Mémoire d'étude Janvier 2003

Diplôme de conservateur de bibliothèque

L'archivage des sites Web d'intérêt régional

HAETTIGER MAGALI

Sous la direction d'Elisabeth Noël
Conservateur des Bibliothèques, enseignante à l'Enssib

Remerciements

Je tiens à remercier Elisabeth Noël pour son aide, ainsi qu'Yvette Weber, responsable de la documentation régionale à la Bibliothèque municipale de Lyon, Xavier Lenoir, responsable du service informatique dans la même bibliothèque et Annie Garden, responsable de la coordination bibliographique. Je tiens également à remercier Julien Masanès, chef de projet pour le dépôt légal des sites Web à la BnF.

Sommaire

Résumé

La conservation de tout document numérique se heurte à deux problèmes majeurs : celui de la fragilité des supports et celui de l'obsolescence technique. L'archivage de sites Web pose, en outre, le problème de la sélection et de l'acquisition des sites, au sein d'une offre pléthorique de documents. Aujourd'hui, plusieurs solutions sont envisagées mais sont toutes développées par de grandes bibliothèques nationales. Quel peut être alors le rôle d'établissements plus modestes dans ce type d'archivage ?

Mots-clès

Sites Web ** Conservation et restauration

Archivage électronique

Abstract

The Archiving of any electronic resource is doubly doubtful: Firstly because of the fragility of the CD's and floppy disks and secondly because of the technical obsolescence. Furthermore, in the case of the archiving of Web sites, it is difficult to select some relevant documents in all the Internet. Today, some solutions exist but they only are developed by national libraries. Then, which kind of role could play the regional libraries in the archiving of the Internet?

Keywords

Web sites ** Conservation and restoration

Electronic filing system

Introduction

Le tout premier mèl reçu dans le monde, le tout premier site Web conçu ne seront jamais exposés dans une bibliothèque ou un musée. La raison de cet écueil n'est pas liée au peu de valeur de ces objets qui appartiennent dorénavant à l'Histoire : Le courrier électronique, comme l'Internet ont à ce point modifié les méthodes de travail et de communication de nos sociétés que la valeur de ces objets prototypes, aujourd'hui, est attestée. La raison pour laquelle personne ne pourra jamais consulter ces documents historiques est des plus simples : Ils n'ont pas été conservés !

Cet aspect peut paraître anecdotique. L'analyse du premier site Web ne nous apporterait peut-être aucune réelle information. Mais qu'en est-il des millions de pages Web produites chaque année ? De nombreux sites Web renferment aujourd'hui des informations qui n'ont aucun équivalent sur support papier. Le Web constitue un média relativement peu coûteux de diffusion et d'édition. De ce fait, il tend à devenir un moyen d'expression largement diffusé au sein de groupes informels, d'associations, de particuliers mais aussi de grandes institutions. A ce titre, la place de l'Internet peut être comparée au rôle qu'a joué au XIX° siècle, la presse. Reflet des opinions, des débats de l'époque et des représentations, la presse constitue un matériel irremplaçable pour connaître une époque et l'on ne saurait envisager aujourd'hui que la presse ne puisse pas faire l'objet d'une conservation.

La mise en parallèle de la presse et de l'Internet n'est pas aussi fortuite qu'il n'y paraît. La presse, comme les sites Web, n'a pas pour fonction première d'être conservée. Le journal, produit dans un papier de qualité exécrable, une fois lu est généralement jeté. Le site Web, de la même façon, n'est pas conçu pour être gardé : il évolue, connaît des versions multiples, change d'adresse et disparaît. Pourtant, au même titre que la presse ancienne, qui, grâce aux efforts des bibliothèques, constitue un témoignage important, les sites Web seront des sources irremplaçables pour quiconque s'intéresse à notre époque.

Nous le savons, la conservation de documents sur supports papier de mauvaise qualité pose problème. Dans le cas des sites Web, ces problèmes de support sont démultipliés. S'y rajoutent des problèmes d'accessibilité ultérieure liés à la structure informatique des sites. Comment alors envisager la conservation des sites Web sur le long terme ? Comment conserver des documents qui évoluent si rapidement et disparaissent à un rythme moyen de six semaines après leur création ? Cette tâche titanesque ne doit-elle être confiée qu'à de grandes bibliothèques nationales ?

En vue d'apporter une réflexion à ces questions, nous allons commencer par aborder les problèmes que pose l'archivage des documents électroniques en général et des sites Web en particulier. Cette première partie devrait permettre d'évaluer l'ampleur de la tâche que

constitue la mise en place d'une politique de conservation de sites. Dans un deuxième temps, nous présenterons les solutions qui sont envisagées à l'heure actuelle par les établissements engagés dans des projets de ce type. Enfin, nous tenterons d'envisager la place que pourraient tenir, au sein du projet d'archivage du Web français, des bibliothèques de taille plus réduite, comme les BMVR ou de grandes bibliothèques municipales. Peut-on ainsi concevoir que ces bibliothèques puissent développer, au niveau local, des projets d'archivage de sites Web régionaux ? Quelle réflexion serait nécessaire pour mettre en œuvre un tel projet ? Les annexes vous permettront d'approcher plus précisément cette dernière question puisque nous y développons l'ébauche d'un projet d'archivage pour la Bibliothèque municipale de Lyon. Dans les annexes, vous trouverez également un tableau plus précis des projets d'archivage en cours en France et dans le monde à travers quelques exemples significatifs. Enfin, l'évaluation des différentes solutions envisagées à l'heure actuelle ainsi que deux exemples de métadonnées vous sont présentés en annexe, comme outils d'aide à la décision.

Partie 1 :Les problèmes généraux liés à la conservation à long terme des sites Web

1. Les problèmes posés par la conservation des documents numériques en général…

L'ère du document numérique est aussi celle de la reproductibilité infinie des documents. Non seulement, la copie d'un document numérique sur un cédérom, un disque dur ou une disquette est simple et rapide, mais encore rien ne permet de différencier la copie obtenue de l'original. Pourtant, rien ne nous garantit que la copie sur cédérom d'un texte écrit sous Word ou celle sur DVD d'un film, pourra être accessible et visible dans cinq ans. Bien plus, tout nous porte à croire que ces mêmes copies ne seront absolument pas accessibles dans cinq ans. C'est ici que se situe sans doute le grand paradoxe de l'ère numérique : Alors que nous sommes entrés dans une ère de reproductibilité infinie de l'information, celle-ci a rarement été aussi fragile.

Les sites Web, en tant que documents numériques, sont touchés par cette fragilité. Celle-ci est véritablement au cœur du problème de la conservation des sites Web et de tout document numérique. Cette fragilité est intrinsèquement liée à la structure même de n'importe quel document numérique. De façon à expliquer pourquoi les sites Web, comme les cédéroms ou les DVD ne seront pas accessibles à long terme, nous allons donc commencer par rappeler quelques éléments sur la structure des documents numériques en général.

1.1. La structure générale d'un document numérique

Contrairement au livre dans lequel le support et le contenu sont intrinsèquement liés, la structure d'un document numérique comporte plusieurs niveaux qui induisent une séparation entre le support (le cédérom, le disque dur…) et le contenu informationnel. Un document numérique, quel qu'il soit, comporte plusieurs « couches »[1] qui, au final, rendent le document accessible et compréhensible par le lecteur:

---

[1] Nous utilisons ici la dénomination de Christine Lupovici In **LUPOVICI, Catherine.** Les besoins et les données techniques de préservation. In *67th. IFLA council and general conference. August 16th-25th. 2001.* [En ligne]
http://www.ifla.org/IV/ifla67/papers/163-168f.pdf

- La couche « physique »

Cette couche du document correspond aux informations relatives au support physique ou de communication du document numérique. Cette information prend la forme d'un format généralement normalisé. C'est ainsi que pour un cédérom le format sera la norme ISO 9660. De ce fait, dans le cas d'un changement de support, par exemple d'un cédérom vers un DVD, ce format correspondant à la couche physique sera modifié.

- La couche « binaire »

Cette couche correspond à l'unité fondamentale d'information du document numérique. Elle se présente sous la forme d'une suite de 0 et de 1, le « 1 » correspondant, au niveau de la machine, à une impulsion électrique de + 5 volts, le « 0 » à – 5 volts.

- La couche « structure »

Ces données binaires sont assemblées selon une structure primitive interprétable par des langages de programmation

- La couche « objet »

Correspond à une transformation des données structurées en objets signifiants pour la couche suivante. Les données au niveau de la couche objet sont régies par un format particulier (JPEG, HTML etc…).

- La couche « application »

Les objets transmis par la couche objet sont alors manipulés par des logiciels, des applications en vue d'être présentés à l'utilisateur. Notons que, dans certains cas, il peut y avoir une correspondance exacte et unique entre l'application et le format (par exemple entre le format PDF et l'application « Acrobat Reader »). Dans d'autres cas, un même format peut être présenté par plusieurs applications, c'est le cas du format JPEG.

Le contenu d'un document numérique est constitué par cet ensemble de données codées. L'altération d'une partie impliquerait donc une perte d'intégrité du document, perdant ainsi, par exemple, des fonctionnalités mais peut-être aussi son contenu informationnel. Ainsi, la conservation unique des codes binaires de l'objet ne signifie-t elle pas forcément un accès ultérieur au document lui-même : on ne saurait se contenter de ce mode de conservation.

Dernière consultation le 07/11/02

1.2. L'obsolescence des techniques

Un document numérique est donc constitué de cet ensemble de couches. Chaque couche nécessite, pour être traitée par l'ordinateur et, *in fine*, être compréhensible pour le lecteur, la médiation d'équipements, de langages, de normes, de logiciels…Ces médiations successives (système d'exploitation,logiciels, applications, périphériques…) conditionnent la consultation du document numérique par l'utilisateur. Cette dépendance technique de l'utilisateur qui constitue une contrainte importante, devient une difficulté majeure dans le cadre d'une politique de conservation des documents numériques. En effet, on estime que le cycle de validité des programmes et des périphériques est de l'ordre de 2 à 5 ans[2]. Passée cette période, un certain nombre de documents ne seront plus accessibles.

En vue d'illustrer ce problème, Julia Martin et David Coleman[3] ont développé l'exemple suivant : Un chercheur commence son travail en 1988 sur un ordinateur IBM 286. Il sauvegarde ses fichiers sur des disquettes de 5 pouces. Dans les années 1990, ce même chercheur fait l'acquisition d'un IMAC, ordinateur qui ne comporte qu'un lecteur de cédéroms. Le chercheur souhaitant consulter ses archives de 1988 sera donc dans l'impossibilité de le faire.

Plus concrètement, les Archives nationales des Etats-Unis ont déjà été confrontées à ce type de difficultés. En 1976, les archives prirent conscience que les enregistrements informatiques du recensement fédéral de 1960 avaient une valeur historique et leur conservation fut donc décidée. Mais le matériel de lecture des fiches perforées n'existait déjà plus. La sauvegarde de ce recensement ne put donc se faire qu'au prix d'un immense travail de restauration.

Les exemples développés ci-dessus n'approchent que l'aspect matériel du problème puisqu'ils ne traitent que d'un problème d'équipement. Mais le problème se pose également au niveau d'une couche plus abstraite de la lecture du document, au niveau des logiciels et des applications nécessaires. Ainsi, la compatibilité ascendante entre les versions successives d'un logiciel peut être assuré, mais ce n'est pas toujours le cas. Par exemple, une nouvelle version de l'application du traitement de texte Word de Microsoft apparaît en moyenne tous les trois à cinq ans. Or ces nouvelles versions ne sont directement compatibles qu'avec la version immédiatement précédente.

---

[2] Chiffres donnés dans la plupart des articles cités en bibliographie.

[3] **MARTIN, Julia et COLEMAN, David.** The archive as an ecosystem. In *Michigan University*. [En ligne]
http://www.press.umich.edu.jep/07-03/martin.html
Dernière consultation le 19/10/02

La fragilité d'un objet numérique face au vieillissement technologique est surtout liée à la dépendance de cet objet à un format, à une application. Or, les formats peuvent être :

- propriétaires ouverts c'est-à-dire définis par une entreprise privée mais dont les spécifications sont ouvertes, connues.
- propriétaires fermés c'est-à-dire définis par une entreprise privée et dont les spécifications sont tenues secrètes par le propriétaire.
- Standards c'est-à-dire produits par un organisme de normalisation (AFNOR, ISO, W3C). Ils sont alors ouverts et non propriétaires.

D'une façon générale, la fragilité des documents est beaucoup plus importante dans le cas où l'objet dépend d'un format ou d'une application propriétaire. En effet, le format PDF (Portable Document Format) est lisible grâce au logiciel gratuit « Acrobat Reader » de la société Adobe. Le format PDF est propriétaire ouvert. Rien ne permet de dire si l'entreprise Adobe ne désirera pas à un moment faire payer son logiciel de lecture PDF - Comme elle le fait déjà pour « Acrobat Writer » qui permet de générer du PDF- et fermer le format PDF dont elle est propriétaire.

D'un point de vue général, il est donc possible de dire que la mise en place d'une politique d'archivage de documents numériques, quels qu'ils soient, nécessite de prendre en considération les aspects techniques liés à l'environnement de chaque document. Mais également, une telle politique ne saurait faire abstraction du contexte économique général dans lequel les équipements et les formats informatiques sont produits.

1.3. La conservation des supports

La question de la conservation des supports est sans doute celle qui rejoint le plus directement les problématiques rencontrées par toute bibliothèque. Cependant, dans le cas qui nous occupe, la faible durée de vie des supports d'information numérique renforce ces problématiques, pourtant classiques en bibliothèques. La question des supports d'archivage est fondamentale dans le cas des sites Web qui devront être enregistrés sur supports et conservés.

Il existe trois types de supports d'information numérique : Les supports magnétiques (les disquettes par exemple), les supports optiques (les cédéroms par exemple) et les supports magnéto-optiques (le Minidisk par exemple). La capacité de stockage des supports est de plus en plus importante. Ainsi le CD commercialisé dès 1983 a une capacité de stockage de 1,2 GO, le DVD commercialisé depuis 1995 peut contenir 4,7 GO.

Pour les CD, les constructeurs annoncent une longévité de 75 à 200 ans selon la nature des composants du disque. Des tests semblent montrer que les CD enregistrables gravés sont plus fragiles que les CD pressés. Loin de l'optimisme des constructeurs, les durées de

vie moyennes annoncées par des laboratoires de recherche pour les CD enregistrables sont de l'ordre de 5 ans contre 10 à 25 pour les CD pressés.

Pour le moment, il n'existe aucune donnée stable sur la durée de vie des supports électroniques, mais la moyenne annoncée ne dépasse que très rarement les dix années. Il semble bien qu'alors que les capacités de stockage des supports électroniques augmentent d'année en année, réduisant d'autant ces mêmes coûts de stockage, la fragilité des supports, elle, augmente.

Comme tout type de support, les conditions environnementales de conservation sont importantes. Ainsi, la norme ISO/CD 16111 sur les conditions de stockage des disques optiques prévoit des conditions variables selon les couches supérieures de substrats composant les disques[4]. Selon la Digital Preservation coalition[5] la température de conservation des supports constitue un facteur important de vieillissement :

Device	25RH 10°C	30RH 15°C	40RH 20°C	50RH 25°C	50RH 28°C
D3 magnetic tape	50 years	25 years	15 years	3 years	1 year
DLT magnetic tapecartridge	75 years	40 years	15 years	3 years	1 year
CD/DVD	75 years	40 years	20 years	10 years	2 years
CD-ROM	30 years	15 years	3 years	9 months	3 months

Figure n°1 : la longévité des supports informatiques en fonction de la température (Tableau présenté par la Digital Preservation Coalition sur son site Web :

[4] Ainsi, à titre d'exemple, les CD comprenant des substrats de verre gravé sont plus résistants puisqu'ils supportent des températures variant de –5°C à 50°C et un taux d'humidité relative variant de 5 à 95%. Au contraire, les conditions de conservation des CD comprenant des substrats en Polymère organique oscillent entre 5 à 20°C et doivent être comprises entre 30 et 50% d'humidité relative.

[5] **DIGITAL PRESERVATION COALITION.** Media and formats . In *Digital Preservation coalition's Website*
http://www.dpconline.org/graphics/medfor/media.html
Dernière consultation le 29/10/02

http://www.dpconline.org/graphics/medfor/media.html#media1. Dernière consultation le 29/10/02

Ce tableau ne saurait faire office d'autorité unique en ce qui concerne la longévité des supports de documents numériques. Par contre, il permet de percevoir rapidement au moins trois éléments essentiels :

- La longévité varie relativement peu en fonction du type technologique du support : c'est ainsi qu'un support magnétique comme les cassettes DLT a la même longévité que le DVD qui est un support optique
- C'est surtout la qualité du support qui semble jouer un rôle déterminant dans la longévité de celui-ci. C'est ainsi qu'un simple cédérom dont la couche supérieure n'est qu'un vernis acrylique verra son espérance de vie largement réduite par rapport à un DVD.
- Enfin, les conditions environnementales de conservation et, parmi elles, la température apparaissent comme un facteur décisif de dégradation des supports.

Pour résumer, la conservation de tout objet numérique se heurte à une double difficulté, la première concerne le rythme d'obsolescence des techniques qui rend hasardeux l'accès ultérieur à l'objet et la seconde provient de la fragilité des supports électroniques sur lesquels l'objet sera archivé. L'archivage des sites Web, puisqu'il s'agit de documents numériques, pose ces problèmes d'obsolescence et de fragilité des supports. Cependant, du fait de leur complexité et du contexte dans lequel ils sont créés, les sites Web posent des problèmes de conservation particuliers.

2. La conservation des sites Web : des problèmes liés à leur complexité

2.1 Des documents multimedias

En tant que documents numériques, les sites Web sont des objets particulièrement complexes. Ils sont en effet des objets intégrant plusieurs types de documents : des textes, des images, des fonctionnalités, parfois du son et des animations. L'ensemble est structuré par des liens hypertextuels, généralement en HTML. Chaque type de document inclus dans un site Web intègre donc des formats différents. C'est ainsi qu'un site Web peut comporter des images en format graphique GIF, des documents textuels en .DOC et des animations au format Flash le tout structuré en HTML.

Selon Peter Lyman[6], une page Web contiendrait en moyenne quinze liens à d'autres pages et cinq objets différents (images, sons ou autres).
Ainsi, les problèmes posés par la dépendance technique aux formats pour tout type de document numérique sont-ils amplifiés dans le cas des sites Web du fait même de leur complexité et de l'intégration de plusieurs formats.

2.2 Des documents instables

L'aspect, la forme du site Web que consulte l'utilisateur peut varier selon le type de navigateur Web qu'il utilise, sa version, mais également selon les caractéristiques de son ordinateur. Cette forme mouvante du document nous interroge sur l'accès à la forme originelle de celui-ci. Ce problème risque de prendre une importance grandissante avec le développement de nouveaux langages comme le XML. En effet, dans le HTML la structuration du document mêle à la fois le contenu et la forme, associant ainsi obligatoirement une structure à une représentation. Avec le XML, le contenu structuré et la présentation des données sont distincts, permettant ainsi des changements formels importants des documents et une plus grande adaptation aux besoins de l'utilisateur. Or, dans le cadre d'une conservation à long terme, la forme des documents peut également avoir une importance puisqu'elle exprime un choix ou une norme stylistique... Ce problème concerne plus généralement ce que l'on nomme Web dynamique et qui recouvre l'ensemble des sites Web dont une partie est générée dynamiquement par le visiteur. Dans le cas d'un site Web dynamique, quelle version, considérée alors comme une sorte d'original, doit-on alors conserver ?
Par ailleurs, le site Web est un document particulièrement instable. Certes, tous les sites Web ne sont pas mis à jour quotidiennement, mais pour autant, il ne faut pas envisager un site Web consulté comme un objet terminé. Le site Web pourrait presque être comparé à un périodique sans périodicité fixe. Ainsi, nous en parlerons plus précisément, apparaît-il illusoire de vouloir conserver toutes les versions de tous les sites Web sur un sujet, même précis, ou sur un territoire donné. Mais alors combien de versions faut-il conserver d'un site Web ? A partir de quel moment considère- t on qu'une modification est à ce point significative qu'elle peut donner lieu à un nouvel enregistrement ?
Outre cette variabilité, un site Web est aussi un document éphémère. Selon l'étude commandée par OCLC[7], la durée de vie moyenne d'un site Web serait de six semaines.

---

[6] **LYMAN , Peter.** Archiving the World Wide Web. In *Clir.* [En ligne]
http://www.clir.org/pubs/reports/pub106/web.html
Dernière consultation le 07/10/02

Enfin, nous faisons chaque jour l'expérience de la variabilité des emplacements des sites Web qui peuvent changer d'URL et de serveurs.

2.3 Les limites de l'objet à archiver

Nous l'avons vu précédemment, en tant qu'objet numérique, le site Web est formé de plusieurs couches correspondant pour chacune d'entre elles à un certain niveau d'abstraction du document. Ainsi, la dimension physique du document sous la forme d'impulsions électriques issues de suites binaires ne constitue pas l'ensemble du document, contrairement au livre imprimé dont l'étendue totale coïncide avec l'objet physique « livre ». De ce fait, la question de l'archivage des sites Web ou de toute forme d'objet numérique correspond à un choix fondateur : à quel niveau d'abstraction de l'objet décide-t-on de commencer l'archivage et plus concrètement, quelle partie du document souhaite-t-on ou plutôt doit-on conserver ? Par exemple souhaite-t-on conserver la présentation du document ? Faut-il conserver également l'ensemble des fonctionnalités du document (outils de recherche, outils de navigation…). Il est bien sûr souhaitable et même nécessaire de conserver, tant que faire se peut, tous ces aspects. Cependant, il faut savoir que, plus l'on se place à un haut niveau d'abstraction du document, intégrant ainsi toutes les couches logiques du document (binaire, structure, objet, application), plus les conditions de sa conservation sont complexes ne serait-ce que parce que l'on intègre dans la conservation plusieurs formats, plusieurs périphériques, plusieurs logiciels.

La variabilité des sites Web renvoie également au problème de la limite de l'objet « site Web » et à sa définition. En effet, puisque le site Web est un objet mouvant, dont le contenu ou la présentation peut varier à tout moment, où se situent alors les limites de celui-ci ? Certainement pas dans sa première version, ni dans son ultime version. Les limites de ce site Web seraient alors constituées par l'ensemble des versions et des modifications qu'a connues ce site.

Enfin, il ne faudrait pas oublier que l'Internet est un réseau et qu'à ce titre, chaque site Web en est à la fois un élément et un vecteur. De façon imagée on pourrait presque dire qu'un site Web est à la fois un espace et un chemin. En effet, les sites Web, pour la plupart, renvoient à d'autres sites par le biais de liens hypertextes. Le fait de renvoyer l'utilisateur à un autre site est un choix du créateur du site et constitue donc un élément du site. Par ailleurs, ces liens hypertextes participent de la navigabilité d'Internet et font partie des utilisations de l'Internet. Faut-il alors archiver à la fois les sites Web eux-même et ceux

---

[7] **OCLC (Online Computer Library Center).** Web characterization. In *OCLC's Website.*
http://wcp.oclc.org
Dernière consultation le 29/10/02

et vers lesquels ils pointent un lien ? Est-ce que ce type d'archivage serait envisageable et pertinent?

