme gouvernemental. C'est l'une des premières tentatives d'archivage considérant l'unité documentaire archivée (le «site web») comme un document d'archives.

Figure 5. Capture d'écran des archives des différents sites des Premiers ministres français (octobre 2006)

Le projet Minerva

En 2000, la Library of Congress entame, dans le cadre du projet Minerva[12], la collecte d'un ensemble de sites web relatifs à des évènements, des thèmes, des sujets ou des domaines particuliers (Arms 2000, Lyman 2002, Léger 2003, Couture et Khouaja 2003-2004). En effet, dans sa politique sur la collecte et l'archivage des publications en ligne, la Library of Congress définit une collection de sites web archivés comme suit : «a group of any number of Web sites that Recommending Officers have selected based on task orders specific to a particular theme, either event-based, subject-oriented, or

domain-based (such as .gov or .edu).» (Library of Congress 2004, URL[13]). Entre août 2000 et janvier 2001, la bibliothèque collecte plus de 800 sites web (voir Figure 6). Les sites aspirés peuvent être recherchés par date, par titre ou par catégorie de site web.

La capture des sites obéit à des critères prédéfinis par l'équipe Minerva et un consultant interne. Par exemple :

- les sites doivent répondre aux besoins informationnels actuels et futurs du public desservi par la bibliothèque,
- ils doivent présenter une valeur informationnelle unique,
- ils doivent contenir des informations scientifiques ou académiques,
- ils doivent contenir des informations à valeur marchande,
- ils courent le danger de disparaître définitivement.

Un robot, Mercator web crawler[14], collecte des échantillons représentatifs de sites web tout en respectant les critères indiqués.

Grâce à ces efforts, la Bibliothèque du Congrès offre actuellement à ses utilisateurs une collection de plus de 36 000 sites web (américains et autres) relatifs essentiellement aux dernières élections présidentielles américaines (Élection 2000, Élection 2002), aux Jeux olympiques de 2002 et aux attentats du 11 septembre 2001.

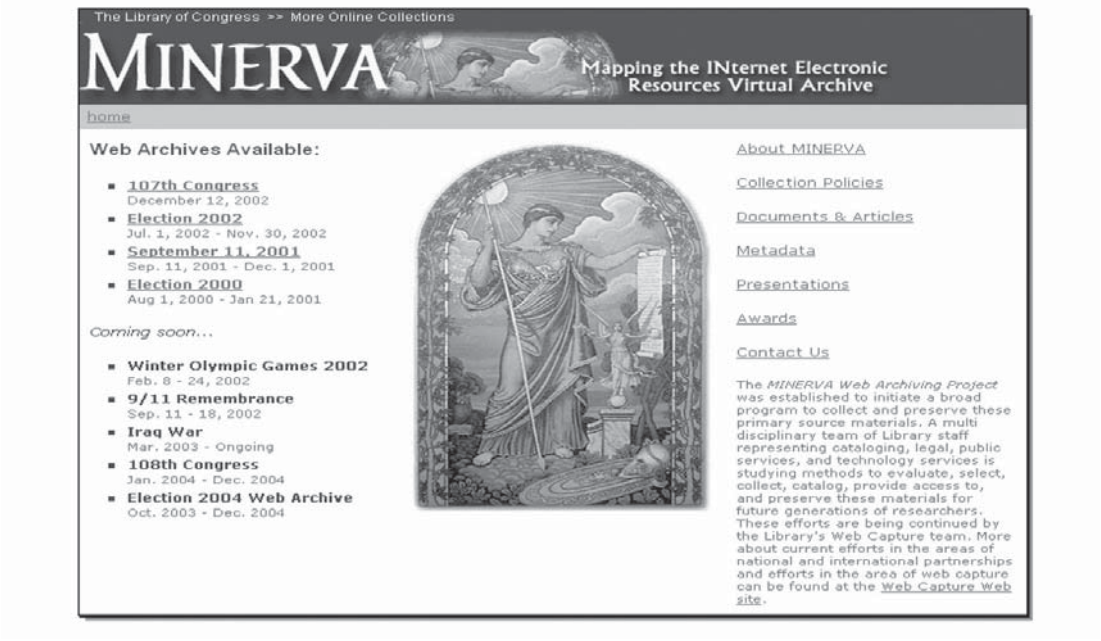

Figure 6. Capture d'écran de la page d'accueil de la collection Minerva (LC octobre 2006)

Initiatives basées sur une approche mixte

L'extension du dépôt légal aux objets informationnels communiqués au public par voie électronique pousse la Bibliothèque nationale de France (BnF) à réfléchir sur la pertinence de collecter le web français. En 1999, la BnF lance une série d'expérimentations visant l'évaluation des différentes approches d'archivage du web afin d'en établir le meilleur choix (Lupovici 2005, BnF 2006, Illien et Game 2006). Au cours

de la même année, la BnF teste la collecte automatique afin de réaliser en juin 2002 la capture d'instantanés de l'intégralité du domaine français. Deux autres instantanés du domaine français sont réalisés à la fin des années 2004 et 2005 au moyen du robot Heritrix et en partenariat de recherche avec la fondation Internet Archive. Ensuite, la BnF procède à la collecte sélective des sites web ayant pour thèmes les élections présidentielles et législatives de 2002 et les élections régionales et européennes de 2004. Entre 2001 et 2002, la BnF expérimente aussi le dépôt des sites web (voir Figure 7). Plusieurs ressources en ligne sont acquises, ce qui exige le développement et la mise en œuvre de méthodes et de procédures d'acquisition au moyen du dépôt légal internet (BnF 2006). Les sites collectés sont consultables sur place dans les salles de recherche de la BnF (Barthe 2005). Comme dans le cas du Canada, la loi française sur le dépôt légal prévoit, à cet égard, la possibilité de reproduire sur tout autre support et par tout autre procédé ces ressources documentaires, offrant ainsi les moyens nécessaires pour leur pérennité et leur exploitation (Haettiger 2003, Barthe 2005, BAC 2006).

Les expériences de la BnF mettent en évidence la nécessité d'une collaboration à l'échelle internationale, du partage des responsabilités à l'échelle nationale, et de la constitution d'une équipe de professionnels qui veille à l'émergence de nouvelles formes documentaires sur le web en plus d'être à l'affût des outils et techniques de collecte et de traitement les plus efficaces en la matière (Hakala 2004).

Ayant testé la plupart des approches, la BnF conclut qu'un modèle d'archivage fondé sur l'une ou l'autre des approches pratiquées n'est pas satisfaisant. L'institution arrête son choix sur une approche pragmatique et intégratrice. Cette approche se base sur l'aspiration à distance des sites web grâce à un logiciel robot spécialisé. Si ce dernier ne peut pas techniquement aspirer un site (ce qui risque de se produire dans le cas des sites web sécurisés ou de nature dynamique), il sera demandé à son éditeur de le déposer volontairement ou d'en communiquer les clés d'accès. L'approche adoptée par la BnF repose ainsi sur deux modèles d'archivage considérés complémentaires. Le premier permet de collecter automatiquement l'intégralité du web français (approche exhaustive automatisée). Le deuxième modèle cible la collecte du web invisible grâce, d'une part, à une démarche de dépôt volontaire par les éditeurs des ressources en ligne et, d'autre part, à une action ciblée des bibliothécaires une fois les autorisations nécessaires obtenues (approche sélective semi-automatisée). Cette démarche permet également de réaliser des collectes thématiques et assure ainsi la préservation des sites de nature éphémère (sites web relatifs à des évènements ou thèmes particuliers dont la durée de vie est limitée). La BnF applique, dans une phase finale, un processus d'évaluation qui détermine, dans la masse des sites collectés, ceux qui seront destinés à un archivage à long terme tout en vérifiant leur intégrité. Les sites sont sélectionnés sur la base de leur notoriété et de la qualité de l'information qui y figure. Le nombre des liens pointant vers un site web et l'analyse sémantique de son contenu forment les deux principales mesures de la qualité des sites web archivés à la BnF. Selon Gharsallah, cette approche «offre l'avantage de marier l'efficacité d'un repérage large et systématique avec la précision et le suivi d'une collecte manuelle» (Gharsallah 2004, URL). De plus, l'approche choisie par la France permet d'établir des priorités au niveau des opérations de collecte et du traitement. Certains contenus à risque, par exemple, ont fait l'objet d'une intervention urgente.

Pour garantir la qualité et l'exploitabilité de ses archives web, la BnF choisit de partager ses activités de développement des collections avec l'INA (Institut National d'Audovisuel), et ce, en fonction de la nature des contenus à collecter. En ce qui a trait aux contenus web, la BnF s'occupe de la collecte et de la sélection des objets informationnels classiques (livres en ligne, publications en série, etc.), tandis que l'INA est responsable de l'archivage des sites relevant de la communication audiovisuelle (Bachimont et al. 2005, Illien et Game 2006).

Figure 7. Capture d'écran de la page web sur le Dépôt légal internet (BnF octobre 2006)

LES PRINCIPALES MÉTHODES D'ARCHIVAGE

L'examen des tendances actuelles en matière d'archivage du web révèle l'existence de deux méthodes principales d'archivage, qui sont largement dépendantes de la nature des sites web ou des ressources en ligne à archiver. La première méthode ne permet que la capture d'images successives des sites web statiques. La deuxième méthode est plus complexe et cible la collecte de sites web de nature dynamique (web invisible).

La méthode des instantanés (*snapshots*)

La méthode des instantanées s'appuie sur la capture d'un grand nombre de copies des sites web. Les aspirateurs des sites web (*harvesters*) sont en mesure de rapatrier l'ensemble des pages web relatives à un site donné ainsi que le tissu de liens pointant vers les pages constitutives (Masanès 2000, Gharsallah 2004, Thomasen 2004). La fréquence des captures peut être établie en se basant sur la fréquence de mise à jour du site lui-même. En analysant les statistiques de mises à jour d'un site donné, le robot modélise son mode de publication et synchronise les captures des instantanés avec ses mises à jour réelles. Le cas du robot WayBack Machine d'Internet Archive en est un exemple (Bachimont et al. 2005). La fréquence des captures peut également être fixée par les membres de l'équipe d'archivage. La collecte peut être alors quotidienne,

hebdomadaire, mensuelle, annuelle, etc. Il va sans dire que, dans certains cas, la collecte des sites web selon des intervalles prédéterminés ne reflète pas l'évolution naturelle des sites aspirés (Gharsallah, Monfort et Chaussard 2003, Couture et Khouaja 2003-2004, Lupovici 2005, BnF 2006, Illien et Game 2006).

Dans un cas comme dans l'autre, la méthode des instantanés permet, dans un premier temps, de conserver des images fidèles de l'état d'un site web à un instant donné tout en sauvegardant l'intégralité des interactions (liens) et des animations présentes sur le site aspiré et, dans un deuxième temps, d'offrir une vue globale sur son évolution dans le temps, et ce, autant au niveau de son contenu, de sa structure que de ses fonctionnalités (voir Figure 8). En effet, la navigabilité parmi l'ensemble des pages constitutives du site web est maintenue en la recréant au niveau interne (Lupovici 2005). Il s'agit de transformer les liens absolus (URL) en des liens relatifs (PURL ou URN[15]), c'est-à-dire des liens permanents opérationnels au niveau local (Bättig 2005). Cette procédure s'avère primordiale puisque les différentes photographies du site sont conservées dans un entrepôt de sites web généralement déconnecté du réseau web. En outre, un système de signature cryptographique est utilisé afin de gérer diverses versions du site aspiré (Bachimont et al. 2005). Alors, en naviguant verticalement entre les différentes strates du même site, il devient facile de repérer les éléments de contenus ajoutés, modifiés ou mis à jour et d'étudier l'évolution du site d'une version à une autre (Gharsallah 2001).

Cette méthode d'archivage devient obsolète lorsqu'elle est appliquée à des sites web sécurisés (à accès limité) ou à des sites web agissant en tant que passerelle documentaire (sites donnant accès à des bases de données en ligne, sites ftp, etc.) Même si elle ne conserve pas le caractère dynamique des sites aspirés, cette méthode demeure une des solutions les plus pertinentes pour assurer une couverture exhaustive du périmètre à archiver. À titre d'exemple, elle permet de collecter des instantanées de tous les sites d'un domaine web particulier (fr, be, ca, uk, etc.) ou encore des sites liés à des évènements ponctuels (des élections présidentielles dans un pays donné).

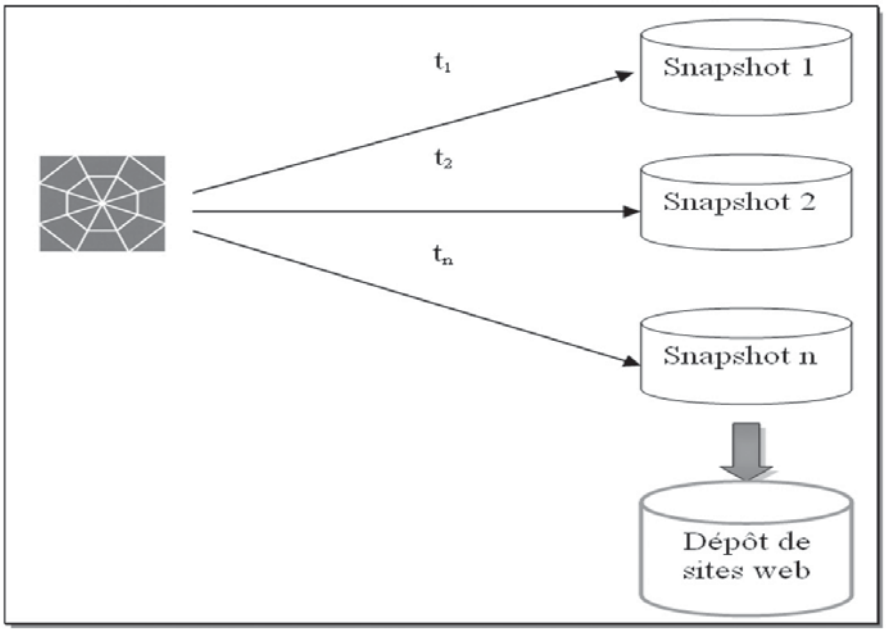

Figure 8. La méthode des instantanés (Chebbi 2006, adapté de Masanès 2002b)