3. Des problèmes liés au contexte de création

3.1 Les conséquences de l'autoédition sur Internet

Le développement de l'Internet a entraîné des changements fondamentaux dans le paysage éditorial mondial. En effet, le développement des réseaux permet un abaissement conséquent des coûts de publication. Le développement de la publication sous forme électronique induit une baisse des coûts de production, ne serait-ce qu'en ne prenant en considération que les économies faites sur les matières premières (le papier, la reliure), sur les frais d'espace d'entreposage des ouvrages. Par ailleurs, la création d'un site Web ne demande pas de compétences en informatique très poussées. En effet, n'importe qui, grâce à un éditeur, peut créer son propre site sans connaître le langage HTML. La simplicité de la diffusion sur Internet à un coût relativement faible fait du Web un formidable outil de développement de l'autoédition. Par contre, le développement de l'autoédition signifie également -et c'est un problème très largement soulevé- que l'on peut tout y trouver, le pire comme le meilleur, sans qu'aucune instance, à l'inverse de l'édition classique, n'ait pu évaluer ni filtrer les sites avant leur mise en ligne. Nous sortons d'un monde où les personnes ayant responsabilités d'édition étaient connues et identifiables, pour aboutir à un nouveau monde dans lequel chacun est potentiellement éditeur. Dans le cadre du dépôt légal, par exemple, il est possible de contraindre chaque éditeur ou imprimeur à déposer son travail, aujourd'hui, pour Internet, cette relation est impensable.

Les facilités de diffusion sur Internet ont des répercussions sur le nombre de sites Web produits et donc, sur le « poids » de l'Internet. Selon l'étude d'OCLC, pour l'année 2001 on ne comptait pas moins de 8 745 000 sites Web. A titre de comparaison, à la fin de l'année 1998, la société Alexa[8], a déposé à la Bibliothèque du Congrès un instantané (*snapshot*) du web pris au début de l'année 1997. Ce « *snapshot* » correspondait à 2 téra-octets d'informations.

Cependant, une certaine méfiance s'impose lorsque l'on parle de la taille de l'Internet. En effet, certains chiffres ne comptabilisent en fait que le nombre d'adresses URL. Or il ne faut pas oublier qu'une même institution peut avoir plusieurs sites et plusieurs URL : C'est le cas par exemple de certaines grandes entreprises internationales qui ont souvent une traduction de leur site Web par pays. Par ailleurs, on ne saurait oublier qu'une partie du

Web n'est pas accessible directement et n'est donc pas comptabilisée dans ces statistiques : C'est ce que l'on appelle le « Web invisible », « *Deep Web* » ou encore « Web profond »[9].

Notons cependant que parmi les rumeurs circulant sur l'Internet, il en est une qui tend à affirmer l'augmentation exponentielle du Web d'année en année. Or, il semble bien que cette augmentation, loin d'être exponentielle, semble se réduire. Ainsi, selon OCLC, l'augmentation du nombre de sites Web est passée de +82% entre 1997 et 1998 à +71% de 1998 à 1999, pour descendre à +52% entre 1999 et 2000 et +18% en 2001.

La taille du Web est donc un obstacle de taille pour l'archivage exhaustif de sites Web. Par ailleurs, il ne faut pas oublier que l'évaluation du nombre de sites est démultipliée par le nombre de versions successives de chaque site.

3.2 L'authentification des sites web

L'authentification des sites Web est, là encore, un problème à considérer lorsqu'il s'agit de mettre en place une politique de conservation. En effet, chaque internaute a été confronté, à un moment ou à un autre, à des difficultés pour évaluer le site qu'il visualise. L'auteur ou le créateur du site n'est pas toujours clairement identifié, ainsi que la date de la dernière modification. Si le créateur est indiqué rien ou peu de choses nous permettent d'être certain qu'il s'agit bien de l'auteur en question[10]. Enfin qu'est-ce qu'un auteur de site Web ? S'agit-il de l'entreprise ou de la personne ayant écrit les pages HTML du site ? S'agit-il du propriétaire du site (dans le cas d'un ministère par exemple ou d'une grande institution) ?

Les pages HTML ne sont, par ailleurs, pas toujours proprement constituées et la consultation du code source ne permet pas toujours d'obtenir des informations pertinentes. C'est ainsi que le titre de la page Web n'est pas toujours significatif et n'est parfois formé que par les premières lignes du texte de la page. La date de modification et la date de création des sites Web ne sont pas toujours bien indiquées. De ce fait, dans le cas de l'enregistrement d'un site Web, il y a toujours une série de confusions possibles entre la date de création du site, la date de sa dernière modification et la date de consultation du site.

---

[8] Voir en annexe 1.1

[9] Le Web invisible est l'ensemble des pages Web qui ne sont pas indexées par des robots (*crawlers*). D'une façon générale plusieurs obstacles s'opposent aux robots : les sites protégés par un mot de passe, les pages interdites au référencement par le fichier « Robot.txt », les bases de données en ligne, les pages dynamiques…

[10] Du point de vue de la sociologie de l'Internet, il est intéressant de voir que toute une sémantique se développe en vue de différencier les sites Web dits « officiels » de ceux qui ne le sont pas.

Notons cependant que la lecture du nom de domaine du site peut apporter un indice important dans l'identification du type de site, (.org , .net…) de sa localisation (.fr, .ca…)et de son objectif (commercial .comou non).
Par ailleurs, les sites Web ne comportent pas d'identificateur unique comme l'est par exemple l'ISBN. Dans le cadre d'une politique de conservation cette absence est problématique. Comment savoir, en effet, que le site Web que l'on archive est bien une nouvelle version, hébergée sur un autre serveur d'un site déjà archivé ? L'utilisation d'un identificateur unique pour chaque site Internet est une nécessité dans le cadre d'un politique de conservation, tout d'abord pour identifier à tout moment et le plus rapidement possible que le site archivé est bien la nouvelle version d'un autre site, ensuite en vue de faciliter la gestion des différentes versions d'un même site, enfin de façon à pouvoir identifier que la version d'un site enregistrée en un temps « T » est bien la suite d'une autre version enregistrée dans un autre établissement.

4. La conservation du Web est-elle possible ?

4.1 Des modes opératoires classiques en bibliothèques

La conservation des sites Web ne modifie pas fondamentalement les modes opératoires mis en place par les bibliothèques. En effet, nous retrouvons dans l'archivage des sites Web trois types d'actions que l'on retrouve dans le traitement de tout document, quel que soit son support :

- L'acquisition
- Le stockage et la conservation
- La valorisation des collections dans laquelle nous plaçons également leur description bibliographique et la mise en place d'outils de recherche.

L'archivage de sites Web passerait donc par un enchaînement d'actions depuis longtemps pratiquées par les bibliothécaires pour tous les autres types de documents. Pourtant ,le cas des sites Web pose des problèmes particuliers à chaque étape de cette chaîne opératoire « classique ».

4.1.1 Au niveau de l'acquisition

Comme pour tout type de documents, la constitution d'un fonds de sites Web archivés doit entrer dans une politique globale d'acquisition. La langue utilisée dans le site, le type de public auquel il s'adresse, les sujets qui y sont traités, le type de site Web (commercial ou

non. Personnel ou non) sont autant de critères pertinents dans la mise en place d'une politique d'acquisition de sites Web.

Cependant, l'application de ces critères de pertinence, ancrés dans une politique d'acquisition plus globale, est problématique. En effet, si l'on décide, par exemple, de mettre en place une politique d'acquisition systématique des sites Web français, encore faut-il définir ce que l'on entend par « site français ». On ne saurait se contenter d'archiver les sites hébergés uniquement sur des serveurs localisés en France. Par ailleurs, il serait insuffisant de n'archiver que les sites du domaine « .fr ». Le seul critère de la langue en vue d'identifier les sites français serait également insuffisant, la francophonie dépassant de loin le cadre français. Doit-on également conserver des sites Web d'entreprises multinationales qui consacrent certaines de leurs pages Web à leur implantation en France ? Quel niveau de granularité doit-on alors prendre en considération ?

A travers cet exemple nous voyons bien que la définition claire d'une politique d'acquisition est parfois difficilement adaptable dans le cadre de l'Internet. Ceci est sans doute particulièrement vrai dans le cas d'une politique centrée sur un critère territorial, dans la mesure où la vocation d'Internet est d'être international, perméable aux frontières et aux notions de territorialité.

La taille du Web constitue l'une des contraintes importantes du repérage des sites Web pertinents. Le Web invisible en est une autre et ce, bien qu'il existe un certain nombre d'outils spécifiques à la recherche de sites Web invisibles[11]. Or il est d'autant plus dommage de ne pas les conserver que ceux-ci comportent parfois des bases de données et des contenus particulièrement intéressants.

Cependant, sans aller aussi loin que le Web invisible, le repérage du Web en général, effectué par les robots est également problématique. En effet, tous les grands moteurs de recherche, quels que soient les critères de pertinence utilisés, n'indexent qu'une faible partie du Web. Selon une étude réalisée par l'Université de Princeton (USA)[12] en avril 1998, seuls 34% des pages Web seraient indexées par les robots. Par ailleurs, il ne faut pas oublier que les moteurs de recherche, même les plus puissants, n'indexent jamais toutes les pages d'un site Web ; la plupart d'entre eux élimine systématiquement les sites

[11] C'est le cas par exemple de « *The invisible Web* » ou encore « *Fossick Europe* » et « *Flipper* » qui sont des métamoteurs spécialisés dans la recherche sur le Web invisible et les bases de données particulièrement

[12] **JACQUESSON, Alain et RIVIER, Alexis.** Bibliothèques et documents numériques : Concepts, composantes techniques et enjeux. Paris : Editions du Cercle de la Librairie. 1999, 377p. Coll. Bibliothèques.

qui ne sont pas remis à jour ; enfin, certains moteurs de recherches appliqueraient des critères de pertinence qui ne sont pas toujours désintéressés[13].

Le fait que l'on ne puisse jamais considérer un site Web comme un document achevé[14], pose le problème de la circonscription réelle et surtout réalisable de la collection de sites Web archivés. En effet, une fois analysée la part du Web que l'on souhaite archiver, à l'intérieur de ce périmètre, chaque version de chaque site Web devrait être conservée. Or plus la « superficie » de Web circonscrite est importante, plus grand est le nombre de sites Web sélectionnés et moins le suivi de chaque site et l'enregistrement de chaque version est, pour le moment, faisable. Ainsi, d'une certaine façon, toute politique d'archivage de sites Web oscille entre une pratique extensive et une pratique intensive de conservation qui pourrait se traduire par le choix suivant : Tout avoir au moins une fois ou bien avoir toutes les versions pour quelques cas.

Par ailleurs, il faut également considérer que la notion de « mise à jour » ou de « nouvelle version d'un site » n'est pas clairement définie. Ainsi, d'un point de vue strictement informatique, le fait de modifier une virgule dans un texte est une mise à jour. Le fait de transformer la mise en forme d'un texte ou la police de caractère doit-il alors être considéré comme une mise à jour du site et, de ce fait, faire l'objet d'un nouvel enregistrement ?

4.1.2 Au niveau de la conservation proprement dite

Les conditions environnementales de conservation des supports électroniques ne constituent pas une question fondamentalement nouvelle en bibliothèque. A ce niveau, le numérique offre un certain nombre d'avantages dans le cadre d'une politique de conservation. Ainsi, en principe, d'un point de vue informatique, on ne saurait différencier une copie de son original. Ensuite, la tendance économique en informatique indique que, parallèlement à l'accroissement des capacités de stockage des supports électroniques, les coûts de stockage ont tendance à baisser. Par ailleurs, si les supports électroniques qu'ils soient optiques, magnétiques ou magnéto-optiques sont fragiles, il est possible de tester leur qualité de façon automatique. Enfin, les supports électroniques, même en nombre

---

[13] Ainsi le célèbre moteur de recherche « *Google* » est en ce moment attaqué en justice par une société concurrente (Searchking) car il aurait dévalué dans ses critères de pertinence l'entreprise plaignante sous prétexte que celle-ci pouvait constituer une concurrence importante.

Cf. **ABONDANCE.** Actumoteurs . lettre du 21 au 25 octobre 2002.[En ligne] In lettre de diffusion du *site Web d'Abondance*

http://www.abondance.com

[14] Dans la mesure où un site peut sans cessefaire l'objet d'une mise à jour.

conséquent, ne nécessitent pas un espace de stockage très vaste. A première vue, donc, la conservation de sites enregistrés sur supports ne poserait pas de problèmes.
Mais le stockage proprement dit des sites Web n'est sans doute pas le problème le plus prégnant d'une politique de conservation. Par contre, beaucoup plus problématique est l'accessibilité ultérieure aux sites Web conservés. Nous l'avons vu le rythme d'obsolescence des techniques contrecarre la consultation à long terme des sites Web archivés : rien ne nous permet de dire que les logiciels et outils disponibles dans quelques années nous permettront de consulter les fichiers informatiques que l'on archive aujourd'hui.

4.1.3 Au niveau de l'accessibilité au fonds

Comme dit précédemment la question de l'accessibilité du fonds sur le long terme est certainement le problème principal de l'archivage des sites Web.
Les facilités d'accès aux sites Web dépendent également du référencement de ceux-ci et des métadonnées utilisées et des outils de recherche proposés à l'utilisateur du fonds archivé (OPAC, moteur de recherche…)
D'un point de vue général, l'accessibilité au fonds de sites Web pose les mêmes problèmes que l'accès aux ressources électroniques en général. Le choix du matériel de consultation, de la mise en ligne des documents, la gestion des accès (avec des tours de cédéroms par exemple) sont autant de questions qui préexistent à l'archivage du Web et qui doivent se poser dans le contexte particulier de la bibliothèque concernée en fonction de ses choix politiques, de ses moyens techniques et financiers, mais également en fonction de contraintes d'ordre juridique.

4.2 Des contraintes transversales

En effet, les problèmes posés par la conservation des sites s'exercent, nous l'avons vu, à chaque étape de la chaîne opératoire de conservation, mais également à un niveau plus global. De ce fait, les contraintes agissant sur une politique d'archivage de sites Web ne sont pas seulement d'ordre technique.

4.2.1 Les contraintes d'ordre juridique

Pour le moment, l'archivage des sites Web français s'effectue à la BnF dans l'esprit et dans l'objectif d'un dépôt légal du Web, mais en l'absence d'une loi. L'article 10 du projet de loi sur la « Société de l'information », présentée en conseil de ministres le 13 juin 2001 a pour but de compléter la loi du 20 juin 1992 sur le dépôt légal.

Cependant, ce projet de loi, s'il permet de clarifier les droits et les devoirs de la BnF en matière de dépôt légal des sites Web, ne résout pas toutes les difficultés relatives à la conservation des sites Web.

En effet, la mise en œuvre d'un plan de conservation des sites Web peut entraîner la modification de ces sites. En effet, dans le cas d'une migration[15] un établissement peut être amené à transformer les formats de certains fichiers ce qui risque de porter atteinte à l'intégrité du site Web. Or le jugement qui s'est tenu le 9 février 1998 au tribunal de commerce de Paris, opposant la société Cybion à la société Qualisteam, a attribué au contenu de pages Web la qualité d'œuvres protégeables au titre des droits d'auteur. Une œuvre se définit, d'un point de vue légal, par son caractère original. Notons que cette originalité a été reconnue dans les tribunaux à des catalogues de présentation de produits et à des dessins d'application industrielle -la protection du droit d'auteur s'attachant à la forme des créations et non à leur fond-. Les sites Web, considérés comme des œuvres, sont donc soumis à la réglementation sur le droit d'auteur. Les droits d'auteur se composent à la fois de droits moraux et de droits patrimoniaux. Les droits moraux protègent l'auteur et son œuvre de toute dénaturation. Or les procédures de conservation des sites Web peuvent entraîner une transformation de ceux-ci, et donc une perte d'intégrité contraire aux droits moraux et plus exactement au droit au respect de l'œuvre[16]. De la même façon, dans le cadre du référencement des sites Web, la bibliothèque peut faire le choix d'intégrer des métadonnées dans le fichier informatique du site Web. Cette intégration *a posteriori* d'éléments pourrait-elle être considérée comme une dégradation du site ? Cette atteinte à l'intégrité des sites Web ne risque-t elle pas d'être considérée comme une atteinte aux droits d'auteur et donc exposer la bibliothèque à une mise en accusation ?

Ensuite, une partie du Web, du Web invisible, n'est consultable que grâce à un mot de passe ou par accès payant. Or souvent, ces sites comportent des informations intéressantes à conserver : des bases de données, des articles en ligne. En l'état, ces sites ne sont pas « aspirables » automatiquement. Le projet de loi sur «la Société de l'information » prévoit donc que, dans le cadre d'un dépôt légal des sites Web, d'autres procédures que l'aspiration puissent permettre à la BnF de recueillir ces sites Web, notamment par le biais d'un dépôt du site par son créateur. Ces sites seront donc intégrés dans l'ensemble des archives du Web français, mais pour autant ils seront encore soumis à

[15] Voir partie 2. La migration est l'une des solutions envisagées aujourd'hui pour conserver les fichiers numériques.. La migration est une opération qui consiste à transformer les formats ou les supports des fichiers numériques lorsque ces formats et supports risquent d'être obsolètes ou altérés. Il s'agit donc de passer d'un format d'origine à un format cible. Ce faisant, la bibliothèque transforme le fichier (le site Web) et cette transformation peut entraîner des altérations du fichier au niveau de son contenu, de ses caractéristiques formelles ou de ses fonctionnalités.

certaines restrictions quant à leur divulgation liées au droit d'auteur. Ce faisant, la bibliothèque aura-t elle l'autorisation de diffuser ces archives de sites Web sur son site Web ? Enfin, outre pour les sites protégés ou ceux aux contenus litigieux pour lesquels une limitation de l'accès sera nécessaire, la bibliothèque devra également se soumettre à la loi n° 78-17 du 6 janvier 1978 intitulée « Informatique et libertés » et, par exemple, veiller à ne pas diffuser de données personnelles protégées sur la vie privée.

Etant donné les conditions d'archivage des documents numériques et le risque d'atteindre à l'intégrité des sites Web, les données d'ordre juridique risquent de devenir une contrainte importante pour les bibliothèques, ne serait-ce que parce que cette contrainte pourrait infléchir les conditions de valorisation des fonds archivés et empêcher, par exemple, la mise en ligne de ces archives.

4.2.2 Des conséquences pour le personnel des bibliothèques

L'arrivée d'Internet dans les bibliothèques au milieu des années 1990 a déjà suscité une série de débats au sein du personnel. Où se situe le rôle du bibliothécaire dans cette offre non sélectionnée de documents ? Faut-il accepter des usages non documentaires, possibles sur Internet ?

Le grand avantage des débats qui se sont tenus et se tiennent parfois encore sur Internet est d'avoir permis aux bibliothèques de s'interroger sur leurs missions fondamentales. Aujourd'hui, comme en témoigne les articles des revues professionnelles, le débat sur l'Internet se concentre aujourd'hui sur les problèmes de conservation et sur les missions patrimoniales des bibliothèques concernées. Ce nouveau débat, associé à une prise de conscience centrée sur le sentiment de perdre chaque jour une documentation intéressante à la disparition de chaque site Web, marque sans doute un tournant dans les débats sur la présence d'Internet dans les bibliothèques. La question n'est peut-être plus de savoir si Internet a sa place en bibliothèque, mais plutôt de savoir comment intégrer Internet dans tout le circuit du document et dans toutes les missions des bibliothèques. D'ailleurs, la mise en place d'une politique d'archivage est certainement le meilleur moyen d'appliquer aux sites Web une véritable politique documentaire. En effet, pour le moment, les bibliothèques se sont lancées dans la mise en place de signets ou d'annuaires de sites Web en vue d'offrir aux usagers une sélection et une description de sites Web liée aux compétences bibliothéconomiques du personnel. Pour le moment, en l'absence d'une politique d'archivage, les bibliothèques ne proposent pas un traitement complet des sites Web. De ce fait, le choix des bibliothécaires, pour l'instant, consiste à montrer (sélectionner) un site Web ou à ne pas montrer : la mise en valeur se fait par un biais

---

[16] **BENSOUSSAN, Alain (dir.).** Internet : Aspects juridiques. Paris : Hermès, 1998, 2° éd. revue

unique à savoir mettre ou ne pas mettre dans l'annuaire, c'est-à-dire l'équivalent d'un libre-accès. Dans cette conjoncture, le désherbage de l'annuaire correspond à l'élimination pure et simple des sites. L'opération de désherbage des sites Web n'est donc pas l'équivalent de celle d'autres documents dans la mesure où il ne comporte pas l'alternative de la conservation.

La mise en place d'une politique d'archivage des sites Web permettra donc de compléter le circuit du document pour l'Internet, ce qui sera sans doute un élément positif d'intégration de l'Internet en bibliothèque et auprès du personnel. Cependant, il faut bien voir qu'en contre-partie, la conservation des sites Web induit une série de contraintes importantes pour le personnel.

En effet, la description bibliographique des sites Web peut tout à fait être effectuée en format MARC, ce qui n'implique pas de compétences nouvelles pour le personnel, habitué à cataloguer des documents complexes. Cependant, cette description des sites peut également prendre la forme de métadonnées. Ces métadonnées peuvent être plus ou moins complexes, mais selon Christian Lupovici[17], la maîtrise des métadonnées ne nécessite pas une formation extrêmement longue. Par contre, le catalogage et l'ajout de métadonnées sont effectués dans des philosophies différentes dans la mesure où l'ajout de métadonnées consiste bien à compléter le document en question.

Les métadonnées nécessaires pour un document numérique archivé intègrent également des informations d'ordre technique (format des fichiers, poids en octets, logiciels ou périphériques nécessaires…). De ce fait, la conservation proprement dite des sites Web nécessite des connaissances informatiques sur les formats et les équipements, leurs caractéristiques et leurs évolutions. La chaîne de traitement dans le cas d'archives de sites Web est particulièrement complexe et nécessite des mesures de prévention effectuées à un rythme beaucoup plus rapproché que dans le cas d'autres documents. Le bibliothécaire en charge d'un fonds de sites Web doit donc être en mesure de savoir si un format est susceptible de devenir prochainement obsolète, dans quelle mesure cette obsolescence mettra en péril le fonds et quelle partie du fonds sera effectivement concernée. Il devra être en mesure de choisir un format cible dans lequel tous les fichiers en danger devront être convertis… Etant donné la complexité de cette chaîne de conservation, l'information

---

et augmentée, 247p.

[17] **LUPOVICI, Christian.** La chaîne de traitement des documents numériques. *Bulletin des Bibliothèques de France.* 2002, t.47, n°1, p.86-91.

[En ligne]

http://bbf.enssib.fr/bbf/html/2002_47_1/2002-1-p86-lupovici.xml.asp

Dernière consultation le 07/10/02

sur les formats et les équipements est une condition nécessaire en vue d'assurer la réussite et la pérennité des archives. On pourrait alors penser que, pour ce type de travail, un informaticien serait certainement beaucoup plus compétent qu'un bibliothécaire. En fait, il est évident que, dans le cadre d'une politique d'archivage de documents électroniques en général, la collaboration des bibliothécaires avec les informaticiens sera une garantie de réussite du projet. cependant, il serait dommage de confier l'ensemble du projet à des informaticiens. Tout d'abord parce que la gestion des archives nécessite des compétences en bibliothéconomie. Ensuite et surtout parce qu'il est nécessaire d'avoir en tête tous les aspects du projet, à la fois les aspects techniques et les aspects documentaires. Imaginons par exemple qu'un format d'image soit en danger d'obsolescence imminente. Pour le remplacer deux formats sont disponibles : le premier permet de conserver la qualité de l'image, mais ne garantit pas de maintenir la mise en forme de l'image dans l'ensemble de la page et les fonctionnalités de l'image (possibilité d'agrandir l'image, de zoomer…) ; Le second altère la qualité de l'image et surtout les couleurs de celle-ci, par contre la mise en page est conservée ainsi que toutes les fonctionnalités. Seul un bibliothécaire connaissant bien le fonds, la qualité de celui-ci, l'intérêt de certains éléments, sera en mesure de choisir entre ces deux formats.