La capture des sites web dynamiques

Par sites web dynamiques, nous faisons référence aux sites web dont le contenu est généré à la volée. Les documents sont stockés dans une base de données et ne s'affichent que suite à une requête bien définie (en remplissant un formulaire électronique, en tapant un mot clé). Un script ou un programme spécifique se charge alors de repérer le contenu correspondant à la requête et de le présenter sous une forme conviviale et personnalisée. Par sites web dynamiques, nous faisons également référence aux sites utilisés comme passerelles pour accéder aux contenus des bases documentaires de différents éditeurs et institutions (Masanès 2000, Masanès 2002a, Gharsallah 2004). Les modalités d'accès à ces contenus sont souvent restreintes et requièrent des mots de passe, une connexion en mode ftp ou IPv6[16] (catalogues de bibliothèques, bases de données documentaires, dépôts de documents institutionnels, sites payants, fichiers locaux, etc.).

Dans les deux cas de figure, le contenu de la page ou du site web n'existe pas en accès libre et immédiat. Il se construit au moment de la consultation ou suite à une procédure d'authentification. Les prototypes d'aspirateurs existants ne sont pas dotés de fonctionnalités permettant un archivage fiable de ce type de sites web. Ces derniers sont généralement aspirés sous forme statique (des instantanés), ce qui n'en reflète pas le vrai contenu ou la structure réelle et n'en respecte pas l'intégrité. Une portion assez importante des contenus documentaires du web profond reste toujours inaccessible pour les robots de moissonnage (Couture et Khouaja 2003-2004, Marill et al. 2004a, Marill et al. 2004b, Thomasen 2004).

Plusieurs groupes de recherche s'intéressent à l'étude de cette problématique, notamment le Consortium International de Conservation de l'Internet (IIPC)[17]. Établie en 2003, cette association entre Internet Archive et 11 institutions de mémoire[18] compte sur la participation active de ses membres pour avancer et approfondir la réflexion sur l'archivage et la conservation des ressources documentaires en ligne (Gatenby 2004, Masanès 2002b, Masanès 2004, McDonald 2005, BnF 2006, Illien et Game 2006). Les travaux de ce groupe débouchent sur trois techniques possibles pour l'acquisition des sites web dynamiques (voir Figure 9). La première solution repose sur une démarche volontaire de la part des éditeurs de sites. Les éditeurs sont responsables de l'archivage de leurs sites web et s'engagent à déposer une copie des contenus archivés auprès d'une institution de mémoire : *Client-side archiving* ou archivage côté client (ACC). L'organisme dépositaire doit gérer les conditions d'accès et veiller au respect des droits d'auteur et de propriétés intellectuelles des produits documentaires versés.

La deuxième possibilité consiste à recueillir tous les documents résultants des transactions et des activités en ligne : *Transaction archiving*, ou, archivage des pages web produites lors de diverses transactions en ligne. C'est une action difficile à mettre en œuvre. Souvent, il faut recueillir le résultat de chaque interrogation ou de consultation d'une base de données en ligne (effectuée par le même utilisateur ou par des utilisateurs différents en même temps). Le résultat affiché sur le site est toujours unique puisqu'il dépend de la requête ou du profil de l'utilisateur. De même, le contenu de la base est sujet à des mises à jour fréquentes. Autrement, il faut archiver toutes les combinaisons et toutes les versions possibles des contenus web. La société Project Computing propose un système qui enregistre toutes les pages générées par les

internautes : le système pageVault[19]. Ce système est efficace à condition que les pages à collecter soient demandées ou consultées au moins une fois (Fitch 2003).

La troisième solution semble être plus réaliste malgré les difficultés techniques qu'elle présente. Cette solution s'applique du côté du serveur : *Server-side archiving*. Le producteur du site communique les clés d'accès aux contenus du site où en autorise l'archivage. L'organisme responsable de l'archivage s'engage alors à collecter non seulement les documents et les données qui existent sur le serveur, mais également, les applications et les bases de données qui permettent de recréer le contenu, la forme et les fonctionnalités des pages web à archiver. Cette solution présente deux contraintes : l'une est relative à la désuétude rapide des applications informatiques, l'autre est liée à la question de l'archivage des bases de données. Ces dernières forment une problématique de taille pour l'archivage du web profond. Des travaux sont en cours sous l'égide de l'IICP visant l'amélioration d'un outil d'extraction et de transformation des bases de données en XML. Ces travaux servent, entre autres, à définir une architecture globale expliquant le mode de versement, d'indexation et d'accès aux ressources archivées. Ils portent, notamment, sur le développement des métadonnées de gestion et de préservation des éléments documentaires archivés (données, liens entre les tables, fonctionnalités des bases de données).

Pour qu'elles soient efficaces, ces trois solutions doivent être appliquées sur un petit nombre de sites. Elles sont onéreuses et lourdes à mettre en œuvre. Leur implantation requiert la collaboration entre différentes parties ainsi que la disponibilité de technologies de pointe et de spécialistes hautement qualifiés (des équipes multi et transdisciplinaires). Il faut reconnaître que les trois propositions avancées ne conservent pas la navigabilité entre les sites. Ces derniers sont archivés en tant qu'unités documentaires indépendantes. Par conséquent, certains éléments d'informations contextuelles sont perdus (Engel 1999). En plus, il demeure encore difficile d'établir un format d'archivage pérenne et commun à toutes les composantes d'un site, même si les robots actuels sont en mesure de reconnaître une grande variété de formats et de langages de balisage. Le champ est ouvert à de futures recherches sur l'interopérabilité et la normalisation des formats.

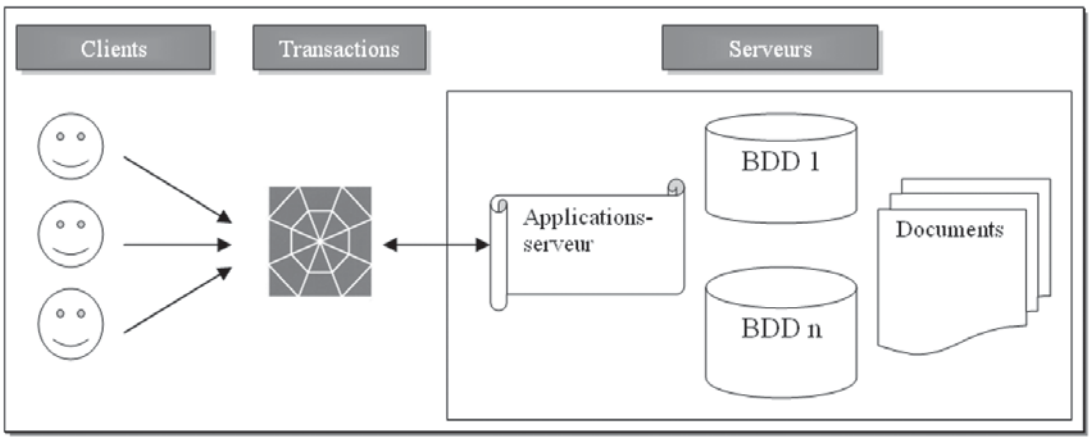

Figure 9. La capture des sites web dynamiques (Chebbi 2006, adapté de Masanès 2002b)

ARCHIVAGE DU WEB : DES LEÇONS À RETENIR

Le choix de l'approche la plus efficace pour archiver et pérenniser le web n'est pas une action facile à mener. Ce choix dépend de plusieurs facteurs instables, imprécis et souvent incontrôlables tels la nature des sites et des ressources en ligne à collecter, le périmètre à couvrir, les ressources technologiques, financières et humaines disponibles, le cadre légal et juridique, les finalités d'archivage, etc. En se référant à de nombreux rapports, il appert que les facteurs d'ordre organisationnel, technique et légal représentent les facteurs les plus imposants et constituent en même temps des défis de taille à surmonter (Masanès 2002a, Masanès 2004, Lyman 2002, Kavcic-Colic 2002, Jimerson 2003, NLA 2004, McDonald 2005, Haigh et Urban 2006).

Au niveau organisationnel, le champ des responsabilités des différents acteurs intéressés par l'archivage du web n'est pas évident à établir. En effet, sur le web coexistent des formes documentaires dont le sort d'acquisition, de gestion et de conservation n'est pas facile à déterminer (Association des archivistes suisses 2002). Même si la plupart de ces ressources sont largement diffusées et peuvent donc être soumises à la loi sur le dépôt légal, certains contenus sont de nature spécifique et leur gestion et leur conservation relèvent plutôt des compétences de gestionnaires des documents d'archives (documents web résultant des transactions en ligne, les sites web institutionnels, etc.) (Léger 2003).

Il est vrai que le web forme un média de convergence entre des documents de nature distincte. Malgré cela, il n'a pas de contours géographiques ou documentaires bien précis : il est difficile de se prononcer sur les frontières exactes d'un domaine web. De même, il n'existe pas encore de consensus sur la définition d'un site ou d'une page web. Il demeure encore délicat de préciser les limites spatio-temporelles d'un document sur le web, d'où l'inévitable redondance au niveau des contenus archivés. Bien que les chevauchements parmi le matériel collecté ne posent pas de contraintes majeures, il serait opportun de définir des territoires d'actions bien spécifiques aux niveaux national et international dans l'objectif de réduire les dépenses et de centrer les efforts sur l'acquisition de sites pertinents et viables. La réalisation d'un projet d'archivage laisse entrevoir des dépenses considérables et mettant à contribution des spécialistes de divers horizons (Lyman 2002, Day 2003, Phillips 2003).

Les contraintes au niveau technique sont, avant tout, liées à la nature et aux propriétés des sites web. Les acteurs du domaine de l'archivage du web sont contrariés par l'inefficacité des logiciels de moissonnage face aux sites web dynamiques. De surcroît, le processus de pérennisation des sites web aspirés grâce à la méthode des instantanés n'est pas sans faille. Des projets de recherche se déploient actuellement pour étudier et évaluer les techniques de préservation utilisées (essentiellement la migration et l'émulation), voire pour développer de nouvelles solutions. La question des métadonnées ainsi que des formats de conservation et de préservation retient l'attention de plusieurs groupes de recherche (Association des archivistes suisses 2002). Il est vrai que les initiatives d'archivage du web visent la préservation des ressources web et leur protection contre les assauts du temps et de la technologie, mais l'objectif de base est de garantir l'accès à ces ressources aux générations futures. L'accessibilité et l'exploitabilité forment un autre écueil d'ordre technique à surmonter. Beaucoup

d'efforts sont à investir dans les techniques de classification, d'indexation et de recherche des ressources collectées.

Quant aux contraintes légales rencontrées, elles consistent en ce que les sites web archivés ne sont pas tous en accès libre, faute d'un cadre réglementaire qui viendrait légaliser l'accès à ces ressources (Kavcic-Colic 2002, Charlesworth 2003). Seul un nombre restreint de pays est doté d'une réglementation qui autorise les institutions de mémoire (bibliothèques nationales, archives nationales ou toute autre institution de ce genre) à recueillir ces contenus et à en permettre l'accès (Muir 2001, Kavcic-Colic 2002, Lupovici 2005). Des contenus de valeur unique sont ainsi perdus à jamais ou demeurent inaccessibles pour une période indéterminée (Gharsallah 2001, Gharsallah 2004).

Dans le cas où ce type de législation existe, les outils et les techniques d'indexation et de recherche constituent un autre obstacle à surpasser. Un survol rapide des pages d'accueil et des prototypes des principaux projets de préservation du web permet de constater l'inexistence ou le peu de moyens et d'outils mis en place pour rechercher les contenus archivés (Beagrie 2003, Masanès 2004, McDonald 2005, Haigh et Urban 2006).

Il faut également retenir que chacune des approches actuellement en usage n'est pas satisfaisante. En effet, malgré l'« assurance » de collecter l'intégralité du périmètre fixé, l'approche exhaustive n'offre aucune garantie par rapport à l'intégrité et à la qualité des sites aspirés. En outre, cette approche ne s'inscrit pas dans un cadre légal ou réglementaire régissant les procédures de collecte, de traitement et de communication des richesses informationnelles collectées. Le volume des sites aspirés constitue de plus un obstacle majeur lors de l'indexation et de la classification de ces ressources, ce qui rend ardu tout procédé d'exploitation ou d'accessibilité (O'Neill et al. 2003, OCLC 2005, Gharsallah 2004, Marill et al. 2004a-b, Masanès 2004). L'approche exhaustive est d'autant moins satisfaisante qu'elle ne se réalise que dans le cadre de collectes massives et automatisées d'instantanés du web, augmentant ainsi le risque de passer à côté de ressources demeurant encore inaccessibles aux robots existants.