La conservation des sites Web nécessitera donc la formation du personnel, mais, la question des compétences n'est pas le seul problème à résoudre. En effet, l'archivage de sites Web peut être un projet extrêmement coûteux en personnel en fonction des solutions choisies[18].

Le critère du personnel est important pour la mise en place de tout projet en bibliothèque, surtout si ce projet engage la bibliothèque sur le long terme. Cependant, dans le cas de la conservation des sites Web et de l'information numérique en général, ce critère est particulièrement important dans la mesure où ce type de projet nécessite une formation poussée et sans cesse renouvelée. Mais aussi parce que les choix opérés par la bibliothèque ont des répercussions sur toute la chaîne de traitement de ces documents et donc sur le nombre de personnes nécessaires.

4.2.3 La question des coûts

Là encore, il s'agit d'un critère important pour la mise en place de tout projet en bibliothèque et ce d'autant plus que nous sommes placés dans un contexte général de réduction des budgets.

Or la question des coûts est particulièrement difficile à traiter lorsque l'on établit un tel projet dans la mesure où tous les grands établissements engagés dans ce type d'archivage

[18] Voir partie 2.

en sont pour le moment à un stade plus ou moins expérimental. Une partie de la technologie nécessaire à l'archivage n'existe parfois qu'à l'état de projet, comme c'est le cas, nous allons le voir, pour la technique de l'émulation. Par ailleurs, les programmes de gestion des archives électroniques, intégrant notamment la question de la migration des documents sont encore expérimentaux. Ils ne sont ni produits industriellement, ni en vente. Ainsi, les systèmes de gestion d'archives électroniques, étant donnée la complexité de la chaîne de traitement, ne peuvent être assumés par des systèmes GED classiques. Pour le moment, le système organisationnel OAIS, dont nous parlerons plus précisément, est celui qui correspond le plus au traitement d'archives électroniques. Or, lorsque la Bibliothèque Nationale d'Australie a lancé un marché en vue d'acquérir une application informatique de gestion répondant aux normes de l'OAIS, elle n'a reçu aucune réponse.

D'autre part, il faut voir que l'archivage sur le long terme de sites Web implique la gestion de volumes informatiques difficiles à évaluer. Par exemple, selon l'étude commandée par OCLC, en 1999 les sites Web français représentaient 2% de l'Internet sur un total évalué, pour la même année, à 4 882 000 sites Web ce qui équivaudrait à 97 640 sites français. Toutefois ce chiffre nous ne dit rien sur le nombre de pages Web concernées ni sur le poids en octets. Par ailleurs, ce nombre ne tient absolument pas compte du Web invisible. Or le coût d'un projet d'archivage dépend également du volume de ces archives. En l'absence d'une offre logicielle standardisée et de calculs prévisionnels sur le volume de documents à archiver, le calcul des coûts d'un projet d'archivage est certes nécessaire, mais impossible à évaluer précisément.

Cependant, il ne faudrait pas oublier que sur Internet de nombreux sites sont accessibles gratuitement. De ce fait, l'acquisition des sites se fera à un coût relativement modeste si l'on déduit les frais d'équipement (investissement) ou de personnel. Toutefois, si une grande part des acquisitions de sites sera gratuite, le traitement de ces sites et leur conservation entraîneront des coûts importants.

La conservation et le maintien de l'accès sur le long terme des sites Web s'avère donc problématique pour les bibliothèques ne serait-ce que parce que la fonction première d'un site Web n'est justement pas d'être archivé : Les sites Web se conçoivent comme des formes de communication volatiles, éphémères. Ceci étant dit, n'est-ce pas là l'une des grandes difficultés des bibliothèques patrimoniales et des services d'archives: conserver également des documents que personne n'aurait pensé intéressant de conserver ? A ce titre, le site Web n'est qu'un type de document éphémère de plus à conserver pour les bibliothèques patrimoniales. Mais, nous l'avons vu il s'agit aussi d'un type de document qui pose des problèmes particuliers.

La question posée précédemment, « la conservation des sites Web est-elle possible ? » est donc une mauvaise question. Puisque les bibliothèques doivent conserver les sites Web, la question est donc plutôt de savoir comment peuvent-elles le faire. La première étape en vue de répondre à cette question consiste certainement à savoir comment les bibliothèques engagées dans ce type de projet font effectivement, comment elles conçoivent cet archivage et quels sont les choix qu'elles ont faits pour commencer ce projet. C'est ce que nous allons voir en deuxième partie.

Partie 2 : Les solutions envisagées à l'heure actuelle

La conservation des sites Web est, nous l'avons vu, particulièrement problématique. Cependant, la plupart des grandes bibliothèques du monde ont pris conscience de la nécessité de commencer une politique d'archivage même imparfaite. Des solutions en vue de pallier les difficultés posées par l'archivage sont donc envisagées et ce sont ces solutions que nous allons tenter d'analyser à présent. Dans cette partie, nous reprendrons donc chaque phase nécessaire du travail d'archivage du Web (l'acquisition, la conservation et le référencement) en développant les solutions envisagées aujourd'hui par les bibliothèques les plus avancées dans ce type de projet. Une présentation des projets de quelques bibliothèques est développée d'une façon plus complète en annexe 1.1.

1. L'acquisition de sites Web en vue de leur conservation

L'acquisition de sites Web comprend deux phases : une première phase qui a pour objectif de circonscrire la zone du Web à acquérir. Cette phase est commune à d'autres pratiques liées à l'Internet et effectuées en bibliothèque. En effet, la constitution d'un annuaire de sites Web ou d'un répertoire de signets nécessite également une première phase d'élaboration en vue de délimiter la partie du Web que l'on souhaite mettre en valeur dans l'annuaire (une discipline scientifique particulière par exemple pour certaines bibliothèques universitaires ou une zone géographique pour une bibliothèque nationale ou à vocation régionale). La seconde phase est proprement liée à une politique d'archivage puisqu'elle consiste à acquérir « physiquement » le site Web, c'est-à-dire à l'enregistrer.

1.1 Circonscrire l'objet

Cette circonscription de la zone et de l'objet à archiver est directement liée à l'objectif du projet d'archivage. La définition de l'objectif, comme pour tout projet, se conçoit en fonction du public visé et de ses besoins, des missions de l'établissement en charge de l'archivage, des ambitions de l'établissement en matière d'archivage.

A l'heure actuelle, les établissements qui s'intéressent à l'archivage du Web sont des bibliothèques nationales, ayant, pour la plupart, la charge du dépôt légal. Les projets d'archivage de sites Web portent donc l'empreinte de cette mission fondamentale des

bibliothèques nationales, avec des points de vue toutefois extrêmement divers d'un établissement à l'autre.

1.1.1 Une délimitation spatiale

Le fait que les projets d'archivage du Web soient essentiellement mis en place à l'échelle des bibliothèques nationales induit une délimitation spatiale du Web concerné par l'archivage. C'est ainsi que la plupart des grands projets mis en place actuellement ont pour objectif d'archiver le Web d'un pays : le Web suédois, finlandais, australien, français… Nous verrons plus tard qu'à l'intérieur de ces grandes délimitations géographiques, le niveau d'exhaustivité recherché varie. Mais pour le moment, il faut bien voir que cette notion de Web national est des plus problématiques.

En effet, l'Internet s'appuie sur une architecture en réseaux qui échappe à toute notion stricte de territorialité. Qu'est-ce que un Web national ? L'ensemble des sites hébergés par des serveurs se trouvant sur un territoire donné ? L'ensemble des sites dont le domaine de l'URL comprend une indication nationale (.fr ; .se…) ? Les sites Web qui, quelle que soit leur localisation « physique » traitent du territoire national ? Les sites Web dont le créateur est un ressortissant du pays ? Dans ce cas comment le savoir ?

Si l'on souhaite établir un rapide parallèle, la « nationalité » d'un ouvrage est également problématique à définir de façon directe : une traduction publiée au Japon de « Macbeth » est-elle un ouvrage japonais ou anglais ? La question peut nous paraître absurde mais dans le cas des sites Web, cette absurdité est bien plus sensible. En effet, si l'on décide de n'archiver que les sites Web dont le nom de domaine est « .fr ». Certes, tous les sites archivés seront des sites Web français, mais pour autant cette collection de sites Web sera-t elle représentative du Web français au moment de l'archivage ? La réponse ne peut être que non. De la même manière, si l'on n'archive que les sites hébergés sur des serveurs français.

Les critères de délimitation nationale du Web diffèrent d'un établissement à l'autre. L' Australie souhaite conserver les sites qui concernent le pays et sont créés par un Australien. La Suède, quant à elle, prend en considération les sites Web dont le nom de domaine est .se ou .nu ainsi que tous les sites Web dont les noms de domaines sont plus génériques mais enregistrés à une adresse ou un numéro de téléphone suédois. Le Canada et la Suède conservent des sites considérés comme étrangers mais traitant de la Suède ou du Canada, alors que l'Australie ne conserve que ceux dont l'une des mentions de responsabilités principales est australienne.

Les critères de délimitation territoriale de l'Internet ne se définissent de façon évidente. Chaque établissement, en fonction de ses objectifs, mais également d'une approche plus philosophique, doit offrir une définition pragmatique et adaptée à son contexte des critères

de territorialité à appliquer au Web. Mais nous allons voir que la définition d'une délimitation ne concerne pas seulement la notion de territoire, mais plus largement la définition même de « site Web ».

1.1.2 Circonscrire l'objet en soi

Le site Web, nous l'avons vu en première partie, est un objet qui échappe à toute définition standardisée. Les caractéristiques de l'équipement (ordinateur, périphériques….) le type de navigateur sont autant d'éléments qui peuvent modifier la forme d'un site consulté. Alors, qu'est-ce que l'original d'un site Web ? Sa forme enregistrée sur le serveur hébergeur ? La forme telle qu'elle existe dans l'ordinateur de celui qui l'a conçu ?

Si la définition de l'objet site Web n'est pas évidente, elle s'avère particulièrement nécessaire dans le cadre d'une politique de conservation. Or les niveaux de complexité inhérents à l'objet rendent cette définition mouvante et toujours partielle. En substance, il s'agit de savoir exactement quel est l'objet que l'on souhaite archiver et définir les propriétés signifiantes principales de l'objet.

Le site Web se caractérise par un contenu mais également par une forme et des fonctionnalités. Or, nous l'avons vu cet ensemble correspond à des niveaux de complexité croissants d'un point de vue informatique. Tous les projets en cours ont pour objectif de conserver tous les aspects des sites Web (forme, contenu, fonctionnalités).

Un site Web, par ailleurs, n'est pas un objet fini dans la mesure où il évolue dans le temps au fil de ses mises à jour, mais également dans la mesure où ses limites ne s'arrêtent pas à son propre contenu puisqu'il établie des liens hypertextes vers d'autres sites Web. Dans le cadre d'une politique d'archivage, l'idéal consisterait à conserver toutes les versions d'un même site ainsi que tous les liens hypertextes externes inclus dans le site. Matériellement cet idéal n'est pour le moment pas envisageable. En ce qui concerne les liens hypertextes externes seuls deux projets les conservent entièrement pour chaque site archivé : il s'agit de la fondation américaine « Internet Archive » et du projet très spécialisé de l' « American Astronomical Society ». La Bibliothèque nationale d'Australie n'archive les liens externes que s'ils correspondent à des critères d'acquisition précis. En suède ces liens ne sont conservés que s'ils renvoient à des sites suédois (c'est-à-dire telle qu'a été définie la notion de territorialité des sites Web dans le contexte suédois.).

En ce qui concerne les versions successives d'un même site, plus le nombre de sites archivés est important, moins le suivi de chaque nouvelle version d'un même site est possible. Si chaque établissement concerné par l'archivage tend vers cet idéal (conserver toutes les versions de chaque site) de manière plus ou moins avouée, la mise en place concrète des projets d'archivage est beaucoup plus pragmatique. Ainsi, la bibliothèque royale de Suède acquiert les sites qu'elle doit archiver en prenant un instantané du Web

suédois à un instant donné (snapshot). Elle effectue en moyenne deux snapshots par an permettant ainsi de recueillir pour certains sites, deux de leurs versions.
De la même façon, doit-on conserver des newsgroups ? S'agit-il de sites Web ? Pour le moment, il semble que la BnF ne s'achemine pas vers un archivage de newsgroups ou de forums de discussion. Par contre, en Norvège, le téléchargement de newsgroups norvégiens a commencé.

1.1.3 Les projets de dépôt légal de sites Web

L'intégration d'un nouveau type de document dans le dépôt légal nécessite une définition claire du document concerné, mais également une définition permettant de circonscrire clairement le périmètre d'action de l'établissement en charge du dépôt légal.
Dans le cas des sites Web cette délimitation et la définition de l'objet sont, nous l'avons vu, problématiques. Chaque établissement est donc amené à adopter une définition pragmatique, adaptée à son contexte et à ses objectifs.
De ce fait, le périmètre d'action, la définition de l'objet, de sa nationalité, varient très profondément d'une bibliothèque nationale à l'autre et cette variabilité dépend en grande partie du contexte légal dans lequel le dépôt légal a été mis en place. En France, par exemple, la loi sur le dépôt légal du 20 juin 1992 définit le champ d'action de la BnF au niveau des supports des documents. C'est ainsi que la loi a pu intégrer sans grandes remises en question, des supports électroniques matériels comme les cédéroms par exemple. Dans le contexte français, l'élargissement du dépôt légal aux ressources en ligne entraînera donc le vote d'une nouvelle loi puisqu'il ne s'agira plus de déposer des supports mais de l'information immatérielle. En Norvège au contraire, la loi sur le dépôt légal de 1989 ne distingue pas clairement les ressources matérielles des ressources immatérielles.
Selon Catherine Lupovici[19] les approches, différentes d'un pays à l'autre, dépendent également du contexte national de l'édition dans chaque pays concerné.
Pour le moment, la plupart des pays engagés dans un projet de dépôt légal des sites Web et des ressources électroniques en ligne, sont dans l'attente d'un changement de législation dans le domaine du dépôt légal. En l'absence d'une loi, les établissements débutent malgré tout des projets d'archivage, mais la mise en œuvre d'une politique importante est pénalisée par l'absence de cadre légal. En effet, l'accès à certains sites est

---

[19] **LUPOVICI, Catherine.** Les stratégies de gestion et de conservation des documents électroniques. *Bulletin des Bibliothèques de France.* 2000, T.45, n°4, p.43-54.
[En ligne]
http://bbf.enssib.fr/bbf/html/2000_45_4/2000-4-p43-lupovici.xml.asp
Dernière consultation le 12/09/02

payant et rien n'oblige, pour le moment ,les éditeurs à déposer leurs sites ou bases de données. C'est le cas également aux Etats-Unis où le projet de dépôt légal des ressources électroniques (CORDS : Copyright office's electronic registration and deposit system) n'est pas encore appliqué. Dans ce contexte, la société *« Law and technology press »* suggéra que si la bibliothèque du Congrès souhaitait obtenir un périodique, elle n'avait qu'à souscrire un abonnement…

En conclusion, la définition du champ d'action d'un établissement en matière de dépôt légal de sites Web ne peut se mettre en place aisément. Dans ce contexte, les difficultés que rencontrent les établissements pour circonscrire leur champ d'action ont plusieurs origines :

- Les difficultés pour appliquer aux sites Web une notion de territorialité
- Les difficultés pour circonscrire et définir l'objet à archiver (qu'est-ce qu'un site Web ? Est-ce que les liens hypertextes externes vers lesquels pointe un site font partie du site ?…)
- Le contexte légal et économique du pays en matière de dépôt légal.

1.2 L'acquisition proprement dite

Plusieurs solutions sont envisageables pour l'acquisition effective des sites Web en vue de leur conservation.

1.2.1 Première solution : la sélection manuelle des sites à archiver

Cette solution a été choisie par la Bibliothèque nationale d'Australie et par la Bibliothèque nationale du Canada. Elle consiste à repérer, sélectionner et acquérir manuellement les sites Web à archiver.

L'exhaustivité en matière d'archivage de sites Web est un objectif impossible à réaliser, même à l'échelle d'un pays. Face à ce constat, les bibliothèques canadienne et australienne ont adopté un point de vue des plus pragmatiques. Elles souhaitent toutes deux conserver des sites qui comportent un intérêt national, mais refusent de viser l'exhaustivité même relative en matière d'archivage. Il s'agit donc pour chacune d'entre elles de sélectionner au plus près les sites Web à archiver en analysant leur contenu. A ce titre, elles mettent en place une politique qui s'apparenterait davantage à une politique d'acquisition qu'à la mise en place d'un dépôt légal.

Comme pour toute politique d'acquisition, un certain nombre de critères d'évaluation et de sélection sont mis à la disposition des acquéreurs. A titre d'exemple, parmi les critères de sélection du projet PANDORA de la bibliothèque nationale d'Australie, on peut noter :

- que le site Web doit avoir été créé par un australien

- que le contenu du site Web doit porter sur l'Australie et sur un sujet d'ordre social, politique, culturel, scientifique, religieux ou économique
- Le contenu du site Web doit être de qualité et apporter une véritable contribution scientifique
- Le contenu informationnel du site Web ne doit pas se trouver par ailleurs sur des supports plus classiques (livre, périodiques…)

Cette solution offre plusieurs avantages. Tout d'abord, elle permet de suivre très précisément le nombre de sites archivés et leur poids en octets ce qui simplifie l'archivage et permet une meilleure prévision des coûts d'archivage. Ensuite, elle offre l'avantage de permettre un suivi plus fin et, notamment, d'envisager de conserver un nombre plus important de mises à jour pour un même site. Enfin, le travail humain de sélection évite les imperfections d'un travail automatique effectué par un robot.

Cependant, cette même solution est aussi hautement problématique. En effet, rien ne nous permet de savoir ce que les usagers de demain rechercheront dans le Web australien d'aujourd'hui. Les critères de sélection que les bibliothèques australienne et canadienne appliquent sont peut-être en décalage avec les attentes futures des usagers. La conservation sur le long terme interdit certainement une sélection trop draconienne, mise en place dès le départ. Par ailleurs, il est vraisemblable que le fonds ne sera pas véritablement représentatif du Web australien.

Enfin, bien qu'aucun chiffre précis ne soit disponible, le nombre vraisemblable de personnes nécessaires pour la sélection et l'acquisition de sites Web est relativement important et surtout les charges de travail liées à l'acquisition s'inscrivent dans la durée puisqu'il s'agit de tâches sans cesse renouvelées.

1.2.2 Le dépôt

Cette solution serait une adaptation au contexte du Web des procédures de dépôt légal existant pour des supports plus classiques (le papier, la vidéo…). Il s'agirait donc de contraindre légalement le créateur du site Web à déposer physiquement celui-ci à la bibliothèque concernée.

Cette solution offre des avantages certains puisqu'elle s'appuie sur des principes et des procédures connues, déjà bien maîtrisées. Cependant, la liste de ses inconvénients est aussi des plus conséquente. Tout d'abord parce que, nous l'avons vu, l'Internet est un espace d'autoédition, démultipliant ainsi le nombre d'interlocuteurs de la bibliothèque. Comment, en effet, envisager que chaque créateur de site Web puisse déposer son site ? Comment envisager un travail de veille et de réclamation de la bibliothèque à l'échelle de dizaines de milliers d'éditeurs ?

Le dépôt de ressources électroniques, même appuyé par un dispositif légal, ne semble, en outre, pas véritablement entré dans les mœurs. A titre d'exemple, depuis 1994, le dépôt de documents électroniques sur support est entré en vigueur en France. Or, quatre année après cette loi, la bibliographie nationale de la BnF ne faisait état que de 1758 documents électroniques, alors que chaque année, quelque 60 000 titres d'imprimés entrent dans la collection[20]. Dans ce contexte, il est bien difficile que le dépôt de sites Web puisse obtenir un succès plus important que le dépôt de ressources électroniques sur support.
Cette solution est, d'ailleurs, à ce point problématique qu'elle n'a été retenue par aucun établissement. La BnF, par contre, intègre le dépôt non pas en tant que solution seule et définitive, mais en tant que recours dans le cas très précis des bases de données, non accessibles par le biais d'un robot.

1.2.3 La collecte automatique

Cette solution passe par l'utilisation d'outils logiciels dits « robots collecteurs » ou « *harvesters* ». Il s'agit en fait de moteurs de recherche qui parcourent le Web, et rapatrient les sites repérés en fonction de critères de pertinence paramétrés dans le robot. Cette solution a été adoptée, entre autres, par la Bibliothèque royale de Suède et la BnF. La fondation privée Alexa procède également à une collecte automatique du Web à l'échelle mondiale. Le produit obtenu par le robot est un « snapshot » c'est-à-dire un instantané du Web ou d'une portion du Web.
Cette solution offre de nombreux avantages. Tout d'abord, elle permet une collecte massive et beaucoup moins discriminante que dans le cas d'une sélection manuelle. Pour le moment, elle constitue le meilleur moyen d'obtenir, sinon la conservation exhaustive du Web, du moins une part représentative de celui-ci. Elle permet aussi, pour l'utilisateur, de conserver les fonctionnalités de navigation du Web et notamment d'une partie des liens hypertextuels externes. En ce sens, les archives ainsi obtenues reproduisent la forme globale du Web. Enfin, cette solution nécessite un personnel nombreux et spécialisé au départ du projet –au moment de la conception de l'outil et des différentes phases de test- mais, par la suite, hormis pour des tâches de maintenance, la charge de travail diminue.
Cependant, si la collecte automatique offre de nombreux avantages et apparaît la plus adaptée au cas du Web, elle est également contraignante et problématique. On a tendance à opposer la grande sélectivité de la collecte manuelle à l'absence de sélection et donc à

---

[20] In **JACQUESSON, Alain et RIVIER, Alexis.** Bibliothèques et documents numériques : Concepts, composantes techniques et enjeux. Paris : Editions du Cercle de la Librairie. 1999, 377p. Coll. Bibliothèques.

une certaine neutralité que serait permise par la collecte automatique. Il serait faux d'opposer ces deux méthodes en considérant que la collecte manuelle est sélective alors que la collecte automatique ne le serait pas. En fait, la collecte automatique intègre une sélectivité mais qui est systématique, intégrée dans le robot. Comme dans le cas de la collecte manuelle, le premier problème que pose donc la collecte automatique concerne les critères de sélection. Un « harvester » fonctionne, au départ, comme un moteur de recherche classique (comme « Google » ou « AltaVista ») : il comporte un « spider » qui parcourt le Web. Comme les moteurs de recherches, les « harvesters » sont paramétrés selon des algorithmes de pertinence qui leur permettent de rapatrier et enregistrer des sites Web pertinents. Ces critères de pertinence, à l'instar des critères de sélection employés dans les collectes manuelles, peuvent poser problème. La BnF, par exemple, possède un robot « harvester » dit « Robot du dépôt légal ». La BnF a décidé d'intégrer parmi les critères de pertinence un indice de notoriété. Ce critère a été popularisé par « Google » et il consiste à évaluer une partie de la pertinence des sites Web en fonction du nombre de liens hypertextes pointant vers le site en question. Concrètement, plus un grand nombre de sites Web établiront un lien hypertexte externe vers un site Web particulier, plus ce site Web se verra attribuer un indice de notoriété important et plus il sera considéré comme pertinent. L'indice de notoriété, qui explique en partie le succès de « Google », ne risque-t il pas de s'avérer problématique dans le cadre d'une politique de dépôt légal du Web français ? Ne risque-t on pas, par exemple, de retrouver toujours les mêmes sites, les plus importants et ignorer des sites moins connus, plus rares ?

Aux problèmes des critères de pertinence s'ajoute celui de l'accès au Web invisible. Les moteurs de recherche n'ont pas accès au Web invisible. Or, les bases de données, certains articles font partie du Web invisible et représentent assurément une documentation particulièrement intéressante. Le moyen le plus simple de prendre conscience de cet écueil est certainement de naviguer dans les archives Internet de la fondation « Internet Archive ». La navigation sur le site d'Internet Archive montre que de nombreuses bases de données, les pages accessibles par mots de passe ou contre paiement n'ont pas été archivées. C'est le cas par exemple des catalogues de bibliothèques en ligne qui, en tant que bases de données, ne sont pas archivés[21] . Selon un article publié dans « Le Monde »[22] 40% du Web serait inaccessible de façon automatique.

---

[21] Notons que parfois, Internet Archive établie un lien avec le catalogue actuel de la bibliothèque. C'est-à-dire que lorsque vous naviguez, par exemple, sur le site de 1997 d'une bibliothèque , si vous accédez au catalogue, vous accédez en fait au catalogue tel qu'il existe en ligne actuellement, ce qui constitue une expérience temporelle relativement intéressante.

[22] « Le dépôt légal du Web, terrain de compétition à la française » . *« Le Monde »* . Le 06/04/02
http://www.-rocq.inria.fr/~abitebou/pub/lemonde02.html
Dernière consultation le 14/11/02

La collecte automatique n'est donc jamais exhaustive dans la mesure où, de par le paramétrage des critères de pertinence, on exclut une partie du Web. Mais aussi, dans la mesure où la machine, le robot, ne sera pas en mesure de repérer et d'indexer un certain nombre de sites Web qui, pourtant, correspondraient aux critères de pertinence. La collecte automatique pose, par ailleurs, un autre problème d'importance. En effet, pour des raisons techniques, il n'est pas possible d'effectuer un très grand nombre de snapshots par an. La bibliothèque Royale de Suède, par exemple, a été en mesure d'effectuer environ deux snapshots par an depuis 1997. Or, entre deux snapshots de nombreux sites ont subi des modifications d'importance. Certains même ont pu apparaître et disparaître sans avoir été, à aucun moment, archivés. Comparé à des acquisitions plus classiques en bibliothèques, c'est un petit peu comme si une bibliothèque n'acquérait que les ouvrages publiés entre le premier janvier et le premier mars d'une même année, les publications publiées en dehors de cette période lui échappant irrémédiablement.

Enfin, la collecte automatique demande beaucoup de moyens et infléchit fortement la mise en œuvre globale du projet d'archivage. En effet, la collecte automatique nécessite d'abord un moteur de recherche puissant assorti de grandes capacités de stockage. En amont de la machine elle-même, le réseau de la bibliothèque doit être particulièrement puissant et rapide dans la mesure où il s'agit non seulement de parcourir de le Web mais encore de rapatrier sur un serveur un poids très important d'informations : à titre d'exemple, le dernier snapshot effectué par la bibliothèque royale de Suède a permis l'acquisition de quelque 31 millions de fichiers.

Par ailleurs, s'il existe dans le commerce des logiciels permettant de créer son propre moteur de recherche par contre, dans le cadre d'une vaste politique d'archivage, il est très difficile d'envisager une offre standardisée. De ce fait, chaque harvester, du fait qu'il intègre des critères de pertinence différents, est un objet unique. Les coûts liés à la phase de développement et de test du robot sont prohibitifs. Ainsi, en vue de débuter le projet, le gouvernement suédois a financé le projet à hauteur de 3 millions de couronnes suédoises en 1996 soit 330 658 euros. Il faut préciser toutefois que, dans ce domaine, la bibliothèque royale de Suède fut relativement pionnière et qu'il est possible d'espérer que les coûts sont aujourd'hui plus bas.

Enfin, le choix d'une collecte automatique a des répercussions sur la conception et l'organisation globale du projet. En effet, étant donné le nombre de sites acquis en un temps relativement court, il apparaît impossible de cataloguer manuellement chacun d'eux.

En conclusion, on peut dire qu'il n'existe pas, en matière d'acquisition de sites Web, de solutions parfaites. Par contre, il existe des solutions adaptées à certains objectifs et à certaines ambitions. Si l'objectif visé est d'acquérir un fonds de sites Web dont on souhaite

suivre plus ou moins précisément l'évolution et dont le contenu fait l'objet d'une sélection pointue, rien ne remplacera les compétences du personnel dans le cadre d'une collecte manuelle. Si l'objectif par contre est de viser l'exhaustivité ou du moins d'acquérir un ensemble représentatif du Web (mondial ou local) à un moment donné, la collecte automatique s'avère la plus satisfaisante.

2 La conservation des documents acquis

Une fois les sites Web acquis se pose alors la question de leur conservation sur le long terme. Or, nous l'avons vu en première partie, du fait que les sites Web sont des documents numériques complexes, cette tâche est hautement problématique. La fragilité des supports mais surtout l'obsolescence des techniques et des formats compromettent l'accès ultérieur des usagers à ces sites Web archivés. Plusieurs solutions sont cependant envisageables.

2.1 La conversion analogique des sites Web

Etant donnés les problèmes techniques posés par la conservation des sites Web et des documents numériques en général, certains n'hésitent pas à envisager la conversion analogique de ces documents problématiques. Il s'agirait concrètement d'imprimer les pages Web sur papier ou sur microforme. Ce choix se fonde sur une idée simple : Alors que l'on ne maîtrise pas la conservation des supports et des documents numériques, les bibliothèques et les services d'archives ont acquis une connaissance certaine de la conservation du papier ou des microformes qui, par ailleurs, se conservent beaucoup plus longtemps que les supports électroniques.

Il est par exemple possible de produire des microformes directement à partir des données binaires d'un document. Cette technique est appelée COM (Computer Output Microfrom).

Le problème de cette conversion est qu'elle ôte aux sites Web toutes leurs fonctionnalités de navigation, de recherche… Le rendu analogique de l'objet est à ce point différent de l'objet originel que l'on peut se demander s'il s'agit bien encore d'un site Web archivé.

Dans le cadre d'une politique d'archivage des sites Web et à ma connaissance aucune bibliothèque n'a envisagé cette solution comme une perspective unique. Par contre, la technique COM est employée par la SNCF en vue d'archiver les plans des rames des nouvelles lignes TGV. Mais il faut dire qu'un plan conservé aux archives de la SNCF n'a pas pour but unique d'être conservé mais également il doit pouvoir être retravaillé ultérieurement. A ce titre, la technique COM très adaptée dans le cas des archives de la

SNCF paraît peu pertinente dans le cas de la conservation de sites Web par une bibliothèque.

2.2 Le musée technique

Comme dit précédemment, la consultation ultérieure des sites Web que l'on conserve aujourd'hui est largement dépendante de l'environnement technique du site Web.
Pour pallier ces difficultés, il s'agirait donc de conserver non seulement le site Web mais également tout l'environnement technique du site (logiciels, périphériques, ordinateur). Cette solution offrirait l'avantage de permettre la consultation des sites dans ses conditions d'origine. Toutefois cette solution n'offre pas toutes les garanties de succès. Tout d'abord parce que la conservation de l'environnement technique suppose son maintien en état, or les pièces de rechange des équipements ne seront certainement plus disponibles. Par ailleurs, les informaticiens ne seront certainement plus formés pour assurer la maintenance d'appareils obsolètes dans des langages de programmation largement dépassés. Enfin, plus prosaïquement, le manque de place pour ce type de collection se ferait certainement et rapidement sentir.

2.3 La migration

La migration consiste à effectuer une transformation plus ou moins importante du document à conserver en suivant l'évolution des techniques. D'une certaine façon les techniques liées à la migration sont relativement connues en bibliothèque. En effet, lorsqu'une bibliothèque veut changer son système informatique, passer à une version plus récente ou encore acquérir un nouveau logiciel plus performant, il s'agit alors d'effectuer, avec le moins de perte de données possible, la récupération des données existant dans l'ancien environnement et leur intégration dans un nouveau.
On peut, globalement, de distinguer deux types de migrations :

- La migration consistant à changer de support physique de stockage. Par exemple il s'agit de migrer un fichier d'une disquette vers un cédérom. Ce type de migration ne présente pas de difficultés majeures dans la mesure où le document lui-même n'est pas véritablement transformé, le train de bits du document n'étant pas altéré. Seule la norme présente dans la couche physique du document est transformée.
- La migration qui consiste à modifier le format ou le codage des données du document. (Par exemple convertir un document écrit en Word 98 en un document composé en Word XP. Ou encore, par exemple, de transformer un document en ASCII en document codé en UNICODE). Cette forme de migration est plus problématique dans

la mesure où il touche aux couches plus abstraites du document et donc au document lui-même.

Il s'agirait donc, dès qu'un format ou un logiciel est amené à disparaître, de convertir le fichier en danger d'obsolescence dans un nouveau format. Ces migrations se feraient donc périodiquement. La migration, à l'heure actuelle, est la seule solution permettant d'envisager la conservation à long terme de sites Web et, d'ailleurs, c'est la solution qui a été retenue par les bibliothèques lancées dans des projets d'archivage du Web. Pour autant, cette solution est aussi problématique à plus d'un titre.

Tout d'abord à partir du moment où l'objet est transformé, il perd de son intégrité, surtout lorsque l'on considère qu'un site Web conservé sur le long terme devra subir plusieurs migrations et donc plusieurs transformations. La perte peut toucher les fonctionnalités de l'objet, mais également sa forme voire son contenu. Ainsi, le Centre National d'Etude Spatial a été contraint récemment de procéder à des migrations car le mode de codage des nombres réels qu'il utilisait était propriétaire et qu'il souhaitait utiliser un mode de codage standardisé. Cette opération peut paraître simple mais il n'a pas été possible, lors de cette migration de garantir l'intégrité de la dernière décimale. Cette altération du document implique donc que le document conservé *in fine* ne peut être véritablement considéré comme un original. Mais, plus grave, à partir du moment où, dans le cadre d'une politique d'archivage de sites Web, une bibliothèque fait le choix de la migration, elle s'autorise à transformer le document qu'elle a pour mission de sauver. Cette situation est problématique au regard des missions d'une bibliothèque patrimoniale, mais également elle a des répercussions juridiques problématiques. Nous l'avons vu dans la première partie de ce présent travail, en touchant consciemment à l'intégrité de l'objet qu'elle conserve, la bibliothèque porte atteinte aux droits de l'auteur de l'objet. D'une certaine façon, l'archivage de sites Web oscille entre deux risques majeurs : d'une part le risque de perdre définitivement la possibilité d'accéder à un fonds si l'on ne lutte pas contre l'obsolescence technique et d'autre part, le risque de transformer irrémédiablement les sites Web archivés.

Par ailleurs, la migration nécessite, pour la bibliothèque, d'être informée d'une manière pointue sur les évolutions des techniques informatiques (logiciels, formats…). Or certaines informations sur des logiciels et des formats propriétaires ne sont pas communiquées. L'information devient, dans le cadre de la migration, un enjeu vital pour la pérennité des archives. Or le personnel des bibliothèques n'est pas formé pour assurer ce travail de veille informatique.

2.4 L'émulation

Contrairement aux trois autres solutions explicitées plus haut, l'émulation n'est pas une solution d'ordre bibliothéconomique, mais bel et bien une solution purement technique et informatique.

Un émulateur est un ensemble de dispositifs et de logiciels permettant d'exécuter sur un certain type d'ordinateur les instructions écrites pour un autre type d'ordinateur. L'ordinateur conserve donc ses caractéristiques mais simule en quelque sorte le comportement d'un autre. Un émulateur peut concerner des ordinateurs de fabriquants différents mais de même génération, mais également des ordinateurs de générations différentes. A titre d'exemple, il est possible d'émuler un C64 (Commodore 64) sorti en 1982 sur un microprocesseur Pentium courant. Le langage JAVA, par exemple, fonctionne en quelque sorte comme dans le cadre d'une émulation. En effet, il peut être lu par n'importe quelle plate-forme informatique mais il ne communique pas véritablement avec la machine mais avec une machine virtuelle dite JVM (Java Virtual Machine).

Il serait donc possible d'accéder aux sites Web dans leur forme originale par le biais d'un émulateur capable d'exécuter des instructions écrites dans des langages obsolètes, pour du matériel obsolète. Cette solution offre l'avantage de conserver l'intégrité de l'objet archivé et de permettre, sur le long terme, d'accéder à toutes les composantes du site Web (son contenu, son apparence et ses fonctionnalités). Pour le moment l'émulation apparaît donc comme une piste de recherche importante, mais non comme une solution envisageable à court terme. Par ailleurs, il faut voir que les émulateurs devront être en mesure de mimer le comportement de plusieurs générations d'ordinateurs et de systèmes d'exploitation. En effet, si l'on considère que les nouvelles versions des grands systèmes d'exploitation apparaissent en moyenne tous les deux ans et en admettant que la compatibilité ascendante ne soit assurée que pour la version immédiatement précédente, cela signifie donc qu'au bout de vingt années, les émulateurs devront être capables d'émuler dix systèmes différents.

D'une façon générale, les chercheurs s'accordent sur le fait que les coûts de l'émulation seraient nettement inférieurs à ceux des migrations successives. Cependant, il faut bien voir que l'émulation est encore un objet de recherche et que le développement d'émulateurs adaptés aux missions d'archivage sera long. La mise en place de produits standardisés ne sera pas non plus immédiate. Or il est probable que les bibliothèques et les services d'archives constitueront le seul marché des entreprises susceptibles de développer ces émulateurs. Dans la perspective où une entreprise déciderait de se lancer dans la conception de tels outils, le prix de vente de ceux-ci risque d'être prohibitif. A titre de comparaison, les bibliothèques sont également le seul marché des éditeurs de revues

en ligne. Or nous savons bien qu'en vue de compenser cette absence de diversification de clientèle, les éditeurs ont largement augmenté les coûts d'abonnement pour les bibliothèques clientes.

En conclusion, au niveau des solutions envisagées pour la conservation des sites Web, l'émulation apparaît comme la plus intéressante dans la mesure où elle permet de conserver l'objet dans toute son intégrité tout en permettant un accès sur le long terme. Cependant, cette option n'est pour le moment pas une solution à proprement parler mais plutôt une perspective. Pour le moment, la seule solution viable demeure la migration en dépit de tous les problèmes que cette technique pose.

3. L'identification et le référencement des sites Web

La mise en valeur d'un fonds en bibliothèque passe obligatoirement par sa description dans un catalogue ou par une autre possibilité de recherche. Un document non référencé, non traité, n'existe pas pour le public. Il en est ainsi pour n'importe quel type de documents. Dans le cas d'une collection de sites Web, le référencement constitue aussi une condition fondamentale de conservation. En effet, dans le cadre d'une conservation par migration, le fait de connaître les formats des documents présents dans le site, le système d'exploitation, le langage dans lequel le site a été écrit sont de informations fondamentales. Elles permettent de repérer les fichiers qui peuvent être mis en danger par une obsolescence logicielle ou matérielle et donc de procéder à une nouvelle migration.

En outre, il est important de savoir que le site A est en fait une nouvelle mise à jour du site B précédemment archivé. Ce repérage des versions successives d'un même site nécessite une identification claire de ceux-ci.

Nous allons donc aborder ces deux aspects de la description des sites Web archivés : l'identification et le référencement.

3.1 L'identification unique

Plusieurs solutions sont envisageables pour attribuer aux sites Web un identifiant unique :

- L'URL (Uniform Resource Location) est un identifiant unique pour chaque site Web mais par contre il ne s'agit pas d'un identifiant persistant dans la mesure où un même site peut changer d'URL c'est-à-dire de localisation. Le choix donc de l'URL comme identifiant unique implique de pouvoir établir un lien entre toutes les versions d'un même site même si au cours de plusieurs remises à jour l'adresse URL a été modifiée.
- L'URN (Uniform Resource Name). Comme l'URL ce nom a été conçu par l'IETF (Internet Engeneering Task Force) qui est une organisation de standardisation. Il s'agit

d'un nom d'identification unique pour chaque site Web et persistant. L'enregistrement de cet URN s'effectue auprès de l'IETF

- Le PURL (Persistent Uniform Resource Locator) a été conçu par l'OCLC. Il s'agit d'un identifiant de localisation comme dans le cas de l'URL, mais qui, contrairement à l'URL, serait persistant. Il se présente sous la forme d'un alias public. Il est créé par un administrateur de site Web et est enregistré comme « propriétaire » de PURL. Il maintient une mise en correspondance du PURL avec une URL.
- Le DOI (Digital Object Identifier) a été élaboré par « l'Association of American Publishers » et la « Corporation for National Research Initiatives ». Il s'agit d'un numéro d'identification unique et persistant ressemblant à l'ISBN. Il doit permettre d'identifier les objets numériques quels qu'ils soient. Ce système a été créé de façon à améliorer la défense du copyright pour les ressources en ligne et à faciliter le commerce en ligne de ressources électroniques.

Les quelques exemples présentés ici se situent à un niveau international. Il faut noter que la Bibliothèque nationale d'Australie, dans le cadre de PANDORA, son projet d'archivage du Web australien, a développé son propre système d'attribution d'un identifiant unique.

3.2 Référencement et métadonnées

Le référencement, la description bibliographique d'un site Web ou de tout autre type de document consiste à produire des métadonnées sur le document. Une métadonnée est une information sur le document. A ce titre, le catalogage en format Marc consiste en un ensemble de métadonnées.

Comme nous l'avons vu, les métadonnées ne doivent pas seulement permettre d'identifier, décrire et localiser un document, mais participent également de leur conservation.

3.2.1 Les types de métadonnées nécessaires

Quel que soit le moyen finalement choisi en vue de référencer les sites Web archivés, il faut retrouver, d'un système à l'autre trois types de métadonnées :

- Les métadonnées descriptives, classiques en bibliothéconomie qui permettent d'identifier entre autre l'auteur, le titre, la date de création du document
- Les métadonnées structurelles ou de conservation Il s'agit de rassembler un ensemble d'informations sur la structure informatique des sites Web : l'arborescence des fichiers, les formats des fichiers, le langage, les code de caractères….
- Les métadonnées de gestion comprennent les informations sur l'histoire du document dans l'institution de conservation. Il s'agit par exemple de savoir à quel moment la ressource a été enregistrée, le nombre de modifications (migrations) qu'elle a subies, les formats successifs dans lesquels les fichiers ont été enregistrés… Ces données

renferment également des données sur la gestion de droits de consultation de la ressource : qui a le droit de la consulter ? Sous quelles conditions ? Quels sont les droits de reproduction des usagers?…

Catherine Lupovici et Julien Masanès[23], dans le cadre du projet européen NEDLIB se sont penchés sur la définition des métadonnées indispensables et minimales en vue de l'archivage de sites Web. Le détail de ces métadonnées est présenté, traduites en français, en annexe. Ces métadonnées sont adaptables et permettent de décrire n'importe quel site. Toutefois, il ne faut pas oublier qu'il s'agit de métadonnées minimales. De ce fait, il manque plusieurs informations comme le nombre de pages du site, les types de documents inclus, l'arborescence du site Web, sa structure… Du moins permettent-elles de se faire une idée du niveau de description minimal qui devra être effectué sur ces sites Web archivés et, surtout, elles montrent bien que le référencement nécessitera certaines compétences et connaissances en informatique.

3.2.2 Les systèmes existants

Plusieurs systèmes existent pour cataloguer des sites Web, parmi eux :

Le catalogage en UNIMARC

En 1997 l'IFLA a développé la norme ISBD(ER) (International Standard for the Bibliographical Description of Electronic Resources) édictant les règles de description bibliographique des ressources électroniques.L'ISBD(ER) comporte 8 zones[24]. A partir de cette norme de description, le catalogage en UNIMARC des ressources électroniques a été rendu possible par le travail de l'UBCIM (Universal Bibliographic Control and International MARC Core Program), groupe de l'IFLA. Un manuel est accessible en ligne depuis août 2000[25].

Ce système offre l'avantage de s'appuyer sur des outils et des pratiques professionnelles existants. Ainsi que dans le cadre du projet australien PANDORA, les sites Web archivés sont catalogués au format MARC. Par ailleurs l'avantage de ce système est de présenter au

---

[23] **LUPOVICI Catherine et MASANES Julien**. Metadata for long term preservation . [En ligne] .In *NEDLIB*
http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm
Dernière consultation le 18/11/02

[24] **BnF.** *«Information pour les professionnels : ISBD(ER) »*
http://www.bnf.fr/pages/zNavigat/frame/infopro.htm
Dernière consultation le 18/11/02

[25] **IFLA.** UNIMARC Guidelines n°6 : Electronic resources. In *IFLA.* [En ligne].
http://www.ifla.org/VI/3/p1996-1/guid6.htm
Dernière consultation le 19/10/02

public un catalogue unifié présentant tous les documents de la bibliothèque quel qu'en soit le type.

Cependant, le catalogage en format MARC pose également de nombreux problèmes. Tout d'abord, l'avantage de l'ISBD(ER) est de pouvoir décrire toutes sortes de ressources électroniques. Cet avantage, dans le cadre d'un archivage aussi spécifique et contraignant que celui des sites Web, devient une limite majeure. En effet, plusieurs éléments sont difficilement adaptables dans le cadre du Web. C'est ainsi que la zone d'édition doit mentionner l'ensemble des mises à jour qu'a connu la ressource électronique décrite. Or nous savons combien il est difficile de pouvoir connaître avec précision ce type d'information pour les sites Web. Par ailleurs, alors qu'il s'agit d'une donnée importante, la taille de la ressource n'est que facultative. De la même façon, les données sur l'environnement informatique nécessaire à la consultation du site Web archivé sont présentées dans le champ des notes (337).

Outre ces difficultés, le catalogage en MARC des sites Web archivés est une opération lourde en personnels. De ce fait, elle ne peut vraisemblablement se concevoir que dans le cadre de projet où le nombre de sites archivés est relativement peu important (dans le cas d'une approche sélective et manuelle de l'acquisition de sites Web), comme c'est effectivement le cas pour la Bibliothèque nationale d'Australie ou du Canada.

Le DUBLIN CORE

Le DUBLIN CORE a été conçu par le NSCA (National Center for Supercomputing Applications) et l'OCLC (Online Computer Library Center) en 1995. La création du Dublin Core devait répondre à un objectif : fournir un standard minimal en vue de produire des métadonnées informatiques sur des documents électroniques essentiellement en ligne. Du fait que c'est un standard minimal, le Dublin Core doit pouvoir être utilisé par n'importe qui (bibliothécaire et documentaliste, mais également créateur de site Web). Le Dublin Core comporte donc 15 éléments sous la forme de balises qui peuvent être encodées, nous le verrons, en HTML mais aussi en XML. Ces 15 éléments[26] concernent le contenu du document, les mentions de responsabilité et le type informatique du document (Format, langage…). Concrètement le Dublin Core se présente de la façon suivante : une balise <meta name=> est suivi du nom de l'élément qui va être décrit. Par exemple <meta name= « DC. Title » content= « Site Web de la ville de Valence »> Donne au moins trois informations :

- Qu'il s'agit d'une métadonnée (<meta name …)
- Que cette métadonnée est en Dublin Core (DC)

[26] Présentation des métadonnées Dublin Core en annexe 2.2

- Que l'élément qui va suivre est le titre du document (DC Title)
- et que ce titre est « Site Web de la ville de Valence » (content= « Soite Wb de la ville de Valence »)
- Le signe > indique que la métadonnée est terminée.