L'approche sélective ou par échantillonnage, souvent manuelle ou semi-automatisée, permet la constitution d'archives web de qualité, exploitables et accessibles[20]. Cependant, cette approche est très onéreuse et requiert du personnel hautement qualifié. Elle est la plus difficile à mettre en œuvre vu la lenteur des démarches de sélection ou d'échantillonnage (demandes d'autorisations, élaboration de critères de sélection et de procédures de collecte, etc.). Sa réussite dépend en grande partie du degré d'ouverture et de collaboration des éditeurs de sites, ainsi que de l'existence d'un cadre légal adéquat. Certains chercheurs critiquent l'approche sélective puisqu'elle se fonde sur l'extraction de fragments de contenus web ou de sites web isolés. Elle ne prend pas en considération tous les éléments contextuels (liens entre les sites, contenus à caractère commercial, certaines fonctionnalités du site) et ne permet pas la constitution de collections de sites web intègres et complets.

Les initiatives les plus récentes font appel à des approches mixtes afin de profiter des avantages et de limiter les inconvénients de chacune des approches en usage. De plus, la tendance actuelle s'oriente vers un archivage décentralisé partagé entre diverses institutions de mémoire. L'expérience de la Bibliothèque nationale de

France en démontre l'efficacité. Une collaboration à l'échelle internationale, la mise en commun des efforts et des expertises et une forte implication dans des groupes de recherche tels que le Consortium International de Conservation de l'Internet (IIPC) ou le Research Library Group s'avèrent nécessaires pour trouver des solutions plus rapides et efficaces, limiter les conséquences d'une évolution incontrôlable de l'objet à archiver et constituer des entrepôts d'archives plus interopérables, évitant ainsi une redondance documentaire inutile et le gaspillage de temps et de ressources.

CONCLUSION

Qu'il représente un phénomène social, culturel, documentaire ou autre, le web (ou du moins certains contenus diffusés sur le web) doit être archivé. L'enjeu d'une telle entreprise est de taille. Deux aspects liés à l'archivage des sites web sont aujourd'hui hautement problématiques, à savoir : l'acquisition ou la collecte des sites web, et leur conservation. Le web est un amalgame très complexe d'objets informationnels et d'acteurs dont les relations et les représentations sont instables, mouvantes, éphémères et faiblement normalisées. Les documents sur le web ne se définissent pas uniquement par leur contenu mais à travers plusieurs autres dimensions : fonctionnalité, structure, esthétique, etc. La capture et la préservation de l'ensemble de ces dimensions ne sont pas encore possibles. Des choix sont à faire et, par conséquent, des pertes sont à assumer.

L'examen de divers projets d'archivage et travaux sur la préservation du patrimoine numérique en ligne révèle deux pistes intéressantes pour surmonter les difficultés technologiques, techniques, légales et organisationnelles liées à l'accessibilité des publications en ligne. La première repose sur la maîtrise et le balisage du mode de conception des sites web. La deuxième consiste en une action proactive de la part des spécialistes d'information sur ces contenus (identification, gestion, conservation, etc.). À ce niveau, des acteurs clés tels que les archivistes sont encore peu impliqués dans la gestion et la préservation d'une portion importante du web, à savoir le web organisationnel. Pourtant, les archivistes disposent des outils et des compétences nécessaires pour maîtriser l'évanescence des contenus web et mettre en place un cadre efficace régissant leur gestion et leur conservation.

L'analyse de la littérature démontre que l'identification, l'évaluation et la conservation des contenus ayant une valeur archivistique, créés ou déposés sur le web ne sont pas encore des pratiques courantes. Certes ces contenus ont acquis des caractéristiques particulières. Toutefois, leur processus de gestion et d'archivage ne semble pas a priori différer des documents produits dans un environnement non réseauté. L'application et l'adaptation des instruments archivistiques tels le plan de classification, le calendrier de conservation et d'autres outils de gestion documentaire ainsi que l'adoption de formats normalisés ou ouverts faciliteront sans doute la rationalisation de la création et de la gestion du web organisationnel.

Aïda Chebbi Candidate au doctorat à l'École de bibliothéconomie et des sciences de l'information (EBSI) de l'Université de Montréal

NOTES

1. HTTP : Hypertext Transfer Protocol. Le protocole HTTP représente un ensemble de règles et de commandes qui définissent comment envoyer et recevoir les informations entre le client et le serveur. Chaque fichier relié à Internet doit disposer de sa propre adresse URL qui contient trois indications : le type de protocole de transfert des données (http, ftp), le nom de la machine hébergeant le fichier et le nom du fichier lui-même.
2. URL : Unified Resources Locator.
3. Electronic Word of Mouth : nouvelle forme de sites web dédiés à la publication des opinions de consommateurs sur un produit donné.
4. Bachimont et ses collaborateurs en dressent une liste exhaustive : les critères linguistiques (sites en langue suédoise), territoriaux (le domaine fr.), thématiques (éducation) et évènementiels (11 septembre, élections présidentielles, jeux olympiques, etc.) (Bachimont et al. 2005).
5. Nous expliquerons cette méthode dans la section 3.
6. Adresse URL du projet : http://www.kb.se/kw3/
7. Même mode de fonctionnement en Grande Bretagne : six institutions de mémoire au Royaume-Uni, membres du UK Web Archiving Consortium, sont responsables d'une collecte sélective (par thème ou sujet) des sites web britanniques. À titre d'exemple, la British Library est responsable de l'acquisition des sites présentant un intérêt culturel, historique et politique. La Wellcome Library doit acquérir les sites dans le domaine médical. Ces institutions doivent obtenir au préalable l'autorisation des éditeurs des sites web. Les Archives nationales assurent l'archivage de plus de 80 sites web gouvernementaux capturés depuis septembre 2003 qui forment l'« UK Government Web Archive ». Des *snapshots* ou instantanés de ces sites web gouvernementaux sont réalisés périodiquement (chaque semaine ou chaque période de 6 mois) et sont présentement accessibles sur internet (Public Records Office 2001, Brown 2003, Day 2003, Waller et Sharpe 2006).
8. La Bibliothèque nationale du Québec est actuellement fusionnée avec les Archives nationales du Québec.
9. La Bibliothèque nationale du Canada est actuellement fusionnée avec les Archives nationales du Canada.
10. En 2005, la section de l'acquisition des publications électroniques est devenue l'Unité du Dépôt légal-Internet.
11. Adresse web à consulter : http://www.archives.premier-ministre.gouv.fr/home_ie.htm
12. Adresse web à consulter : Mapping the Internet : the Electronic Resources Virtual Archive. http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html
13. URL signifie ici une référence sur le web.
14. Développé par la compagnie Compac SRC.
15. PURL pour Permanent Unified Resource Locator, URN pour Uniform Resource Name : c'est un identificateur unique et permanent qui remplace un URL et qui est utilisé comme une référence fiable et stable. Le remplacement des URLs par des PURLs facilite la mise à jour automatique des références, permet de réduire le temps requis pour cette mise à jour et garantit un accès durable à l'objet archivé (Bättig 2005).
16. IP version 6 : Le protocole Ipv6 permet d'interrelier des ordinateurs et autres appareils (téléphones mobiles, appareils photos, etc.) connectés à des sous-réseaux importants ou des usagers particuliers (institutions universitaires, par exemple). Une adresse Ipv6 se compose de 39 caractères permettant d'identifier le sous-réseau et l'hôte à l'intérieur du réseau.
17. IICP : International Internet Preservation Consortium. http://netpreserve.org
18. L'IICP est coordonné par la Bibliothèque nationale de France. Il comprend la Bibliothèque du Congrès, la British Library et les bibliothèques nationales d'Australie, du Canada, du Danemark, de la Finlande, d'Islande, d'Italie, de Norvège, de Suède ainsi que la fondation Internet Archive.
19. Adresse web à consulter : Project computing. pageVault. http://www.projectcomputing.com/products/pageVault/
20. Archives est ici synonyme d'entrepôts de pages web.