A titre d'exemple, voici le description en Dublin Core du site Web de l'enssib :

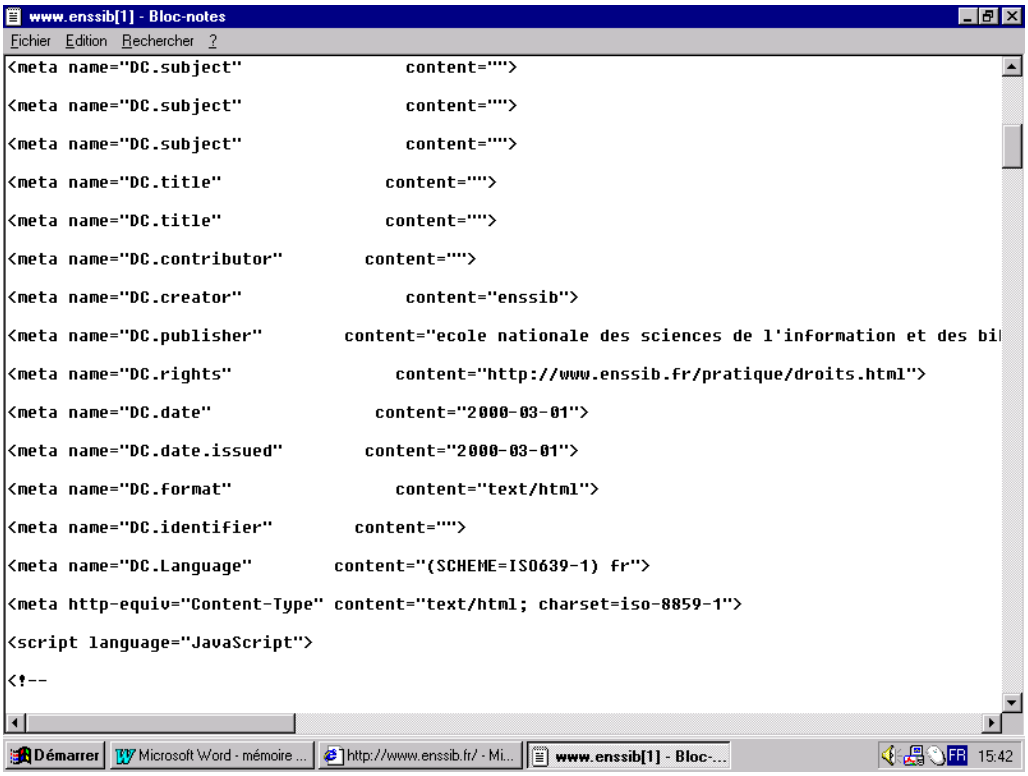

Figure n°2 : Extrait du code source du site Web de l'Enssib

Parmi les nombreux avantages qu'offre le Dublin Core, il faut d'abord mettre en avant sa simplicité. Par ailleurs, il existe d'ores et déjà des outils susceptibles de générer automatiquement un référencement en Dublin Core de sites Web à partir de recherche de mots effectuées dans le texte du site : C'est le cas par exemple du logiciel *« Klarity »*[27]. Par ailleurs, en dépit de sa simplicité, le Dublin Core a été créé en vue de permettre des recherches de documents complexes. Notons que la Bibliothèque Royale de Suède utilise le Dublin Core pour décrire les sites Web qu'elle archive. Il ne nécessite ni un nombre important de personnes, ni une formation particulièrement poussée.

Toutefois, dans le cadre d'une politique d'archivage de sites Web, le Dublin Core ne renferme pas toutes les informations nécessaires. Ainsi, les métadonnées de « format » et de « type » ne suffisent pas à décrire l'environnement technique du site Web de façon satisfaisante. Par ailleurs, la notion d'auteur et donc de responsabilité, si importante dans

[27] **Klarity**
http://www.klarity.com.au/
Dernière consultation le 18/11/02

le catalogage, n'est abordée ici que par celle, plus vague, de « créateur ». Cette dernière est parfois choquante pour les bibliothécaires, bien qu'en toute honnêteté elle corresponde davantage au contexte éditorial du Web.

Pour le moment, malgré sa simplicité, les créateurs de sites Web ne référencent que fort peu leurs sites Web en Dublin Core. De ce fait, pour le moment, il serait illusoire de penser que les bibliothèques pourront récupérer à la fois le site Web et ses métadonnées Dublin Core écrites par le créateur du site. Il faut dire que tous les grands moteurs de recherche ne lisent pas forcément le Dublin Core. Par ailleurs, si tant est que les créateurs référencent leurs sites Web, le feront-ils de façon satisfaisante ?

Les DTD de description

Le XML (eXtensible Markup Language) est un langage à balises issu du SGML et supposé détrôner le HTML sur le Web. Ce langage a été créé par le W3C (World Wide Web Consortium). Sans entrer dans les détails, il est possible néanmoins de dire que le XML est un langage qui offre de nombreuses fonctionnalités. Ainsi, contrairement au HTML qui associe obligatoirement un contenu à une présentation, le XML distingue ces deux éléments ce qui permet pour un même document d'appliquer de nombreuses présentations. Le XML permet de créer de nouvelles balises. Du fait de son extensibilité, le XML est un langage pouvant décrire n'importe quel type de document. Par ailleurs, le XML permet de définir le contenu sémantique d'un document et hiérarchise les données sous la forme d'une arborescence. Concrètement, le XML devrait permettre d'optimiser les recherches sur Internet en développant un véritable Web sémantique par une profonde structuration des données.

Dans le cadre du XML, il faut faire la différence entre les documents XML bien formés et ceux valides. Un document XML bien formé se contente d'appliquer les règles de base du langage XML. Un document valide, par contre, respecte non seulement les règles de base du langage XML mais encore une DTD particulière. On pourrait dire, pour simplifier, qu'une DTD constitue des règles de grammaire structurant le document.

L'avantage du XML, dans le cas précis de la description de documents, est qu'il est possible d'utiliser pour chaque grand type de document une DTD adaptée, tout en se fondant sur le même langage XML. Pour simplifier, on peut dire que le XML permet une description unifiée mais adaptée à chaque type de document. L'EAD (Encoded Archival description), par exemple, est une DTD qui permet de décrire des fonds d'archives particulièrement complexes[28].

---

[28] En effet, des archives ne sont pas constituées d'objets indépendants qui se suffisent à eux-mêmes. Une lettre fait partie d'un dossier qui lui même fait partie d'un fonds… La description de ces fonds d'archives nécessite donc de

Le Bibliothèque du Congrès a développé une DTD particulièrement intéressante : le « METS »[29] (Metadata Encoding and Transmission Standard). Cette DTD a pour objectif de décrire et gérer des documents au sein de bibliothèques électroniques. Elle comporte à la fois des métadonnées descriptives, des métadonnées de gestion, les structures hiérarchiques au sein des collections électroniques.

Le XML offre de nombreux avantages dans le mesure où il est un outil souple et adaptable. Il est ainsi possible d'envisager, par l'utilisation d'une DTD particulière, un outil véritablement adapté au référencement de sites Web archivés. Notons qu'une partie des métadonnées pourra être générée automatiquement. Par contre, l'utilisation du XML de façon systématique pose le problème de la formation du personnel.

4. Vers l'adoption d'une norme ?

Nous avons vu la diversité des solutions qui s'offraient à toute institution susceptible de conserver des sites Web. L'archivage des sites Web existe, pour l'heure, à l'état de projets plus ou moins avancés d'un établissement à l'autre. A ce titre, sur tous les points que nous avons développés plus haut, il n'existe pas, à proprement parler, une norme. Par contre, en ce qui concerne l'organisation globale de l'archivage, une norme ISO est en cours d'élaboration. Elle sera basée sur un système d'organisation des archives numériques créé par la NASA : l'OAIS (Open Archival Information System). Il s'agit donc d'un modèle organisationnel d'archives numériques et non pas d'un système propre à l'archivage des sites Web.

L'OAIS développe toutes les phases de traitement des données numériques à archiver. Les données se présentent sous la forme de « paquets d'informations » qui contiennent à la fois les données brutes et des métadonnées.

L'OAIS divise l'ensemble du traitement en plusieurs procédures appelées entités qui chacune assure le traitement d'un certain type de paquet :

1. **L'entité d'ingestion** : Cette entité de l'institution accepte un paquet d'informations appelées SIP (Submission Information Package). Les SIP sont celles fournies par le

---

pouvoir décrire plusieurs types d'objets (lettres, feuillets, carnets…) en un seul outil et de hiérarchiser ces informations en ensembles et sous-ensembles. C'est justement ce que permet l'EAD.

[29] **Library of Congress.** METS Metadata encoding and transmission standard. In *Library of Congress.* [En ligne]
http://www.loc.gov/standards/mets
Dernière consultation le 04/11/02

producteur de l'information (créateur de site Web, snapshot…) à l'OAIS. L'entité a alors pour mission de recevoir les SIP et de les transformer en AIP (Archival Information Package). Les AIP sont formées de l'association des SIP avec des données appelées PDI (Preservation Description Information) qui sont en fait des métadonnées de description sur les données informatiques à archiver (ici les sites Web)

2. **L'entité d'archivage** : Cette entité reçoit les AIP et s'occupe de leur conservation. C'est cette entité qui s'occupe des différentes migrations que doivent subir les AIP, de vérifier l'état de conservation des données et des supports.

3. **L'entité de gestion des données** : Cette entité permet l'accès à la fois aux métadonnées descriptives sur les documents ainsi qu'aux documents archivés et aux métadonnées de gestion.

4. **L'entité administrative** : Cette entité a la responsabilité de tout le système et de la gestion des droits d'accès aux documents et aux métadonnées.

5. **L'entité de planification de conservation** : Cette entité s'occupe du fonctionnement de l'environnement informatique de l'OAIS et donne des recommandations pour que les documents préservés demeurent accessibles sur le long terme, en dépit de l'obsolescence technique.

6. **L'entité d'accès** sert d'interface avec les usagers. Elle fournit à l'usager des DIP c'est-à-dire une partie des AIP correspondant à la requête de l'usager.

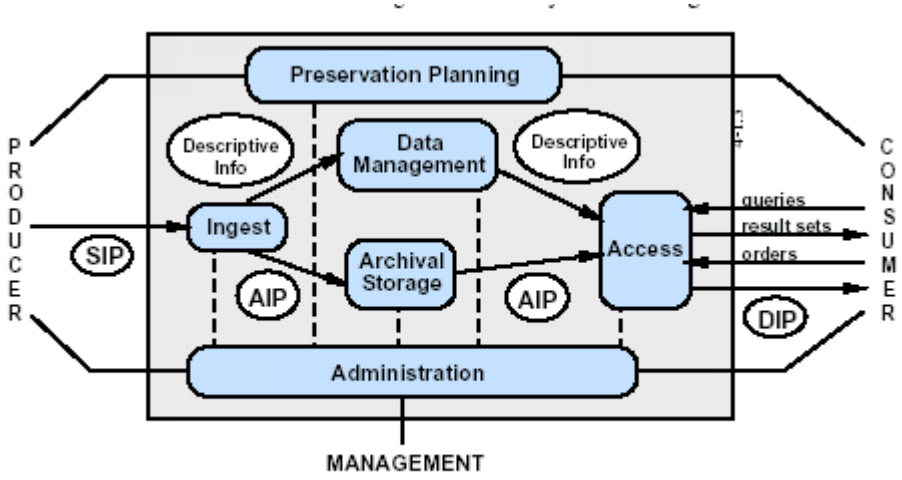

Figure 4-1: OAIS Functional Entities

Figure n°3: Organisation générale de l'OAIS

Le modèle OAIS peut paraître complexe. Dans les faits, il s'agit seulement d'un système de base qui peut être mis en application au sein de n'importe quel projet d'archivage numérique. En effet, au sein de chaque entité les tâches peuvent être assurées par un agent ou en automatique par une machine. Le choix également des métadonnées qui doivent être intégrées dans chaque paquet est également laissé à l'institution d'archivage. Par contre l'intérêt principal de ce modèle est de rappeler l'importance d'associer toute donnée numérique à des métadonnées. Par ailleurs, ce modèle a le mérite de découper l'ensemble de l'organisation de l'archivage en quelques tâches (entités).

Pour l'archivage des sites Web, des solutions sont donc d'ores et déjà envisagées. Cependant, il est important de noter que les grands projets développés pour le moment se situent peu ou prou dans l'optique de la mise en place d'une forme de dépôt légal des sites Web. A ce titre, ces projets sont conduits par de grandes bibliothèques nationales voire par des consortiums de bibliothèques[30]. En outre, on ne saurait oublier qu' à bien des égards, nous sommes encore, pour l'archivage des sites Web, à une ère expérimentale. Toute expérimentation, surtout lorsqu'elle nécessite d'importantes compétences et des moyens informatiques importants, engendre des coûts conséquents. Du fait de ces coûts et du caractère expérimental des projets, pour le moment du moins, seules les bibliothèques nationales et quelques fondations privées se sont penchées sur la question de la conservation des sites Web. Comment alors envisager le rôle d'une grande bibliothèque municipale dans ce type de projet ? Quelle serait la pertinence d'une telle participation ? Quels seraient les risques, en terme de coûts et de personnels, encourus par cette bibliothèque municipale ? Quel en serait l'intérêt pour la bibliothèque ? C'est à cette série de questions que nous allons tenter de répondre dans la troisième et ultime partie.

[30] Voir la présentation de quelques projets en annexe 1.1

Partie 3 :L'archivage des sites Web à intérêt régional : Vers une conservation des sites Web en bibliothèques municipales ?

La réflexion sur la conservation de sites Web commence à dépasser le cadre de la BnF. Si de nombreux bibliothécaires sont aujourd'hui conscients de la nécessité de conserver le Web, la mise en pratique d'une politique d'archivage, au sein d'établissements plus petits que la BnF, doit faire l'objet d'une réflexion et ce, surtout lorsque l'on prend en considération le coût que risque d'engendrer une telle politique. Commencer une réflexion sur le sujet revient sans doute à jauger à la fois l'intérêt d'effectuer la conservation de sites Web en local mais également les paramètres et contraintes qu'il faut prendre en considération dès la conception du projet.

1. Pourquoi envisager l'archivage des sites Web à un niveau local ?

L'intérêt de la conservation des sites Web à un niveau régional se décline par rapport à l'établissement et à ses missions, mais aussi par rapport au Web local et enfin au public concerné.

1.1. Les missions des bibliothèques

Les politiques patrimoniales font partie intégrante des missions de tous les types de bibliothèques. C'est ainsi que la responsabilité des bibliothèques municipales en matière de patrimoine constitue l'article 8 de la « Charte des bibliothèques » (1991). Quant aux bibliothèques universitaires, comme le rappelle Daniel Renoult[31] elles ont, elles aussi, une mission patrimoniale importante.

Cependant, compte tenu des coûts engendrés par l'archivage des sites Web et sa complexité technique, il paraîtrait délicat et certainement peu pertinent que toutes les bibliothèques se lancent dans ce type de projet. En effet, à cause des contraintes techniques et financières de ce type de projet seuls les établissements importants, qui ont

[31] **RENOULT, Daniel.** Les bibliothèques dans l'université. Paris : Editions du Cercle de la Librairie. 1994. 358p.

placé la question patrimoniale comme axe politique fondamental devraient être concernés par ce type de projet. C'est le cas, par exemple de la bibliothèque municipale de Lyon.

1.2 La qualité du Web local

Nous avons déjà parlé de l'intérêt et de la nécessité d'archiver des sites Web en général, compte tenu de leur fragilité et de leur volatilité. Mais qu'en est-il de l'intérêt régional d'un site Web ?
Il ne suffit sans doute pas de conserver les sites Web hébergés par des serveurs situés dans la région pour former une collection de sites Web d'intérêt régional. En effet, un site Web peut présenter un intérêt régional parce qu'il a été créé dans la région, parce qu'il parle de la région...
Par ailleurs, la richesse du Web d'intérêt local ou régional est des plus variables. Il est difficile d'obtenir des statistiques fiables sur la répartition régionale du Web français. Cependant, parmi les statistiques consultées, la région Rhône-Alpes occupe, par exemple, une excellente position. C'est ainsi que selon l' AFNIC (Association Française pour le Nommage Internet en coopération), La région Rhône-Alpes occupait en août 1999, la seconde place pour le nombre de sites Web dans le domaine .fr. Par contre, dans ce même classement, elle se place loin derrière la région Ile-de-France qui a enregistré au moins trois fois plus de sites en domaine .fr :

	Région	domaines .fr
1	Ile de France	14565
2	Rhônes alpes	4219
3	PACA	1763
4	Nord-pas de calais	1638
5	Pays de la loire	1628
6	Midi-pyrénées	1126
7	Alsace	1060
8	Aquitaine	1016

Figure n° 4: classement des régions françaises par nombre de sites Web du domaine .fr

Cette liste est déjà ancienne et ne prend en considération que le domaine .fr. La CCI de Lyon (Chambre de Commerce et de l'Industrie) a pour sa part répertorié 3112 entreprises de la région

ayant un site Web. Par ailleurs, selon une étude de Benchmark group effectuée en 1998, Lyon et Grenoble figuraient parmi les dix premières villes françaises en nombre d'internautes :

Rang	Ville
1	Paris
2	Toulouse
3	Lyon
4	Marseille
5	Strasbourg
6	Grenoble
7	Lille
8	Nantes
9	Montpellier
10	Rennes

Figure n°5: classement des villes françaises par nombre d'internautes

Outre ces considérations quantifiées, le Web rhônalpin présente une extrême variété aussi bien du point de vue des contenus que d'un point de vue formel. A titre d'exemple, les sites Web de municipalités sont particulièrement représentatifs de cette richesse du Web. En effet, de nombreuses municipalités, parfois de taille très réduite, ont aujourd'hui leur site Web. Il arrive souvent, surtout pour les petites villes et les villages, qu'un particulier se charge en bénévole de ce travail. De ce fait, ce type de site Web comporte toutes les caractéristiques d'un site Web personnel. Il est d'ailleurs intéressant de noter

que ces sites présentent un intérêt particulier dans la mesure où ils permettent à des structures très réduites de se faire connaître. Ce que ces municipalités donnent à voir est extrêmement variable : certaines mettent l'accent sur le tourisme, d'autres sur le commerce et l'industrie. Certains villages insistent sur la présence d'une école et d'un tissu associatif dense. Certains sites ne semblent s'adresser qu'au visiteur de passage (un touriste potentiel), d'autres font du site Web un outil de communication et d'information pour leurs habitants. La mise en page est parfois très intéressante puisque certains sites montrent principalement des images de la commune, parfois même la photographie du maire ! D'ores et déjà, ce type de sites nous apprend beaucoup sur la vie politique municipale ; sur la façon dont se construit et se manifeste l'identité collective municipale. Perdre cette information serait d'autant plus tragique que la plupart des mairies et collectivités qui communiquent sur le net n'ont pas les moyens d'éditer un bulletin municipal sur support papier. De ce fait, ces sites Web représenteront parfois un témoignage unique et irremplaçable.

Les sites Web de municipalités présentent un intérêt régional évident. Mais la définition de l'intérêt régional n'est pas sans poser problème pour d'autres types de sites Web. D'un point de vue général on pourrait considérer qu'un document présentant un intérêt régional reflète une situation ou des idées d'un territoire donné. Dans le cadre de l'édition sur support traditionnel, il est relativement aisé d'effectuer un repérage de documents présentant un fort intérêt régional. Prenons l'exemple de la documentation sur les entreprises d'une région. Les rapports d'activité, rapports de stagiaires dans ces entreprises, les annuaires d'entreprises, les articles par exemple, sont des documents intéressants puisqu'ils reflètent une situation régionale économique à un moment donné. Le repérage et l'acquisition de ces documents peuvent être effectués sans grands problèmes. Si l'on essaye de composer une collection équivalente sur le Web, les difficultés augmentent. En effet, la plupart des grandes entreprises régionales ont leur propre site Web sous le nom de domaine . Com très facilement repérables. Toutefois, il existe de grandes disparités entre les sites Web d'entreprises. Certains sites informent véritablement l'internaute sur l'entreprise, ses activités, ses projets et offrent donc un témoignage fondamental sur la vie économique de l'entreprise et de la région dont elle fait partie. Par contre, d'autres sites se présentent plutôt comme des catalogues ou des publicités. Par exemple, on est en droit de s'interroger sur l'intérêt même national que représenteraient l'enregistrement et la conservation du site « Amazon.com » ! La difficulté consisterait donc dans le repérage de sites Web ceux qui présentant ou non un intérêt régional défini. Précisons tout de suite qu'une telle sélection, fine, ne sera pas toujours possible.

Il en est des sites commerciaux, comme des sites Web personnels. Comme nous l'avons vu, certains sites de municipalités peuvent entrer dans la catégorie des sites Web

personnels, pourtant ils représentent un véritable intérêt régional. Nous savons que ce n'est pas le cas pour de nombreux sites Web personnels.
Il ne s'agit pas, à ce point de l'analyse, de donner des recommandations en terme de sélection de sites Web à archiver, mais plutôt de montrer que l'application de critères sélectifs fondés sur une définition établie de l'intérêt régional et adaptés à des documents plus classiques (périodiques, monographies…) ne sera pas évidente dans le cas des sites Web. Nous le verrons, il est cependant fondamental de s'appuyer sur une telle définition, variable d'un établissement et d'un contexte à l'autre, pour envisager un tel archivage.

1.3 L'intérêt de l'archivage par rapport aux publics

Etant donné que l'on se place du point de vue de la conservation de documents, l'intérêt que représente l'archivage de sites Web ne se conçoit pas véritablement au présent et nul ne saurait envisager ce que le public de demain recherchera dans les sites Web conservés, ni quels types de sites seront intéressants. Cependant, comme dit précédemment, la communication via un site Web tend à remplacer, pour certaines structures, un mode de communication sur support plus classique. Comme la presse, les tracts, nombreux sont les documents qui n'ont révélé leur intérêt que bien après leur diffusion. En 2001, par exemple, la Bibliothèque municipale a reçu un don important de tracts, d'affiches et de documents divers témoignant des événements de mai 68 à Lyon. Cette documentation, informelle, sur des supports souvent fragiles, constitue aujourd'hui une source d'analyse et d'étude pour les chercheurs. Alors que ces documents divers étaient, à l'époque, produits et distribués à foison, ils sont devenus, trente ans plus tard, rares. Peut-être s'agit-il d'une des grandes leçons de l'histoire des bibliothèques : c'est ce qui est le plus courant, le plus quotidien, et dont l'intérêt apparaît sur le moment des plus relatifs qui devient, par la suite, un document rare.
S'agit-il pour autant de tout conserver ? S'agit-il, par exemple, de conserver des sites Web commerciaux ? Si la prudence et la peur de voir disparaître certains témoignages pourraient amener certains bibliothécaires à le penser, les contraintes d'un tel objectif l'interdiront.
Il est donc périlleux de prétendre prévoir quel type de sites Web sera le plus recherché par le public dans quelques dizaines d'années. Par contre il est possible de parier sur la qualité d'archivage qui sera demandé. En effet, il paraît vraisemblable d'envisager que les usagers rechercheront le même confort et les mêmes potentialités offerts par le Web actuel. En effet, l'aspect formel des sites, les possibilités de navigation dans le Web archivé seront certainement recherchés par les usagers. A ce titre, il est difficile d'envisager de se contenter de conserver une version imprimée des sites Web. Le choix d'un archivage de qualité orienterait donc les bibliothèques vers un haut niveau d'exigence de conservation,

puisqu'il s'agirait de conserver à la fois les contenus, l'aspect et la plupart des fonctionnalités notamment de navigation hypertextuelle des sites Web.

2. Paramètres, contraintes et décisions : les choix à opérer pour un projet d'archivage du Web à un niveau local

La question de l'intérêt étant posée, il s'agit maintenant d'aborder plus concrètement celle du projet d'archivage lui-même.

2.1 Les paramètres dont il faut tenir compte

Deux types de paramètres sont à considérer en amont de toute réflexion sur un projet d'archivage des sites Web. Certains paramètres sont internes à l'établissement : les missions et l'existant. D'autres paramètres concernent plutôt le contexte français et l'action de la BnF.

2.1.1 Des paramètres internes

Il s'agit presque d'un cliché : tout projet doit se fonder sur l'existant. La question des missions de l'établissement est une donnée fondamentale à prendre en considération. Nous l'avons déjà dit, compte tenu des coûts de l'archivage, seules des bibliothèques considérant dans leurs missions prioritaires la question de la conservation de vraiment se sentir concernées par un projet à grande échelle. D'autre part, la définition précise, au sein de chaque établissement concerné, de ses missions et objectifs doit permettre de délimiter le champ d'application de l'archivage. Par exemple, dans le cas qui nous occupe, il s'agit d'appréhender ce que l'établissement définit comme document d'intérêt régional. Il s'agit bel et bien de penser une politique d'archivage de sites Web comme la continuité de la politique d'acquisition globale de l'établissement. Ce retour vers les missions et objectifs définis de l'établissement sera la condition préalable pour former une collection de sites adaptée et pour délimiter, même de façon grossière, la portion du Web concernée par cet archivage. Par exemple, un CADIST a l'habitude d'acquérir et de conserver des documents en langue étrangère, il sera donc intéressant pour un tel établissement d'archiver des sites Web sur son sujet d'excellence dans des langues différentes. Ce n'est absolument pas le cas pour le fonds régional d'une bibliothèque classée qui conserve en

général des documents en langue française ou régionale et donc le champ d'application sur le Web se limitera vraisemblablement au Web français[32].

La question de l'existant de l'établissement est bien entendu une donnée à prendre en considération. Du point de vue du personnel par exemple, il semble difficile d'envisager qu'un établissement ne possédant pas son propre service informatique puisse se lancer sans risques dans une telle aventure. La conception du projet, comme la maintenance des outils en place nécessitent des compétences en informatique importantes. La sécurité des documents archivés en dépend. Outre le service informatique, le nombre et les compétences du personnel de bibliothèque constituent, nous allons le voir, une donnée importante pour déterminer les options possibles pour l'archivage.

Outre le personnel, l'équipement informatique est un paramètre des plus importants. Au minimum, l'établissement doit bénéficier d'une connexion en réseau de très haut débit. Un réseau déficient ne permettrait pas l'acquisition massive de sites Web parfois volumineux. D'autre part, il ne faut pas oublier que le réseau devra supporter à la fois l'acquisition, mais également les utilisations quotidiennes du personnel et des usagers. De façon caricaturale, il ne s'agirait pas de bloquer le réseau pour l'acquisition de sites Web. L'architecture informatique de l'établissement est un argument de poids qui, comme le nombre du personnel impliqué, oriente les choix en matière d'archivage. A titre d'exemple, l'exécution d'un « snapshot » du Web suédois sollicite le réseau de la Bibliothèque nationale de Suède pendant au moins deux mois. Enfin, toujours du point de vue de l'équipement informatique, l'archivage de sites Web nécessite, au moins à moyen terme, l'acquisition d'un serveur dédié.

2.1.2 Les paramètres externes

2.1.2.1 La tutelle et la question des financements

Nous savons qu'il est, à l'heure actuelle, extrêmement difficile d'évaluer les coûts de l'archivage de sites Web. La direction des Archives de France[33] s'est toutefois penché sur la question des coûts d'enregistrement de fichiers informatiques. Leur simulation est basée sur des enregistrements sur CD-R et sur une moyenne de 100 Go enregistrés par mois ce

---

[32] Il est possible au contrire d'imaginer qu'un fonds régional d'une bibliothèque municipale souhaite compléter ses collections par l'acquisition de sites Web étranger sur sa région. Ce type d'optique n'est pas le propos de cette étude, mais est tout à fait envisageable bien que difficile à mettre en application.

[33] **DHERENT, Catherine.** Les archives électroniques : Manuel pratique. In *Site Web de la Direction des Archives de France.* In *Site Web du ministère de la Culture et de la Communication.* [En ligne].
http://www.archivesdefrance.culture.gouv.fr/fr/archivistique.index.html
Dernière consultation le 23/09/02

qui est certainement inférieur aux quantités qui doivent être gérées dans le cadre de l'archivage de sites Web[34]. La simulation se présente donc ainsi :

➢Pour 100 Go à stocker par mois sur des CD-R (2 euros l'unité). En considérant qu'il faut 30 minutes de traitement pour chaque CD, à partir d'un matériel amorti sur 3 ans. La maintenance équivaudrait à 15% du prix du matériel avec un coût horaire de l'opérateur s'élevant à 15 euros (charges comprises)

➢on aboutit à un coût de 21 euros au Go soit un total de 27 000 euros par an en comptant des frais généraux de 6000 euros.

Il faut bien entendu voir que cette simulation ne tient pas compte des frais d'acquisition (ou de capture), des frais de référencement, des frais de migration ni des frais de mise en accès, qu'il s'agisse des dépenses d'investissement ou des frais de fonctionnement.

Notons que certains logiciels, nécessaires pour l'archivage, sont disponibles gratuitement. Cependant, même gratuite, l'installation de ces logiciels dans des ensembles informatiques aussi complexes que ceux des bibliothèques entraîne généralement un travail important de la part des informaticiens de la bibliothèque. Il s'agit généralement de s'assurer que le logiciel offre une compatibilité optimale avec le reste du système , ce qui nécessite souvent des opérations de paramétrage voire même de programmation, opération qui ont un coût salarial.

L'archivage sera donc une opération coûteuse. C'est d'ailleurs l'un des aspects qui risque de s'avérer problématique pour des négociations en vue de l'obtention de crédits supplémentaires pour mettre en place ce projet. En effet, la navigation sur Internet - Lorsqu'elle est effectuée sur réseaux haut débit et si l'on en déduit les frais d'investissement et de maintenance du réseau - coûte relativement peu. Or comment faire accepter à la tutelle financière que ce mode de communication relativement peu coûteux, permettant d'accéder à des informations souvent gratuites, va devoir nécessiter des budgets conséquents en vue d'assurer sa conservation et que ces frais seront en totalité pris en charge par des fonds publics ? En effet, la disproportion de coût existant entre le prix de l'accès simple au document et le prix de ce même document conservé (et donc accessible sur le long terme) risque d'être un sujet difficile dans le cadre de négociations.

2.1.2.2 La BnF

La BnF s'est lancée depuis quelques années dans un vaste projet de dépôt légal des sites Web. Nous sommes loin de la mise en place effective d'un tel dépôt légal, ne serait-ce que parce que la loi sur la société de l'information n'a pas encore été votée. Toutefois, la BnF

---

[34] 1 Go équivaut à une quantité de 5 à 50 images de format TIFF ou à 200 à 2000 pages en PDF

possède déjà un moteur de capture et de recherche et s'est lancée dans la « collecte » des sites Web du domaine .fr.

Il ne s'agit donc pas de refaire, à un niveau local, ce que la BnF veut faire avec davantage de moyens et de personnels. Toutefois, il ne faudrait pas *a contrario* que ce projet de dépôt légal ne réponde pas aux attentes du public des bibliothèques régionales. C'est ainsi, par exemple, que pour le moment, dans le projet de la BnF, les archives constituées ne seront vraisemblablement pas mises en ligne sur Internet. L'une des raisons de ce choix est d'ordre légal puisque la BnF souhaite recevoir le dépôt de bases de données payantes ou protégées en ligne. Il serait contraire à la loi de proposer un accès gratuit et libre aux bases archivées sur le site Web de la BnF, alors que ce n'est pas le cas sur Internet. En l'état actuel et compte-tenu de cette contrainte, la valorisation des sites Web archivés risque d'être hautement problématique. Les usagers de province devront-ils se déplacer à la BnF pour consulter les archives du Web de leur région?

Les critères de sélection de la BnF peuvent également ne pas répondre aux attentes des publics en régions. En effet, dans le projet actuel, les sites Web personnels ne sont pas acquis par la BnF. Or, nous l'avons vu, certains sites Web de municipalités, très intéressants d'un point de vue régional, sont des sites Web personnels. Il en est de même pour les sites Web d'associations.

Enfin, ne serait-il pas risqué, étant données les difficultés que pose l'archivage des documents électroniques en général, que cette mission ne soit confiée qu'à la BnF ? Ne serait-il pas intéressant d'envisager une conservation partagée ?

Sans parler d'une réelle collaboration, la BnF a lancé en mars 2002, sur la liste de « Biblio.fr » un appel à contribution à destination des autres bibliothèques françaises. Lancée dans une vaste campagne d'archivage des sites Web liés aux élections législatives, la BnF a demandé alors aux bibliothèques françaises de lui transmettre des suggestions de sites Web locaux traitant de la campagne électorale. Si cette démarche ne saurait être perçue comme un projet d'archivage partagé, elle témoigne du moins d'une prise de conscience des responsables du dépôt légal des sites Web de la nécessité de compléter une collecte automatique par une expertise locale, humaine sur les sites Web locaux.

Pour le moment et à notre connaissance, la collaboration entre la BnF et les autres bibliothèques françaises ne fait pas l'objet d'un projet particulier, sinon sous la forme d'opérations ponctuelles telles que l'appel à contribution présenté plus haut. Il ne s'agit pas ici de présenter des modèles d'organisation directement applicables, mais plutôt quelques pistes de réflexion visant à une participation plus importante et constante des bibliothèques françaises dans ce projet global. Il nous semble que l'on peut envisager plusieurs niveaux de collaboration. Chaque niveau correspond à un degré supérieur

d'implication des bibliothèques. Il faut bien préciser que ces pistes ne sont que théoriques et que la BnF ne s'est prononcée sur aucune d'entre elles :

- Au premier niveau, les bibliothèques envoient à la BnF des suggestions de sites Web à archiver[35]. En échange, la BnF offre une interface de consultation sécurisée aux usagers des bibliothèques concernées. Il ne s'agirait donc pas de placer les archives du Web sur Internet, mais bien de permettre, à un niveau local, une consultation au sein des bibliothèques en province. On peut, à l'intérieur de ce premier niveau, concevoir deux sous-niveaux :
    - Ou bien les bibliothèques concernées envoient leurs suggestions par mèls. La BnF analyse alors les propositions et les intègre ou non aux commandes du robot de capture
    - Ou bien les bibliothèques, via une interface sécurisée, interviennent directement sur le robot et lui transmettent leurs commandes. Cette deuxième solution serait certainement plus facile à appliquer pour la BnF, mais implique un accord fort avec les bibliothèques concernées, notamment sur des critères de sélection. Par ailleurs, il faut bien voir que seules quelques bibliothèques pourraient vraisemblablement intervenir de cette manière sur le système de la BnF sous peine de le surcharger.

  Analyse de ce niveau : dans ce cas, les bibliothèques ne participent pas véritablement à la conservation proprement dite. Par contre, elles obtiennent un accès pour la consultation. Du point de vue de la BnF, ce système, bien que lourd à gérer, peut permettre de profiter d'un repérage local plus fin. Par contre, pour les bibliothèques concernées, cette solution nécessite que le personnel ou du moins une partie, ait une connaissance approfondie et suivie du Web local, ce qui n'est pas toujours évident.
- Au deuxième niveau, les bibliothèques concernées suggèrent à la BnF certains sites à archiver comme dans le premier niveau. Cependant, au lieu de se contenter d'une interface de consultation, les bibliothèques reçoivent de la BnF la partie du Web qui les concerne sous la forme d'une cassette DLT ou d'un envoi en réseau. Cette partie du Web correspondrait à ce que les bibliothèques définissent comme d'intérêt régional (ou autre). Il s'agirait donc pour elles de conserver la partie du Web qu'elles auraient acquise et archivée si elles en avaient eu la possibilité (C'est-à-dire si elles avaient eu leur propre robot). La bibliothèque a alors en charge la conservation de cette partie

[35] Précisons que ces suggestions ne concerneraient que les sites Web qui n'ont pas fait l'objet d'une collecte et que ne sont pas déjà archivés.

locale du Web en parallèle avec la BnF qui conserve la responsabilité de l'archivage de la totalité du Web français.

Analyse de ce niveau : de notre point de vue, cette solution offre de nombreux avantages. Elle permet de mettre en place une conservation et une mise en valeur partagées. La BnF y gagne une expertise locale qui lui permettra de compléter efficacement les collections et les bibliothèques concernées obtiendront la maîtrise d'une partie de ces collections ce qui, pour une valorisation à un niveau local, est fondamental. Par contre la conservation de cette partie du Web représentera une charge lourde pour les bibliothèques. C'est pourquoi, il paraît peu vraisemblable que toutes les bibliothèques puissent participer à un tel projet. Par ailleurs, il s'agirait de négocier avec la BnF la périodicité des livraisons. Nous verrons également et sans doute plus concrètement les problèmes que peuvent engendrer ce type d'organisation dans les propositions conçues pour la bibliothèque municipale de Lyon, en annexe.

- Au troisième niveau : il s'agit de reproduire pour les sites Web, le modèle organisationnel du dépôt légal pour les imprimés. En d'autres termes, la BnF s'occuperait de conserver l'ensemble du Web français, alors que les bibliothèques ayant en charge le dépôt légal imprimeur s'occuperaient de récolter et conserver les sites Web hébergés par les serveurs localisés dans la région dont ils ont la responsabilité. Les bibliothèques possèderaient donc leur propre robot d'acquisition et s'occuperaient de la conservation des sites Web obtenus selon une périodicité d'enregistrement à définir.

  Analyse de ce niveau : Il s'agit donc d'appliquer un modèle déjà maîtrisé ce qui est assez avantageux lorsque l'on s'occupe d'un projet aussi novateur. Toutefois, ce niveau d'organisation, même s'il correspond au plus haut degré de participation et de collaboration des bibliothèques n'est pas, nous semble-t il, le plus intéressant. Tout d'abord, il faut bien voir que les collections obtenues ne seraient pas forcément cohérentes. Ensuite, d'un point de vue matériel, la multiplication de robots collecteurs sur le territoire et de procédures de collectes risque de surcharger le réseau national et nécessiterait donc un calendrier très contraignant de collectes. Ensuite, nous savons les difficultés de personnel qu'engendrent en région la maîtrise du dépôt légal imprimeur. La plupart des services du dépôt légal sont en sous-effectif et la tendance n'est pas, pour la BnF, à un accroissement de sa contribution en personnel[36] .

[36] En effet, pour le dépôt légal imprimuer, la BnF offre une participation aux bibliothèques concernées. Cette participation prend notamment la forme concrète de personnels. Or, aucun projet d'augmentation de cette contribution

D'autres possibilités sont envisageables. Il est par exemple possible qu'au premier et deuxième niveau, certaines bibliothèques s'occupent de la collecte régulière de certains sites Web locaux particulièrement mouvants. En effet, la BnF s'engageant dans un processus de collecte automatique, l'acquisition des sites sera relativement espacée. Une bibliothèque peut, à ce titre, souhaiter suivre plus finement certains sites Web, les enregistrer plus souvent et en déposer une copie à la BnF par exemple. Il est également possible que la BnF demande aux bibliothèques concernées de s'occuper des négociations de dépôt avec les créateurs locaux de bases en ligne. Ces bases, nous l'avons vu, ne peuvent être acquises automatiquement. De ce fait, la BnF négocie, en l'absence d'une loi, le dépôt de ces bases directement auprès des créateurs. Ce travail long et fastidieux pourrait être effectué par les bibliothèques localement concernées. Notons toutefois que cette contribution serait lourde pour les bibliothèques municipales. Ce type de projet ne pourrait s'insérer qu'au sein des bibliothèques associées au titre du dépôt légal et ne devrait normalement être effectué que par du personnel d'Etat. Enfin, ce type de contribution nécessiterait des formations lourdes pour l'établissement puisque des compétences en informatiques sont indispensables pour ce type de travail.
Enfin, il est également envisageable qu'une bibliothèque effectue elle-même et en dehors de toute coopération son propre archivage.

2.1.2.3 Le cadre légal

Le cadre législatif est une contrainte forte, à prendre en considération pour l'archivage des sites Web. Le cadre législatif peut intervenir au niveau de l'archivage par au moins deux biais différents :

- Le dépôt légal qui est inclus dans le projet de loi sur la société de l'information et qui fixera les responsabilités de la BnF
- Le droit d'auteur qui est certainement l'aspect le plus problématique pour l'archivage. Les difficultés que pose le droit d'auteur pour l'archivage concernent relativement peu le problème de l'accès massif à des sites protégés (par mot de passe, accès payant…) qui peut être facilement résolu en ajournant pour un temps donné la diffusion des quelques sites Web concernés. Par contre, le problème principal que pose la conservation de sites Web pour les droits de leurs auteurs est que la bibliothèque, en vue d'assurer la pérennité des sites se trouvera dans l'obligation de transformer le site, c'est-à-dire le fichier informatique. Bien entendu, nous le verrons plus précisément, il s'agira pour l'établissement d'archivage de limiter les risques de porter profondément atteinte aux fichiers et donc aux sites Web, sans quoi l'archivage n'aurait plus aucun

---

n'a été présenté par la BnF. En effet , on ne saurait oublier que la BnF subit des contraintes fortes liées au dépôt légal

sens[37]. Cependant, il faut être bien conscient qu'en l'attente de solutions nouvelles (l'émulation notamment), la migration et donc la transformation des fichiers seront une obligation. Le risque de porter atteinte au bien que l'on préserve contredit les missions patrimoniales des bibliothèques mais également le droit d'auteur. Nous pourrions presque dire que le cadre législatif, à l'heure actuelle, interdit presque la conservation des sites Web, du moins tant que la migration s'impose comme solution unique.

3. Vers l'élaboration d'un projet

Que l'archivage s'organise ou non en collaboration avec la BnF ou toute autre structure, un établissement en vue d'élaborer un tel projet est amené à se poser un certain nombre de questions.

Même si, à l'heure actuelle, les projets d'archivage sont encore à un stade plus ou moins avancé d'expérimentation, les établissements désirant s'aventurer dans ce type de projet ne sont pas dépourvus de bases. Ainsi, au niveau organisationnel, le modèle OAIS est un fondement important et normatif sur lequel tout projet doit se fonder : il ne s'agit pas d'inventer ce qui a déjà été bien pensé !

3.1 Missions et champ d'action

Comme pour tout travail d'acquisition, la bibliothèque doit savoir ce qu'elle veut obtenir. Il s'agit, plus ou moins d'établir les critères de sélection et d'acquisition de sites Web. Notons que ces critères seront certainement différents de ceux que les établissements conçoivent pour un annuaire ou un répertoire de signets. Par exemple, dans le cas de la documentation régionale à la bibliothèque municipale de Lyon, du fait que l'annuaire donne un accès libre et non limitatif aux sites Web d'intérêt régional sélectionnés, certains sites sont éliminés d'office de cette sélection. Il s'agit par exemple de sites Web de groupes sectaires religieux ou politiques. Or, dans le cadre d'une collection d'archives, il est tout à fait possible de limiter les droits d'accès à ce type de sites Web problématiques. Ces sites Web ne sauraient faire partie de signets mais par contre, il serait certainement dommage de ne pas les conserver.

D'un point de vue général, pour mettre en place cette politique d'acquisition, l'établissement devra se fonder sur au moins trois questions :

éditeur.

[37] Voir annexe 3 la gestion des risques pour la bibliothèque municipale de Lyon

- Les missions de l'établissement. Nous l'avons vu, comme pour tout projet, l'archivage des sites Web doit entrer dans les missions et les objectifs de l'établissement.
- La politique d'acquisition de l'établissement. Dans le cas qui nous occupe, par exemple, il s'agit de se fonder sur la définition que donne l'établissement de « l'intérêt régional ». Le fait de se baser sur la politique d'acquisition de l'établissement doit permettre par exemple de savoir si les sites Web commerciaux feront l'objet d'une acquisition.
- Une définition du site Web. Il s'agit par exemple de savoir si l'on considère qu'un forum de discussion, une liste de diffusion entrent dans la catégorie des sites Web.

Le but est donc d'établir une grille large de sélection, adaptée à la politique de l'établissement de façon à circonscrire, du moins théoriquement, la portion du Web concernée par le projet d'archivage.

Par exemple, voici un échantillon des questions auxquelles il faudra répondre :

Conservation des sites commerciaux ?	OUI	NON	Réserves
Conservation de sites Web personnels	OUI	NON	Réserves
Conservation des listes de diffusion ?	OUI	NON	Réserves
Conservation de bases de données	OUI	NON	Réserves
Conservation de tous les formats ?	OUI	NON	Réserves
Conservation de sites payants	OUI	NON	Réserves
Niveau de granularité souhaité ?			
Critères d'acquisition (à développer)			
Langue du site Web			
Niveau du contenu			

Ce tableau est loin d'être complet, ne serait-ce que parce qu'il doit être conçu en fonction de la politique d'acquisition générale de l'établissement et de ses objectifs. Vous trouverez en annexe un tableau plus complet effectué pour la bibliothèque municipale de Lyon.

Un tel tableau permettra donc d'établir des critères larges de sélection à appliquer au Web. Ainsi dans le cas où tous les sites commerciaux seraient bannis de la collection sans aucune réserve, il sera inutile de conserver les sites Web dans le domaine .com.

Enfin, dans le cas d'une collaboration avec la BnF, il sera ainsi possible de se rendre rapidement compte des attentes qui, pour l'établissement, ne seront pas comblées par le projet de la BnF, de façon à envisager, le cas échéant une méthode de remplacement.

3.2 Modalités d'acquisition

Nous avons déjà vu les modalités d'acquisition de sites Web :

- La sélection manuelle a le désavantage d'être incomplète mais permet d'obtenir un fonds cohérent. Il s'agit d'une option intéressante pour un établissement dont une partie du personnel est déjà rôdé à l'acquisition de sites Web pour un annuaire ou des signets par exemple. Cette option entre également dans le cadre d'une éventuelle collaboration avec la BnF.
- La sélection automatique est beaucoup plus complète que la sélection manuelle. Toutefois cette option est à déconseiller. En effet, elle nécessite un fort investissement de la part des informaticiens et un temps long de développement des outils et du paramétrage du robot puisqu'il s'agit de retranscrire en algorithmes des critères de sélection et d'acquisition. Dans le cas d'une acquisition de sites Web d'intérêt régional par exemple, il faut bien voir qu'il n'existe pas de nom de domaine dédié aux zones géographiques plus locales que le niveau nationale : il n'existe pas de nom de domaine pour la région Rhône-Alpes ou PACA ! La sélection, au sein du domaine .fr, des sites Web d'intérêt régional, nécessiterait donc un paramétrage complexe. Par ailleurs, la sélection automatique nécessite un réseau informatique très puissant. Cependant, un tel projet est envisageable mais serait plus long et plus coûteux à mettre en place.
- Le dépôt n'est concevable que pour le cas particulier des bases de données en ligne et des sites qui relèvent du Web invisible. Cependant, en dehors des obligations liées au dépôt légal, la bibliothèque devra certainement s'acquitter d'un abonnement pour les sites Web payants. Par ailleurs, ce type d'acquisition sera lourde en personnel.

3.3 Les modalités de conservation

En dehors de la conservation sur support analogique (version imprimée ou microformes) qui ne constitue pas à proprement parler un archivage de sites Web, la migration apparaît comme le seul système qui, à l'heure actuelle, permet d'envisager une consultation pérenne des sites Web. Il ne s'agit pas pour autant de rejeter l'émulation comme nous le montrons dans le projet pour la bibliothèque municipale de Lyon placé en annexe de ce document.

La migration nécessite un haut niveau de connaissance du personnel sur les formats informatiques. A ce titre, il serait tout à fait intéressant de mettre en place au niveau national et international un cadre d'échange d'informations et de connaissances sur les formats. C'est ce que tente de mettre en place la BnF. Il serait ainsi possible de savoir

rapidement quels formats risquent d'être obsolètes, ceux qui devraient les remplacer…L'information sur les formats est véritablement la clé du succès de la migration et il serait dommage que chaque établissement, dans son coin, soit contraint de refaire sans cesse les mêmes recherches.