BIBLIOGRAPHIE

ARMS, W. Y. 2000. Collecting and Preserving Open-Access Materials on the Web : A Proposal to the Library of Congress from Cornell University. [En ligne]. http://www.cs.cornell.edu/wya/LC-web/proposal.html

ARVIDSON, A., K. PERSSON et al. 2000. *The Kulturarw3 Project – The Royal Swedish Web Archiw3e – An example of «complete» collection of web pages*. 66th IFLA Council and General Conference, 13-18 August, Jerusalem, Israel. [En ligne]. http://www.ifla.org/IV/ifla66/papers/154-157e.htm

ASSOCIATION DES ARCHIVISTES SUISSES (AAS). 2002. *Archivage des documents électroniques dans l'administration publique – Perspectives et besoin d'actions 2002–2010 : Étude stratégique globale pour la conservation à long terme des documents électroniques en Suisse. Rapport de synthèse*. Berne. [En ligne]. http://www.pwc.ch/user_content/editor/files/publ_public/pwc_archivage_documents_f.pdf

BACHIMONT, B., T. DRUGEON et al. 2005. Documenter et partitionner une archive du web : vers le dépôt légal d'un domaine média. ICHIM 05 – Digital Culture & Heritage / Patrimoine & Culture Numérique, 21-23 septembre, Paris, Bibliothèque nationale de France. *Archives & Museum Informatics Europe*. [En ligne]. http://www.archimuse.com/publishing/ichim05/Bachimont.pdf

BARRY, R. 2004. Web sites as recordkeeping & «recordmaking» systems. *The Information Management Journal* (novembre/décembre) : 26-32.

BARTHE, E. 2005. Dépôt légal des pages web : bientôt! Les archives des sites web français seront consultables au rez-de-jardin de la BnF. *Precisement.org : un blog pour l'information juridique*. [En ligne]. http://www.precisement.org/blog/article.php3?id_article=3

BÄTTIG, Y. 2005. *Politique URN de la Bibliothèque nationale suisse*. Bibliothèque nationale suisse. [En ligne]. http://www.snl.admin.ch/slb/dokumentation/normen_und_regelwerke/index.html?lang=fr&download=M3wBPgDB/8ull6Du36WcnojN14in3qSbnpWVZGqbnE6p1rJgsYfhyt3NhqbdqIV+bay9bKbXrZ6lhuDZz8mMps2go6fo

BEAGRIE, N. 2003. *National Digital Preservation Initiatives : An Overview of Developments in Australia, France, the Netherlands, and the United Kingdom and of Related International Activity*. Library of Congress. [En ligne]. http://www.clir.org/pubs/reports/pub116/contents.html

BIBLIOTHÈQUE ET ARCHIVES CANADA (BAC). 2006. *Renseignements sur les publications électroniques : Consignes pour archiver une publication HTML en ligne*. [En ligne]. http://www.collectionscanada.ca/collectionelectronique/003008-300-f.html

BIBLIOTHÈQUE ET ARCHIVES NATIONALES DU QUÉBEC (BAnQ). 2005. *Dépôt légal des publications diffusées sur Internet*. [En ligne]. http://sgdl.banq.qc.ca/EXTRANET/depotlegal/dl.html

BIBLIOTHÈQUE NATIONALE DE FRANCE (BnF). 2006. *Dépôt légal Internet : les étapes du projet*. [En ligne]. http://www.bnf.fr/pages/infopro/depotleg/dli_intro.htm

BROWN, A. 2003. *Preserving the digital heritage : building a digital archive for UK Government records*. Online Information 2003 Proceedings. [En ligne]. http://www.nationalarchives.gov.uk/documents/brown.pdf

BRÜGGER, N. 2005. *Archiving Websites. General Considerations and Strategies : General Considerations and Strategies*. The Centre for Internet Research. [En ligne]. http://www.cfi.au.dk/publikationer/archiving/.pdf

CATHRO, W., C. WEBB et al. 2001. *Archiving the Web : The PANDORA Archive at the National Library of Australia*. Preserving the Present for the Future Web Archiving Conference, 18-19 June, Copenhagen. [En ligne]. http://www.nla.gov.au/nla/staffpaper/2001/cathro3.html

CHARLESWORTH, A. (2003). *Legal issues relating to the archiving of Internet resources in the UK, EU, USA and Australia : A study undertaken for the JISC and Wellcome Trust*. Version 1.0. Centre for IT and Law. University of Bristol. [En ligne]. http://www.jisc.ac.uk/uploaded_documents/archiving_legal.pdf

CLAPPERTON, M. 2004. *L'implantation du dépôt légal des publications diffusées sur Internet du gouvernement du Québec. L'usager : ses exigences et nos perspectives*. 31e congrès de l'ASTED, 25-27 octobre, Québec. [En ligne]. http://www.banq.qc.ca/documents/a_propos_banq/nos_publications/a_rayons_ouverts/ASTED_dl_internet.ppt

COUTURE, C. et B. KHOUAJA. 2003-2004. La gestion et l'archivage des sites Web institutionnels. *Archives* 35, 3/4 : 17-41.

CUNNINGHAM, A. et M. PHILLIPS. 2005. Accountability and accessibility : ensuring the evidence of e-governance in Australia. *Aslib Proceedings New Information Perspectives* 57, 4 : 301-317.

DAY, M. 2003. *Collecting and preserving the World Wide Web : A feasibility study undertaken for the JISC and Wellcome Trust*. Version 1.0. UKOLN, University of Bath. [En ligne]. http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

DENNIS, A. R. 1998. Lessons from Three Years of Web Development. *Communications of the ACM* 41, 7 : 112-113.

DOLLAR CONSULTING. 2001. *Archival preservation of Smithsonian Web resources : strategies, principles, and best practices*. Smithsonian Institution Archives. [En ligne]. http://www.si.edu/archives/archives/dollar%20report.html

ENGEL, S. 1999. *Context is everything : the nature of memory*. New York, W.H. Freeman.

ESCHENFELDER, K. R. 2004. Behind the Web site : an inside look at the production of Web-based textual government information. *Government Information Quarterly* 21 : 337-358.

ESCHENFELDER, K. R. 2005. *The openness of government websites : toward a socio-technical government website evaluation toolkit*. [En ligne]. http://iis.

syr.edu :8380/dspace/bitstream/2291/19/1/Eschenfelder%2520Miller%2520Institutions-Government.pdf

FITCH, K. (2003). *Web site archiving – an approach to recording every materially different response produced by a website*. AusWeb03. The thirteenth Australasian Word Wide Web Conference, 5-9 July. [En ligne]. http://ausweb.scu.edu.au/aw03/papers/fitch/

FOO, S. et E. P. LIM. 1997. Managing World Wide Web publications. *Asian Libraries* 6, 3/4 : 166-177.

FROST, J. P. 2001. Web technologies for information management. *The Information Management Journal*, octobre : 34-37

GATENBY, P. 2004. *Collecter et gérer les ressources Internet pour un accès pérenne : moissonnage et directives pour l'aide à la conservation* (Actions de l'ICABS 3.3 et 3.4). World Library and Information Congress. 70th IFLA General Conference and Council, 22-27 August. Buenos Aires, Argentina. [En ligne]. http://www.ifla.org/IV/ifla70/papers/026f_trans-Gatenby.pdf

GHARSALLAH, Mehdi. 2001. Pour que la mémoire ne flanche pas. *Archimag* 145 : 28-29.

GHARSALLAH, Mehdi. 2004. *Dépôt légal des publications électroniques et préservation patrimoniale du web français*. [En ligne]. http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/13/11/sic0000131101/sic00001311.pdf

GHARSALLAH, Mehdi, Jérome MONTFORT et Audrey CHAUSSARD. 2003. *Panorama mondial de l'archivage du web*. Cellule Wall-On-Line – Ministère de la Région Wallonne. [En ligne]. http://archivesweb.wallonie.be/apps/spip/IMG/pdf/SC0335_09b.pdf

GHARSALLAH, Mehdi, Jérome MONFORT et Audrey CHAUSSARD. 2004. *Typologie des sites web*. Cellule Wall-On-Line – Ministère de la Région Wallonne. [En ligne]. http://egov.wallonie.be/docs/egov_rw/archivage/Typologie.pdf

GHITALLA, F. et D. BOULLIER. 2004. Le web ou l'utopie d'un espace documentaire. *Revue I3* 4, 1 : 173-189.

HAETTIGER, M. 2003. Vers la conservation des sites web régionaux. *BBF* T48, 4 : 77-84.

HAIGH, S. et M. URBAN. 2006. *Bâtir une infrastructure de conservation numérique (Ébauche- le 2 mai 2006). Vers une stratégie canadienne sur l'information numérique*. Bibliothèque et Archives Canada.

HAKALA, J. 2004. Archiving the web : European experiences. *Electronic Library and Information System* 38, 3 : 176-183.

HARRIES, S. 1999. *Capturing and managing electronic records from websites and Intranets in the government environment*. DLM Forum '99. European citizens and electronic information : the memory of the Information Society, 18-20 octobre, Bruxelles. [En ligne]. http://ec.europa.eu/archives/ISPO/dlm/fulltext/full_harr_en.htm

ILLIEN, G. et V. GAME. 2006. Le dépôt légal d'Internet à la Bibliothèque nationale de France : Cadre juridique, modèle de collecte, évolutions des métiers. *BBF*, 3 : 82-85.

JIMERSON, R. C. 2003. Deciding what to collect. *OCLC Systems & Services* 19, 2 : 54-57.

KALLINIKOS, J. 2005. The order of technology : complexity and control in a connected world. *Information and Organization* 15, 3 : 185-202.

KAVCIC-COLIC, A. 2002. *Archiver le Web : quelques perspectives juridiques*. 68th IFLA Council and general conference, August 18-24, Glasgow. [En ligne]. http://www.ifla.org/IV/ifla68/papers/116-163f.pdf

KOEHLER, W. 1999. An analysis of Web page and Web site constancy and permanence. *Journal of the American Society for Information Science* 50, 2 : 162-180.

LEFURGY, W. G. 2001. Records and Archival Management of World Wide Web Sites. *The Newsletter of the Government Records Section of the Society of American Archivists* 2 (Avril). [En ligne]. http://www.mybestdocs.com/lefurgy-w-grn0104.htm

LÉGER, D. 2003. *L'implantation du dépôt légal des publications diffusées sur Internet à la Bibliothèque nationale du Québec* : (Texte d'accompagnement de la présentation PowerPoint DLPI20030620.ppt.) Bibliothèque nationale du Québec. Section du dépôt légal. [En ligne]. http://www.banq.qc.ca/documents/ressources_en_ligne/pgq/DLPI20030620.pdf

LELEU-MERVIEL, S. 2004. Effets de la numérisation et de la mise en réseau sur le concept de document. *Revue I3* 4, 1 : 121-140.

LIBRARY OF CONGRESS. 2004. *Collections policy statement : Web capture & archiving*. [En ligne]. http://www.loc.gov/acq/devpol/webarchive.html

LUPOVICI, C. 2000. Les stratégies de gestion et de conservation préventive des documents électroniques. *Bulletin des Bibliothèques de France* 45,4 : 43-54.

LUPOVICI, C. 2005. *La collecte automatique du web : l'expérience de la Bibliothèque nationale de France*. World Library and Information Congress : 71th IFLA General Conference and Council. *Libraries – A voyage of discovery*, August 14th – 18th. Oslo, Norway. [En ligne]. http://www.ifla.org/IV/ifla71/papers/074f_trans-Lupovici.pdf

LYMAN, P. 2002. *Archiving the World Wide Web. Building a National strategy for digital preservation : issues in digital media*. Washington, Library of Congress. [En ligne]. http://www.clir.org/pubs/reports/pub106/pub106.pdf

MARILL, J., A. BOYKO et al. 2004a. *Web Harvesting Survey*. Version 1. In Site des Netpreserve.org, International Internet preservation consortium. [En ligne]. http://netpreserve.org/publications/iipc-r-001.pdf

MARILL, J., A. BOYKO et al. 2004b. *Tools and Techniques for Harvesting the World Wide Web*. Proceedings of the Joint ACM/IEEE Conference on Digital Libraries (JCDL'04), June 7-11. Tucson, Arizona, États-Unis. [En ligne]. http://ieeexplore.ieee.org/iel5/9280/29473/01336207.pdf

MASANÈS, J. 2000. *L'archivage des sites Internet*. Rapport de stage, Diplôme de conservateur de bibliothèques, École nationale supérieure des sciences de l'information et des bibliothèques, Département de la bibliothèque numérique de la Bibliothèque nationale de France. [En ligne]. http://www.enssib.fr/bibliotheque/documents/dcb/rsmasanes.pdf