La migration doit se conduire selon une gestion des risques : risques de ne plus pouvoir consulter le site Web, mais également risque de profondément altérer le site Web concerné. Nous montrons en annexe le plan de calcul des risques que nous avons imaginé pour la bibliothèque municipale de Lyon.

En ce qui concerne les supports d'archivage, si les CD-R présentent l'avantage d'être peu coûteux, leur durée de vie est aléatoire. D'autre part, leur accessibilité nécessite l'emploi de juke boxes ou de prêts manuels. Enfin, leur capacité de stockage est relativement limitée (environ 650 Mo). Les cassettes DLT, au contraire ont une importante capacité de stockage (100 à 200 Go) mais ne permettent pas un accès facile dans la mesure où elles nécessitent l'utilisation d'un équipement particulier (lecteur de cassettes DLT) dont le coût est important. Il serait de ce fait difficile d'envisager qu'un établissement puisse proposer plusieurs ordinateurs ainsi équipés à ses publics. L'utilisation d'un serveur dédié apparaît comme l'une des solutions les plus intéressantes mais uniquement dans le cas où la quantité archivée est imposante. Par contre il sera nécessaire de sécuriser ce serveur dont l'architecture devra être stable (A ce titre, un serveur UNIX apparaît comme une bonne solution).

3.4 Le référencement

Au niveau du référencement, les choix opérés sur les points précédents ont une influence capitale sur les options qui s'offrent alors à l'établissement concerné. Ainsi, le catalogage des sites Web en Unimarc n'est envisageable que pour une sélection manuelle de sites peu nombreux.

D'un point de vue général, il s'agit bien souvent d'établir un choix entre deux types de compétences. Dans le cas d'un référencement automatique avec capture par un robot, le projet sollicitera les compétences des informaticiens. Dans le cas d'une collecte manuelle et d'un référencement en Unimarc le personnel de la filière bibliothèque sera davantage sollicité.

3.5 Les questions de mise en valeur

La question de la forme que prendra la consultation des sites Web archivés est déterminante pour le projet. Si par exemple, l'établissement envisage un accès à la totalité de la portion du Web archivée, l'accès via un serveur s'impose. Si par contre, il s'agit

d'obtenir plusieurs versions d'un même site Web, la conservation sur CD-R peut être envisagée.

Le questionnement sur la mise en valeur rejoint finalement celui sur le champ d'action de la bibliothèque et sur ses critères d'acquisition. Souhaite-t on obtenir une masse représentative du Web sur un domaine ou un territoire donné ? Dans ce cas, un enregistrement par snapshot, qu'il soit effectué par l'établissement lui-même ou par la BnF apparaît comme la meilleure solution. Souhaite-t on suivre certains sites et les enregistrer plus fréquemment ? Peut-être alors faut-il en faire la commande au robot de la BnF ou alors envisager l'enregistrement de ces sites repérés manuellement. Ces deux approches ne sont bien entendu pas contradictoires.

4. Tableau synthétique

Il n'est pas question de donner ici la trame d'un projet qui serait adapté à toute situation, mais plutôt de montrer un certain nombre de grandes phases et d'options possibles présidant à la réflexion pour la mise en place d'un archivage de sites Web :

- **Phase 1 : les critères de sélection**

  Il faut prendre en considération à la fois les missions du service ou de l'établissement concerné et sa politique d'acquisition. L'objectif étant de composer des critères d'acquisition de sites Web à archiver.

- **Phase 2 : Evaluation grossière du champ concerné**

  A partir des critères de sélection, on évalue approximativement la part du Web qui entrerait dans ces critères de sélection. Il est possible de s'appuyer sur les données de l'AFNIC, mais également sur des annuaires spécialisés ou généraux. L'idéal serait bien entendu de connaître également le poids en octets de cette masse approximative, mais cette donnée, à ce stade, sera très difficile à obtenir.

- **Phase 3 : choix des modes d'acquisition**

  A partir de cette évaluation grossière, plusieurs possibilités sont offertes :

1. Le nombre de sites est relativement peu important (100- 300) : il est possible d'envisager une acquisition manuelle. Ceci implique donc l'acquisition d'un logiciel de capture simple et des acquéreurs formés à la sélection de sites Web.
2. Le nombre de sites est trop important. Dans ce cas trois possibilités :

   2.1 une acquisition manuelle demeure envisageable à la condition d'avoir un personnel nombreux

2.2 L'établissement peut envisager d'acquérir son propre robot de collecteur ce qui sera coûteux en investissement, en personnel spécialisé (informaticiens), en temps (temps de développement et de test de l'équipement).

2.3 Une collaboration avec la BnF peut être souhaitable (surtout collaboration de niveau 2 où l'établissement reçoit de la BnF une copie de la portion du Web français qui l'intéresse). Il s'agit donc de négocier avec la BnF les modalités de cette collaboration.

- **Phase 4 : réflexion complémentaire sur la collection et sur la mise en valeur**

Cette phase ne suit pas véritablement la précédente. Elle doit permettre d'affiner les choix opérés lors de la phase 3 et d'anticiper sur ceux de la phase 4.

A ce moment, l'établissement doit réfléchir à certains détails de la collection de sites Web qu'il souhaite obtenir. Il s'agit par exemple de savoir selon quelle périodicité les sites Web seront enregistrés. Dans tous les cas proposés en phase 3, il ne sera pas possible de suivre toutes les transformations d'un même site Web. Par contre, dans le cas d'une collaboration avec la BnF de niveau 1 ou 2, il est tout à fait possible que l'établissement effectue un suivi plus précis et des enregistrement plus fréquents d'une sélection de sites Web. Cette décision doit alors être précédée d'une phase d'analyse et de sélection dont nous donnons un exemple en annexe pour la bibliothèque municipale de Lyon.

Toujours à ce stade de la réflexion, il est intéressant de s'interroger sur le mode de mise en valeur que l'on envisage pour les sites Web archivés. Envisage-t on un accès par CD-R ? Dans ce cas il faut envisager une organisation de cet accès par Juke-Boxes par exemple….

Il est intéressant également de réfléchir sur les modalités de recherche que l'on souhaite offrir pour les usagers futurs. Ceux-ci devront-ils passer par le catalogue de la bibliothèque ? Comment le catalogue présentera-t il les différentes versions d'un même site Web ? Comme un périodique par exemple ? Au contraire, la recherche s'effectuera-t elle via un moteur de recherche ?

- **Phase 5 : Référencement**

Les choix opérés lors des phases 3 et 4 déterminent largement les options possibles à ce niveau de la réflexion.

A l'heure actuelle, le catalogage en Unimarc n'est possible que dans le cas d'une sélection manuelle et pour un nombre relativement réduit de sites, à moins que l'établissement n'ait un nombre très conséquent de catalogueurs.

La mise en place d'un moteur de recherche peut nécessiter un coût supplémentaire de sous-traitance ou de personnel informaticien.

- **Phase 6 : le choix des supports**

  Ce choix dépend de la mise en valeur que l'on envisage en phase 4 mais également de données techniques comme la durée de vie du support, la capacité de stockage et le coût à l'unité.

- **Phase 7 : l'organisation du traitement**

  L'organisation générale du traitement des documents doit respecter la norme OAIS. A l'intérieur de l'application du modèle, il s'agit alors de répartir les tâches tout d'abord entre les agents humains et les tâches effectuées automatiquement par des agents logiciels. Par ailleurs, les tâches des agents humains doivent être réparties entre différents types de compétences : informatiques ou documentaires. Les choix opérés en amont conditionnent bien entendu ces différentes répartitions. Ainsi, dans le cas d'une sélection manuelle avec catalogage en Unimarc, il est évident que la charge de travail humain et documentaire sera largement supérieure.

  L'organisation doit également permettre d'envisager les différents niveaux de responsabilité des agents.

En conclusion, une grande bibliothèque municipale, une BMVR ou une bibliothèque universitaire souhaitant travailler sur un projet d'archivage de sites Web, sera confronté aux mêmes questions que les grands établissements nationaux en charge de ce type de projet. Cependant, même si nous avons montré l'ensemble des possibilités offertes aux établissements, il faut bien voir que, de façon évidente, le niveau de faisabilité n'est pas le même selon que l'on se trouve à la BnF ou dans des établissements de taille plus réduite, du moins pour le moment. L'archivage de sites Web est aujourd'hui dans une phase pionnière. Il n'existe que très peu de normes en la matière et les outils informatiques nécessaires ne sont pas produits industriellement, mais plutôt au cas par cas. Faut-il pour le moins considérer que les grandes bibliothèques municipales n'ont aucun rôle à jouer dans ce type de projet ? Bien au contraire. L'expertise des bibliothèques municipales au niveau régional, les connaissances acquises sur les régions, le public, les collections et les acteurs du Web local sont indispensables pour un projet global de conservation des sites Web à intérêt régional. Par ailleurs, il serait dommage de ne pas envisager une mise en valeur des sites Web archivés au niveau local. Cependant, ne serait-il pas risqué pour un établissement de se lancer seul dans ce type de projet même localisé ? Il est possible que d'ici quelques années, l'expérience acquise par les établissements nationaux en matière d'archivage, le développement d'outils informatiques adaptés, puissent permettre aux bibliothèques municipales d'envisager plus facilement la mise en place d'un tel projet. S'agit-il alors d'adopter une attitude attentiste sur le domaine ? En fait, il serait fondamental que ces bibliothèques puissent participer aux projets de la BnF en matière

d'archivage du numérique et ce, d'autant plus que le projet de dépôt légal du Web français, en est encore à une phase de préparation et alors que l'obligation légale de résultats ne tardera pas à venir[38]. Le niveau de participation des bibliothèques locales reste à définir, mais il serait important qu'elles puissent faire état de leurs attentes et de leurs compétences à agir.

[38] Le vote de la loi sur la société de l'information et sur le dépôt légal du Web est prévu pour janvier 2003 à l'Assemblée Nationale.

Conclusion

La courte histoire de l'Internet a suscité bien des débats au sein des bibliothèques. Aujourd'hui, il semble bien que nous assistions à une nouvelle étape dans ces discussions. La question de la légitimité de l'Internet en bibliothèque et du rôle des bibliothécaires face à la diversité de l'offre sur Internet, largement débattue, laisse peu à peu la place à une autres problématique : celle de la conservation. A ce titre, il est intéressant de voir que les polémiques autour du Web renvoient presque immanquablement à un questionnement plus fondamental sur les missions même des bibliothèques. La question de la mise à la disposition du public de postes connectés, celle de la constitution ou non de signets, se rapportaient au rôle fondamental des bibliothèques en matière d'offre de documents sélectionnés et de politique d'acquisition. Aujourd'hui, les problèmes d'archivage reflètent la contradiction ressentie par les bibliothécaires dans le fait d'offrir des documents sur un très court terme, sans possibilité de recul, démentant ainsi les missions patrimoniales des bibliothèques.

Les questionnements sur la légitimité de l'Internet et sur sa conservation ne sont pas antinomiques, bien au contraire. En effet, l'absence de pérennité des sites Web, leur caractère éphémère, participent de la difficulté que nous éprouvons à considérer ceux-ci comme légitimes. Quelle valeur accordons-nous à ce qui ne laisse aucune trace ? L'intérêt documentaire, artistique ou historique de certains objets ne se donne pas à voir directement. L'analyse *a posteriori* de ces objets en révèle parfois tout l'intérêt. La légitimité et l'intérêt d'un document viennent aussi de ces lectures successives mais encore faut-il que ces documents soient conservés.

Le défi que pose la conservation de sites Web mais également de tout document numérique est d'abord d'ordre technique. Nous l'avons vu, la structure même des documents informatiques, leur dépendance à des programmes condamnés à une obsolescence rapide, rendent leur conservation sur le long terme problématique. Il est d'ailleurs intéressant de voir qu'alors que le vocabulaire informatique utilise en abondance la notion de mémoire (capacité de mémoire, mémoire vive…), l'informatique est sans mémoire, ou du moins exempte de mémoire sur le long terme.

Enfin, pour le moment, la conservation des documents électroniques induit des opérations qui contredisent en partie les missions patrimoniales des

bibliothèques. A l'heure actuelle la migration est l'opération permettant la sauvegarde des fichiers informatiques. Or, elle nécessite parfois une transformation des fichiers à sauvegarder. Les bibliothèques seront amenées à faire des choix parfois risqués : il s'agira par exemple de sacrifier les fonctionnalités de certains documents de façon à en sauvegarder le contenu ou l'apparence. Ces choix pragmatiques démentent les missions de préservation de biens inaliénables et, en ce sens, constituent une rupture épistémologique et surtout éthique pour les bibliothèques.

Compte-tenu de toutes ces difficultés, qui doit alors procéder à la conservation du Web ? Des fondations privées comme Internet Archive, avec le risque de voir se développer un accès payant à ces archives ? Les bibliothèques nationales et elles seules ? Pour notre part, si rien ne saurait remplacer l'expertise des grandes bibliothèques nationales, il serait dommage que sur un même territoire, d'autres bibliothèques-relais ne participent pas à cette entreprise.

Enfin, plus que l'archivage du Web, c'est bien la question de la conservation des documents électroniques en général qu'il s'agit de poser. Les bibliothèques se sont lancées dans de vastes opérations de numérisations d'œuvres fragiles de façon à en assurer une meilleure diffusion. Ne faut-il pas également réfléchir à la conservation de ces documents électroniques sous peine de se trouver dans l'obligation de réitérer des opérations de numérisation coûteuses et risquées pour les documents d'origines ? Les bibliothèques conservent parfois d'importantes collections de manuscrits, de brouillons d'écrivains et de correspondances. L'informatique a aujourd'hui pénétré tous les domaines y compris ceux de l'art et les écrivains d'aujourd'hui n'hésitent pas à utiliser des traitements de texte. Il est donc vraisemblable que les fonds d'archives de demain renfermeront des « manuscrits électroniques », des mèls… A ce titre, quel peut être le statut d'un « manuscrit électronique » sans ratures, ni brouillons successifs ? Cette question anodine est peut-être le fonds d'un problème plus fondamental : Comment envisager une politique de conservation dans un cadre où les pratiques d'écriture et de production de documents sont aujourd'hui bouleversées ?

Bibliographie

Généralités

ASCHENBRENNER, Andreas. Long-term preservation of digital material: Building an Archive to preserve digital cultural heritage from the Internet (Master thesis). [En ligne]
http://www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/
Dernière consultation le 13/11/02

BRODIE, Nancy. Collaboration entre les bibliothèques nationales en vue de conserver l'information numérique. *Nouvelles de la Bibliothèque nationale.* Mars 1999, vol.31, n°3-4. [En ligne]
http://www.nlc-bnc.ca/9/2p2-1993-03-f.html
Dernière consultation le 12/09/02

BULLOCK, Alison. La conservation de l'information numérique : ses divers aspects et la situation actuelle. *Flash Réseau.* 1999, n°60. [En ligne]
http://www.nlc-bnc.ca/9/1/p1-259-f.html
Dernière consultation le 12/09/02

DHERENT, Catherine. L'archivage à long terme des documents électroniques en France. In *7° conférence Microlib 2000. Lisbonne, Juin 2000.* [En ligne]
http://www.archivesdefrance.culture.gouv.fr/fr/notices/archi2.html
Dernière consultation le 12/09/02

DHERENT, Catherine. Les archives électroniques : Manuel pratique. In *Site Web de la Direction des Archives de France.* In *Site Web du ministère de la Culture et de la Communication.* [En ligne].
http://www.archivesdefrance.culture.gouv.fr/fr/archivistique.index.html
Dernière consultation le 23/09/02

ESTERMANN, Yolande et JACQUESSON, Alain. Quelles formations pour les bibliothèques numériques ? *Bulletin des Bibliothèques de France.* 2000, t.45, p.4-17.
[En ligne]

http://bbf.enssib.fr/bbf/html/2000_45_5/2000-5-p4-estermann.xml.asp
Dernière consultation le 07/10/02

GAUDIN , Frédérique. Quelle normalisation pour les documents numérisés en vue d'une conservation et d'une consultation à long terme ? *Document numérique.* 2000, vol.4, n°3-4, p.199-217.

HAIGH, Susan. Glossaire des normes, des protocoles et des formats liés à la bibliothèque numérique. *Flash Réseau.* Mai 1998, n°54. [En ligne]
http://www.nlc-bnc.ca/9/1/pl-253-f.html
Dernière consultation le 12/09/02

HODGE, Gail et CAROLL, Bonnie C. Digital electronic archiving: The state of art and the state of practice. In *ICSTI.* [En ligne]
http://www.icsti.org/conferences.html
Dernière consultation le 20/10/02

JACQUESSON, Alain et RIVIER, Alexis. Bibliothèques et documents numériques : Concepts, composantes techniques et enjeux. Paris : Editions du Cercle de la Librairie. 1999, 377p. Coll. Bibliothèques.

KENNEY, Anne R. European libraries create framework for networked deposit library. *CLIR issues.* Mars-Avril 2001, n°20. [En ligne]
http://www.clir.org/pubs/issues/issues20.html#european
Dernière consultation le 18/10/02

KLEINBERG, J et LAWRENCE, S. The structure of the Web. In *Cornell University.* [En ligne]
http://www.cs.cornell.edu/home/kleinber/sci01.pdf
Dernière consultation le 06/11/02

LEE, Kyong-Ho et al. The state of art and practice in digital preservation. *Journal of Research of the National Institute of Standards and Technology.* Janvier-février 2002, vol.107, n°1, p.93-106. [En ligne]
http://www.nist.gov/jres
Dernière consultation le 07/10/02

LUPOVICI, Catherine. Les stratégies de gestion et de conservation des documents électroniques. *Bulletin des Bibliothèques de France.* 2000, T.45, n°4, p.43-54.
[En ligne]
http://bbf.enssib.fr/bbf/html/2000_45_4/2000-4-p43-lupovici.xml.asp
Dernière consultation le 12/09/02

LUPOVICI, Catherine. Les besoins et les données techniques de préservation. In *67th. IFLA council and general conference. August 16th-25th. 2001.* [En ligne]
http://www.ifla.org/IV/ifla67/papers/163-168f.pdf
Dernière consultation le 07/11/02

LUPOVICI, Catherine. Les principes techniques et organisationnels de la préservation des documents numériques. Actes du 31° congrès de l'ADBU à l'Université de Provence, le 14/09/01. In *Site de l'ADBU.* [En ligne]
http://www-sv.cict.fr/adbu/actes_et_je/je2001/cathLUPO_140901.html
Dernière consultation le 07/10/02

LUPOVICI, Christian. La chaîne de traitement des documents numériques. *Bulletin des Bibliothèques de France.* 2002, t.47, n°1, p.86-91.
[En ligne]
http://bbf.enssib.fr/bbf/html/2002_47_1/2002-1-p86-lupovici.xml.asp
Dernière consultation le 07/10/02

LYMAN , Peter. Archiving the World Wide Web. In *Clir.* [En ligne]
http://www.clir.org/pubs/reports/pub106/web.html
Dernière consultation le 07/10/02

MANNERHEIM, Johan. The WWW and our digital heritage: The new preservation tasks of the library community. In *66th. IFLA Council and general conference. Jerusalem, Israel, 13- 18th August*
http://www.ifla.org/ifla/IV/ifla66/papers/158-157e.htm
Dernière consultation le 17/09/02

MARTIN, Julia et COLEMAN, David. The archive as an ecosystem. In *Michigan University*. [En ligne]

http://www.press.umich.edu.jep/07-03/martin.html
Dernière consultation le 19/10/02

OCLC (Online Computer Library Center). Web characterization. In *OCLC's Website.*
http://wcp.oclc.org
Dernière consultation le 29/10/02

PASCON, Jean-Louis et POTTIER, Isabelle. Archivage électronique : Aspects technique et juridique. Paris : AFNOR, 2000, 83p. Coll. « AFNOR pratique ».

SANTANTONIOS, Laurence. La bibliothèque de Babel Web. *Livres Hebdo.* Vendredi 15 février 2002, n°457, p. 62-63.

THIBODEAU, Kenneth. Building the archives of the future : Advances in preserving electronic records at the National Archives and Records Administration.*D-Lib Magazine.* Février 2001, vol.7, n°2. [En ligne]
http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html
Dernière consultation le 17/09/02

THOMAS , Michel. Archivage électronique et normalisation. *Document numérique.* 2000, vol.4, n°3-4, p.219-232.

WELF-DAVELAAR, Titia Van der. Long term preservation of electronic publications. *D-Lib Magazine.* Septembre 1999, vol.5, n°9. [En ligne]
http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html
Dernière consultation le 17/09/02

Les expériences d'archivage de sites Web

En Australie

CATHRO, W., WEBB, C. et WHITING, J. Archiving the Web: The PANDORA archive of the National Library of Australia. In *National library of Australia.* [En ligne]

http://www.nla.gov.au/nla/staffpaper/2001/cathro3.html
Dernière consultation le 07/10/02

National Library of Australia. Pandora Archive: PANDAS Manual. In *The National Library of Australia.* [En ligne]
http://pandora.nla.gov.au/manual/pandas/general.html
Dernière consultation le 07/10/02

National Library of Australia. Digital library of Australia. In *the National Library of Australia.* [En ligne].
http://www.nla.gov.au/dsp/
Dernière consultation le 17/10/02

National Library of Australia. National strategy for provision of access to australian electronic publications: A national library of Australia position paper. In *The National Library of Australia.* [En ligne].
http://www.nla.gov.au/policy/paep.html#eleven
Dernière consultation le 17/10/02

Aux Etats-Unis

BREWSTER, Kahle. The Internet Archive. *RLG DigiNews.* Juin 2002, Vol.6, n°3. [En ligne].
http://www.rlg.org/preserv/diginews/diginews6-3.html
Dernière consultation de 12/10/02

Internet Archive. Internet Archive. [En ligne]
http://www.archive.org/
Dernière consultation le 04/01/02

Library of Congress. The national digital information infrastructure preservation program. In *Library of Congress.* [En ligne]
http://www.digitalpreservation.gov/ndiipp/
Dernière consultation le 02/12/02

En France

MASANES, Julien. The BnF'sproject for Web archiving. Contribution for the *European Conference* on digital libraries (ECDL) 2001: What's next for digital deposit libraries? Darmstadt, 8 Septembre 2001. [En ligne]
http://bibnum.bnf.fr/ecdl/2001/france/sld001.htm
Dernière consultation le 04/11/02

En Grande-Bretagne

JENKINS, Clare. Presentation of the CEDARS project. In *CEDARS.* [En ligne]
http://leeds.ac.uk/cedars/
Dernière consultation le 17/10/02

SHENTON, Helen. From talking to doing : Digital preservation at the British Library. In *RLG*. [En ligne].
http://www.rlg.org./events/pres-2000/shenton.html
Dernière consultation le 19/10/02

En Suède

MANNERHEIM, J., ARVIDSON, A et PERSSON, K. The Kulturawr3 project. Contribution for the 66th IFLA general conference, Jerusalem, 13-18 août 2000. In *IFLA.* [En ligne]
http://www.ifla.org/IV/ifla66/papers/154-157e.htm
Dernière consultation le 17/09/02

Royal Library of Sweden. Kulturarw3 Heritage project. In *Royal library of Sweden.* [En ligne]
http://www.kb.se/kw3/ENG/Default.htm
Dernière consultation le 06/11/02

Les solutions techniques envisagées pour l'archivage

BEARMAN, David. Reality and chimeras in the preservation of electronic records. *D-Lib Magazine.* Avril 1999, Vol.5, n°4. [En ligne]
http://www.dlib.org/dlib/april99/bearman/04bearman.html
Dernière consultation le 07/10/02

BORTZMEYER, Stéphane et PERRET, Olivier. Versionnage : Garder facilement trace des versions successives d'un document. *Document numérique.* 2000, vol.4, n°3-4, p. 253-264.