MASANÈS, J. 2002a. Towards Continuous Web Archiving : First Results and an Agenda for the Future. *D-Lib Magazine* 8, 12. [En ligne]. http://www.dlib.org/dlib/december02/masanes/12masanes.html

MASANÈS, J. 2002b. *Préserver les contenus du Web*. 4ème journées internationales d'études de l'arsag, Paris, Bibliothèque nationale de France. [En ligne]. http://bibnum.bnf.fr/conservation/migration_web.pdf

MASANÈS, J. 2004. Report on the 4th International Web Archiving Workshop (IWAW), 16 September, Bath, United Kingdom. *D-Lib Magazine* 10, 11. [En ligne]. http://www.dlib.org/dlib/november04/masanes/11masanes.html

McCLURE, C. R. et J. T. SPREHE. 1998. *Guidelines for electronic records management on state and federal agency websites*. The National Historical Publications and Records Commission. [En ligne]. http://slis-two.lis.fsu.edu/~cmcclure/guidelines.pdf

McDONALD, J. 2005. *Initiatives internationales concernant l'information numérique : Revue des initiatives internationales pertinentes*. Bibliothèque et Archives Canada. [En ligne]. http://www.collectionscanada.ca/scin/012033-400-f.html

MUIR, A. 2001. *Legal deposit of digital publications : a review of research and development activity*. Joint Conference on Digital Libraries 2001. Roanoke, Virginia.

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION (NARA). 2005. *NARA Guidance on Managing Web Records*. [En ligne]. http://www.archives.gov/records-mgmt/pdf/managing-web-records-index.pdf

NATIONAL ARCHIVES OF AUSTRALIA (NAA). 2001. *Archiving Web Resources : Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government*. [En ligne]. http://www.naa.gov.au/recordkeeping/er/web_records/archweb_guide.pdf

NATIONAL ARCHIVES OF AUSTRALIA (NAA). 2005. *Online Australian Publications : Selection Guidelines for Archiving and Preservation by the National Library of Australia*. [En ligne]. http://pandora.nla.gov.au/selectionguidelines.html

NATIONAL ARCHIVES OF AUSTRALIA (NAA) et COMMONWEALTH OF AUSTRALIA. 2001. *Archiving Web Resources : A Policy for Keeping Records of Web-based Activity in the Commonwealth Government*. Government Recordkeeping Standards and Policy. [En ligne]. http://www.naa.gov.au/recordkeeping/er/web_records/archweb_policy.pdf

NATIONAL ARCHIVES OF AUSTRALIA (NAA) et COMMONWEALTH OF AUSTRALIA. 2001. *Archiving Web Resources : Guidelines for Keeping Records of Web-based Activity*. Government Recordkeeping Standards and Policy. [En ligne]. http://www.naa.gov.au/Images/archweb_guide_tcm2-903.pdf

NATIONAL LIBRARY OF AUSTRALIA (NLA) and PARTNERS. 2006a. *PANDORA Australia's web archive. Legal deposit*. [En ligne]. http://pandora.nla.gov.au/legaldeposit.html

NATIONAL LIBRARY OF AUSTRALIA (NLA) and PARTNERS. 2006b. *PANDORA archive size and monthly growth. Statistics as at 26 October 2006*. [En ligne]. http://pandora.nla.gov.au/statistics.html

NATIONAL LIBRARY OF AUSTRALIA (NLA). 2004. *Themes emerging from Archiving Web Resources : Issues for Cultural Heritage organisations* (Canberra, 9 – 11 November). [En ligne]. http://www.nla.gov.au/webarchiving/

NATIONAL LIBRARY OF AUSTRALIA (NLA). 2002. *Managing web resources for persistent access*. [En ligne]. http://pandora.nla.gov.au/pan/36282/20030701/www.nla.gov.au/guidelines/2000/persistence.html

NATIONAL OFFICE FOR THE INFORMATION ECONOMY. 2002. *Keeping Government Publications Online : A Guide for Commonwealth Agencies. Why is it important to ensure long-term access to on-line publications?* National Archives of Australia. [En ligne]. http://www.nla.gov.au/guidelines/govpubs.pdf

OCLC ONLINE COMPUTER LIBRARY. 2005. *Size and growth statistics*. [En ligne]. http://www.oclc.org/research/projects/archive/wcp/stats/size.htm

O'NEILL, E. T., B. LAVOIE et al. 2003. Trends in the Evolution of the Public Web 1998 – 2002. *D-Lib Magazine* 9, 4. [En ligne]. http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

PATRIMOINE CANADIEN et K. SHEARER. (2001). *Préservation de l'information numérique, Ressources utiles : Organismes internationaux clés possédant des ressources liées à la préservation des documents numériques*. [En ligne]. http://www.chin.gc.ca/Francais/Contenu_Numerique/Preservation/ressources.html

PÉDAUQUE, R. T. 2003. *Document : forme, signe et medium, les re-formulations du numérique*. [En ligne]. http://archivesic.ccsd.cnrs.fr/docs/00/06/21/99/PDF/sic_00000511.pdf

PENNOCK, M. et B. KELLY. 2006. Archiving Web Site Resources : A Records Management View. *WWW 2006*, May 23-26. Edinburgh, Scotland. International World Wide Web Conference Committee (IW3C2). [En ligne]. http://www.dcc.ac.uk/docs/WWW2006_Archiving_Web_Site_Resources.pdf

PHILLIPS, J. T. 2003. The challenge of web site records preservation. *The Information Management Journal* janvier/février : 42-48.

PUBLIC RECORD OFFICE. 2001. *Managing web resources : Management of electronic records on websites and intranets : an ERM toolkit*. Version 1.0. [En ligne]. http://www.nationalarchives.gov.uk/documents/website_toolkit.pdf

SHEPHERD, M., C. WATTERS, et A. KENNEDY. 2004. Cybergenre : Automatic identification of home pages on the web. *Journal of Web Engineering* 3, 3-4. [En ligne]. URL :http://users.cs.dal.ca/~shepherd/pubs/JWE040722.pdf

STATISTIQUES CANADA. 2005. *Utilisation des technologies de l'information et des communications par les entreprises et les gouvernements (Entreprises qui possèdent un site Web sur Internet)*. [En ligne]. http://www40.statcan.ca/l02/cst01/econ146c_f.htm

THOMASEN, B. H. 2004. *Tests of software and strategies for micro-archiving websites*. Centre for Internet Research, University of Aarhus. [En ligne]. http://www.cfi.au.dk/publikationer/archiving/test.pdf

WALLER, M. et R. SHARPE. 2006. *Mind the gap : Assessing digital preservation needs in the UK*. Digital Preservation Coalition. [En ligne]. http://www.dpconline.org/docs/reports/uknamindthegap.pdf

WATSON, I. 1999. Internet, intranet, extranet : managing the information bazaar. *Aslib Proceedings* 51, 4 : 109-114.