GRANGER, Stewart. Emulation as a digital preservation strategy. *D-Lib Magazine.* Octobre 2000, vol.6, n°10. [En ligne]
http://www.dlib.org/dlib/october00/granger/10granger.html
Dernière consultation le 17/10/02

LAWRENCE, W. Gregory et al. Risk management of digital information: A file format investigation. In *CLIR's Website.* [En ligne].
http://www.clir.org/pubs/reports/pub93/contents.html
Dernière consultation le 19/10/02

WHEATLEY, Paul. Migration: a CAMILEON discussion paper. In *CAMILEON.* [En ligne]
http://www.ariadne.ac.uk/issue29/camileon/
Dernière consultation le 19/10/02

Les Métadonnées

BnF. Information pour les professionnels: Description bibliographique internationale normalisée des ressources électroniques ISBD(ER. In *Site de la BnF.* [En ligne]
http://www.bnf.fr/pages/zNavigat/frame/infopro.htm
Dernière consultation le 07/10/02

CEDARS (Curl Exemplars in Digital ARchiveS). Metadat for digital preservation: the CEDARS project outline specification. In *CEDARS.* [En ligne].
http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html
Dernière consultation le 19/10/02

DHERENT, Catherine. Une DTD pour la description des fonds d'archives et collections spécialisées, l'EAD. In *Site des Archives de France.* In *Site du ministère de la Culture et de la communication.* [En ligne]
http://www.archivesdefrance.culture.gouv.fr/fr/archivisitique/pr%E9sentationEAD.html
Dernière consultation le 21/10/02

DUVAL, Erick et al. Metadata principles and practicabilities. *D-Lib Magazine.* Avril 2002, Vol.8, n°4. [En ligne]
http://www.dlib.org/dlib/april02/weibel/04weibel.html
Dernière consultation le 19/10/02

IFLA. UNIMARC Guidelines n°6 : Electronic resources. In *IFLA.* [En ligne].
http://www.ifla.org/VI/3/p1996-1/guid6.htm
Dernière consultation le 19/10/02

Library of Congress. METS Metadata encoding and transmission standard. In *Library of Congress.* [En ligne]
http://www.loc.gov/standards/mets
Dernière consultation le 04/11/02

LUPOVICI, Catherine et MASANES, Julien. Metadata for long-term preservation. In *NEDLIB.* [En ligne]
http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm
Dernière consultation le 04/11/02

L'OAIS

Consultative Committee for Space and Data System (CCSDS). Reference model for an Open Archival Information System (OAIS): blue book. In *CCSDS.* [en ligne]
http://www.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf
Dernière consultation le 10/09/02

HODGE, M. Gail. Best practices for digital archiving. In *University of Michigan.* [En ligne].
http://www.press.umich.edu/jep/05-04/hodge.html
Dernière consultation le 18/09/02

HOLDSWORTH, D. A blueprint for representation information in the OAIS model. In *CEDARS' website.* [En ligne].
http://www.personnal.leeds.ac.uk/~ecldh/cedars/iee00.html
Dernière consultation le 17/09/02

HUC, Claude. Le modèle de référence pour les systèmes ouverts d'archivage. *Document numérique.* 2000, vol.4, n°3-4, p.233-251.

LAVOIE, Brian. Meeting the challenges of digital preservation: The OAIS reference model. In *OCLC's website.* [En ligne].
http://www.oclc.org/research/publications/newsletter/repubs/lavoie243/
Dernière consultation le 17/09/02

Les aspects juridiques

BENSOUSSAN, Alain (dir.). Internet : Aspects juridiques. Paris : Hermès, 1998, 2° éd. revue et augmentée, 247p.

CAHEN, Murielle. Aspiration de sites Web. In *Avocat Online.* [En ligne]
http://www.murielle-cahen.com/p_aspiration.asp
Dernière consultation le 19/09/02

Table des annexes

Annexe 1 : Présentation de projets d'archivage de sites Web

Cette annexe vous dresse un tableau de présentation de quelques grands projets d'archivage de sites Web.

Annexe 1-1
Quelques exemples de projets nationaux et internationaux d'archivage de sites Web

1. Le Cas de la BnF

L'objectif de la BnF our l'archivage de sites Web est d'aboutir à la mise en place d'un dépôt légal du Web français. De ce fait c'est la notion d'exhaustivité qui fait office de fil conducteur du projet.

Le projet de la BnF se développe à partir de deux procédures d'acquisition de sites Web :

- L'acquisition massive du Web français directement accessible par le biais d'un robot qui établit des snapshots
- Le dépôt de certaines bases de données et sites Web appartenant au Web invisible et échappant à toute collecte automatique.

La définition du Web français s'appuie sur plusieurs critères : la langue, le nom de domaine sachant qu'il ne s'agit pas d'une donnée suffisante, l'adresse du site… Ces critères doivent permettre de fixer le périmètre du Web français.

1.1 Le robot de la BnF

Le robot de la BnF est appelé « moteur du dépôt légal ». Cet outil a été développé en collaboration avec l'INRIA et la « start-up » « Xyleme ».

Cet outil a un certain nombre de fonctions :

- Le repérage : il s'agit de dresser la liste des pages et des sites français et en estimer l'importance en fonction du nombre de liens pointant vers eux. Cette estimation s'appuie donc sur l'indice de notoriété du site Web, méthode également utilisée par le moteur de recherche « google » pour classer les réponses pertinentes (c'est ce que l'on appelle l'indice de notoriété). Cette fonction de repérage doit permettre en outre de recenser les obstacles à une collecte automatique c'est-à-dire tous les sites relevant du Web invisible.
- L'échantillonnage : A partir de ces informations, le moteur établit automatiquement un échantillon représentatif du Web français actuel. Cet échantillon archivé constituera le premier degré de l'archive d Web français.
- La notification : à partir de la liste des sites Web dont l'acquisition automatique serait impossible (Web invisible), le moteur notifie aux services traditionnels du dépôt légal l'existence de ces sites importants.

1.2 Le dépôt

Le dépôt de sites Web ne concerne donc que la liste des sites Web notifiés par le robot. N'oublions pas que près de 40% des sites Web ne sont pas accessibles par un robot. A partir de ce moment c'est le personnel du dépôt légal du Web qui prend le relais de l'électronique. La chaîne opérationnelle se présente alors de cette façon :

- Phase 1 : identification et contact. Cette phase consiste en un repérage des éditeurs des sites Web en question. Ces personnes serviront d'interlocuteurs. A ce stade les compétences requises sont des compétences documentaires et une bonne connaissance des acteurs du Web. Pour le moment, en l'absence d'un loi sur le dépôt légal de sites Web et donc d'une obligation de déposer, la BnF transmet à ces interlocuteurs une demande d'accord.
- Phase 2 : phase contractuelle. Une fois l'accord obtenu, une convention est signée par les éditeurs. Cette convention, conçue par les services juridiques de la BnF n'existe, là encore, que parce que les éditeurs ne sont soumis à aucune obligation légale de déposer.
- Phase 3 : la phase de spécification. Cette phase correspond à un entretien pour déterminer exactement la nature et l'étendue de ce qui sera déposé par l'éditeur. Il s'agit donc de tenir compte de l'analyse des acquéreurs mais également des contraintes techniques du site Web. L'extrême diversité dans l'architecture des sites Web et l'hétérogénéité des formats induisent la nécessité de procéder à un entretien particulier pour chaque site. A ce niveau, des compétences documentaires et informatiques sont requises.
- Phase 4 : livraison. Le site est alors livré sous la forme d'un CD, d'un DVD ou par FTP.
- Phase 5 : validation. Il s'agit de vérifier l'état de la livraison au niveau logique comme au niveau physique.
- Phase 6 : archivage. Une fois vérifié, le site est sauvegardé sur cassette DLT.

Vue la complexité de ces différentes phases, l'archivage de chaque site nécessite trois jours de travail.

1.3 l'avancée du projet

Comme nous l'avons vu, l'absence d'une loi sur le dépôt légal des sites Web rallonge les procédures de dépôt pour les sites Web qui ne peuvent être collectés automatiquement. Pour le moment, cette extension du dépôt légal au Web est intégrée au projet de loi sur la société de l'Information.

D'autres aspects restent encore à définir. Par exemple, la forme que prendra l'interface de consultation des archives n'est pas encore décidée. En tout cas, pour le moment, il est prévu que les archives ne seront pas consultables en ligne. Le référencement des sites et leur traitement devraient être effectués en XML.

L'intérêt de l'organisation de l'archivage par rapport à d'autres projets est qu'il compense les insuffisances d'une collecte manuelle par une collecte automatique (snapshot). De la même façon, l'incomplétude de la collecte automatique est compensée par le dépôt manuel.

Pour le moment, la BnF s'est lancée dans plusieurs actions d'archivage qui ne correspondent pas encore à l'application complète d'un dépôt légal. Parmi les grandes campagnes d'archivage effectuées, la BnF s'est lancée dans la collecte et l'archivage des sites Web électoraux pendant les élections législatives et présidentielles de 2002. Dernièrement, elle a effectuée la capture et l'archivage des sites Web du domaine .fr.

2. Un exemple de sélection manuelle : Le cas Australien

La bibliothèque nationale d'Australie s'est penché sur le problème de la conservation de documents numériques dès 1994. Le projet PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) concerne plus particulièrement l'archivage des sites Web.

2.1 Des critères de sélection

Face à l'offre pléthorique de sites Web et aux difficultés d'assurer un archivage exhaustif, la bibliothèque Nationale a fait le choix de sélectionner les sites Web qu'elle souhaite acquérir. Cette sélection manuelle se fait selon des critères clairement énoncés, entrant de ce fait dans une politique d'acquisition. Avant tout, les sites sélectionnés doivent concerner l'Australie et doivent être créés par un Australien (Ou du moins un Australien doit faire partie des mentions de responsabilités du site). Le site doit en outre porter sur un sujet d'ordre social, politique, culturel, religieux, scientifique ou économique. Par contre la localisation du serveur hébergeant le site n'a aucune importance. Une priorité absolue est donnée aux publications dont le contenu scientifique est prouvé et dont l'intérêt se conçoit sur le long terme. La qualité du contenu fait donc l'objet d'un examen attentif dans la mesure où un site dont le contenu ne donnerait qu'une information superficielle ne sera pas retenu.

Certains sujets particuliers à la culture australienne font l'objet de critères de sélection particuliers. C'est le cas, par exemple, des sites sur les Aborigènes, le centenaire de l'Etat fédéral etc.

Le contenu du site Web sélectionné doit également apporter un supplément informationnel par rapport à des publications sur support papier.
Les sites Web personnels ne sont que très rarement sélectionnés. La valeur scientifique du site doit être constatée et le contenu informationnel ne doit pouvoir être trouvé dans aucun texte imprimé.
Enfin, le choix d'archiver les liens hypertextes vers d'autres sites Internet s'effectue en fonction des critères de sélection exposées plus haut : Si le site Web vers lequel pointe le site archivé correspond aux critères de sélection, il sera archivé et dans le cas contraire il ne le sera pas.
La périodicité d'enregistrement des sites sélectionnés s'effectue là encore en fonction des types de publication définis par la bibliothèque et d'un coefficient de changement attribué par la bibliothèque.

2.2 Organisation

La collecte et la préservation des sites Web s'effectue sur la base d'un réseau associant la Bibliothèque nationale, la bibliothèque de Victoria, la bibliothèque d'Australie du Sud, la Library and information service of Western Australia et ScreenSound Australia. Quant à la bibliothèque de Tasmanie elle organise son propre projet d'archivage des sites tasmaniens(*« Our Digital Island"*[39]). La bibliothèque de Galles du Sud développe également son propre projet d'archivage.
La bibliothèque nationale a développé son propre système en vue d'attribuer à chaque site australien un identificateur unique et persistant.
Chaque site Web archivé est catalogué au format MARC.
La liste des archives de PANDORA annonce 3187 sites[40] (à multiplier par le nombre de versions enregistrées par titre).
La conservation proprement dite passe pour le moment par des migrations successives, mais la bibliothèque souhaite, dès que les outils existeront, utiliser également des émulateurs.
Depuis août 2002, la bibliothèque a implémenté un système informatique de gestion des archives de PANDORA, appelé PANDA.

---

[39] Our Digital Island
http://www.statelibrary.tas.gov.au/odi/
Dernière consultation le 05/11/02

[40] **PANDORA** Liste de tous les titres disponibles
http://pandora.nla.gov.au/alpha/ALL
Dernière consultation le 05/11/02

3. Une acquisition par snapshot : Les bibliothèques d'Europe septentrionale et le cas particulier de la Suède

Le NWA (Nordic Web Archive) est la réunion de plusieurs projets d'archivage du Web pour les pays nordiques. L'objectif de ce groupe de bibliothèques est de développer des outils d'archivage communs et de rationaliser les initiatives locales dans un ensemble plus vaste. Le Danemark, la Norvège, l'Islande La Finlande et la Suède sont les membres de ce groupe de coordination de projets.

Le Kulturarw 3 Project est le projet d'archivage des sites Web de la Bibliothèque royale de Suède. Il s'appuie sur une collecte automatisée des site Web suédois dans l'esprit d'un dépôt légal des sites Web.

Le périmètre d'acquisition du projet concerne tous les sites Web du domaine national (.se), mais aussi tous les sites Web ayant des noms de domaine plus générique (.com ; .net) enregistrés à une adresse ou un numéro de téléphone suédois. Enfin tous les sites dont le domaine se termine en .nu. En complément, la bibliothèque conserve également des sites étrangers parlant de la Suède ainsi que des sites proposant des traductions de textes de la littérature suédoise. La collecte concerne les sites Web de tout type sans restriction. Par contre, les sites de newsgroups, les archives ftp ou encore les bases de données sont considérées comme non prioritaires dans la collecte. Le web invisible n'est pas receuilli puisque les limites de l'acquisition correspondent à celles du robot.

L'acquisition se fait grâce à un robot qui effectue un snapshot complet du Web dans le périmètre indiqué ci-dessus. La bibliothèque effectue environ deux snapshots par an depuis 1997, date du premier snapshot.

Le catalogage des sites Web au format MARC n'est pas la priorité de la bibliothèque, préférant une mise en valeur du fonds basée sur les fonctionnalités actuelles du Web (navigation hypertextuelle, recherche plein-texte…).

La description des sites s'effectue sous la forme de métadonnées Dublin Core en XML.

L'identificateur des sites Web est le numéro URN.

Le dernier snapshot a permis l'acquisition de 31 millions de fichiers.

4. Le cas de la Library of Congress et Internet Archive

« Internet Archive » est une association à but non lucratif fondée en 1996. Son objectif est de créer la bibliothèque patrimoniale de l'Internet. Aujourd'hui « Internet Archive » est en grande partie financé par Alexa une entreprise américaine aujourd'hui rachetée par Amazon.com et spécialisée dans l'archivage numérique.

Internet Archive est associée à de nombreux établissements publics américains, dont la bibliothèque de Congrès. C'est ainsi qu'en 1998, Alexa a remis à la bibliothèque un snapshot effectué au début de l'année 1997.
L'objectif d'Alexa et d'Internet Archive est de conserver l'ensemble du Web mondial. La wayback machine, développée par Alexa, sert d'interface à tout internaute désirant consulter les archives du Web sur Internet Archive. L'archive comprend aujourd'hui 100 Térabytes. Une douzaine de snapshots a été effectuée depuis sa mise en place en 1996. La fréquence de l'actualisation des enregistrements de sites est de six mois environ.
La collecte des sites Web s'effectue donc sans aucune sélection même d'ordre territorial. Par contre, les archives ne comprennent pas le contenu des sites appartenant au Web invisible.
Dans le cadre du projet Minerva et en relation avec Alexa, la bibliothèque du Congrès a engagé un vaste projet de collecte des sites Web américains liés aux élections de l'année 2000. Dans le cadre de ce même projet, la bibliothèque du Congrès a créé un fonds d'archives numériques sur les événements du 11 Septembre 2001. Ces deux collections sont d'ailleurs hébergées sur le site Web d'Internet Archive.

5. Quelques cas de consoritums

De façon à faire face aux difficultés de l'archivage électronique, plusieurs grandes bibliothèques se sont associées dans de vastes structures. Ces groupements permettent un partage d'expérience des plus intéressants. Par ailleurs, ils constituent parfois un moyen de réduire certains coûts, notamment de développement.

5.1 CEDARS

CEDARS (CURL exemplars in digital archives) est un projet qui a débuté en 1998 et qui a pris fin en mars 2002. Il ne s'agit pas à proprement parlé d'un projet international puisqu'il est centré sur les îles britanniques. Il regroupe à la fois le consortium CURL qui rassemble plusieurs grandes universités anglaises, écossaises et irlandaises, et d'autres grandes universités comme Leeds, Cambridge et Oxford. L'objectif de CEDARS était d'apporter une réflexion et une expertise sur l'archivage électronique en général. L'un des sous-groupes de CEDARS, CAMILEON (Creative Archiving at Michigan and Leeds emulating the old on the new) a conduit un projet autour de l'émulation. Il s'agit d'un projet anglo-américain unissant les universités de Leeds et du Michigan.

5.2 NEDLIB

NEDLIB (Networked European deposit Library) est un projet européen regroupant plusieurs grandes bibliothèques nationales (BnF, Norvège, Finlande, Allemagne, Pays-Bas,

Portugal, Suisse et Italie). Le chef de projet est la Koniklijke Bibliotheek (Pays-Bas). Ce groupement a été formé en 1998 sous l'égide de l'European Commission's Telematics Application Programm.
L'objecti du projet est de concevoir des outils organisationnels et informatiques de façon à mettre en place un dépôt légal des publications électroniques de toute sorte. NEDLIB a notamment mis en place un système organisationnel appelé DSEP fondé sur l'OAIS.

5.3 RLG et OCLC

En mars 2000, RLG (Research Libraries Group) entament une collaboration avec OCLC (Online Computer Library Center) sur l'archivage des ressources électroniques. L'objectif est de permettre à plusieurs grandes bibliothèques universitaires et nationales de partager leurs réflexions sur ces questions d'archivage. Il s'agit donc de mettre en place une recherche collaborative. RLG édite la revue Diginews qui est une revue électronique traitant exclusivement des questions liées à l'archivage de ressources électroniques.
L'objectif sous-tendu par cette collaboration est d'atteindre un consensus sur l'archivage et d'harmoniser les pratiques. C'est ainsi que de nombreux travaux de cette association s'appuient sur l'OAIS.

Annexe 2 : Outils d'évaluation de solution possibles pour l'archivage de sites Web

Cette annexe doit compléter la deuxième partie de ce mémoire en présentant deux exemples de métadonnées et en dressant un tableau synthétique reprenant tous les avantages et les inconvénients de chaque solution présentée en deuxième partie.

2.1 Des exemples de métadonnées minimales

1. Les métadonnées présentées par le projet NEDLIB

1.1 Métadonnées structurelles

- **Equipement informatique spécifique**

  = Description de configurations et d'équipements non standards

  Obligatoire que dans le cas où un une configuration ou un équipement non standards sont nécessaires pour lire le fichier. Non répétable

- **Microprocesseur spécifique**

  = Instructions particulières pour l'installation d'un microprocesseur ou d'un coprocesseur

  Obligatoire seulement dans le cas où le document ne peut être consulté que par le biais d'une application nécessitant des instructions d'installation particulières. Répétable

- **Equipement multimédia particulier**

  = Description d'équipements multimedia non standards

  Obligatoire seulement dans le cas où la consultation du fichier nécessiterait une application qui utilise obligatoirement un périphérique particulier. Répétable

- **Equipement périphérique particulier**

  = description d'un matériel périphérique non standard.

  Obligatoire seulement dans le cas où un périphérique particulier s'avère indispensable pour la consultation d'un fichier

  Sous éléments :

  - **Nom**
  - **Version**

- **Système d'exploitation**

  = Description du système d'exploitation à partir duquel une application peut fonctionner en vue de consulter le fichier.

  Sous-élements :

  - **Nom**
  - **Version**

- **Interpréteur et compilateur**

= Description du programme nécessaire en vue d'exécuter un autre programme écrit dans un langage complexe (compilé ou interprété).

Obligatoire seulement si nécessaire

Sous-éléments :

- **Nom**
- **Version**
- **Instructions**

- **Format objet**

  = description du format du fichier

  Obligatoire et répétable dans les cas de format complexes incluant, par exemple une DTD

  Sous éléments :

  - **Nom**
  - **Version**

- **Application**

  = Nom et version de l'application permettent de consulter le fichier

  Obligatoire

  Sous-éléments :

  - **Nom**
  - **Version**

1.2 Métadonnées descriptives

- **Référence**

  = Informations permettant d'identifier le contenu du fichier

  Sous-éléments :

  - **Créateur**
  - **Titre**
  - **Date de création**
  - **Editeur**
  - **Identifiant**
  - **URL**

- **Identifiant**

  = numéro d'identification unique et persistant du site Web

  Sous-éléments :

  - **Numéro**
  - **Méthode de construction de l'identifiant**

- **Agence responsable de l'identification**

- **URL**

  = adresse URL du site Web

  Champ optionnel

  Sous éléments :

  - **URL**
  - **Date de validation de l'URL**

- **Authentification**

  = Données permettant d'authentifier le document

  Champ optionnel

  Sous éléments :

  - **Checksum** (le cheksum est inclus dans le protocole d'échange TCP-IP et permet de contrôler de contrôler la taille des fichiers transmis)
  - **La signature digitale**

- **Checksum**

  = Informations sur l'utilisation d'un checksum

  Champ optionnel

  Sous éléments :

  - **Checksum**
  - **Algorithme**

- **Changements subis par le site Web archivé**

  = décrit les changements qu'a subis le site Web archivé

  Champ obligatoire et répétable

  Sous éléments :

  - **métadonnée principale concernée**
    - **Date de la transformation**
    - **Ancienne valeur (avant transformation)**
    - **Nouvelle valeur (après transformation)**
    - **L'outil de conversion**
    - **Eversibilité**
  - **Outil de conversion**
    - **Nom**
    - **Version**
  - **Réversibilité**

  - **Autres métadonnées concernées**

= toutes les métadonnées qui lors de ce changement ont subi des transformations alors que ces changements n'étaient pas l'objet principal visé de cette transformation du fichier

2. Les métadonnées DUBLIN CORE

- Titre signifiant
- Créateur
- Sujets et mots-clés
- Description du contenu
- Editeur (entité responsable de la diffusion de la ressource)
- Contributeur (entité ayant contribué à la création du contenu de la ressource)
- Type (nature ou genre du contenu de la ressource
- Format (matérialisation physique ou digitale de la ressource)
- Identifiant (Une référence non ambiguë à la ressource)
- Source (référence, s'il y a lieu, à une autre ressource à partir de laquelle la présente aurait été dérivée)
- Langue
- Relation (Référence à une autre ressource ayant un rapport avec la présente)
- Couverture (Portée dans le temps et dans l'espace de la ressource)
- Droits (informations sur les droits liés à l'utilisation de la ressource)

Annexe 2.2 : Tableau d'évaluation récapitulatif des solutions envisagées pour l'archivage des sites Web

Cette partie a pour objectif de faire une synthèse des solutions envisagées pour l'archivage de sites Web dans la deuxième partie de ce mémoire.

Les modalités d'acquisition des sites Web

Nom de la méthode	Type de sélection	Niveau d'exhaustivité	Coût en matériel	Coût en personnel	Avantages particuliers
Sélection manuelle	Manuelle	Relativement faible	Peu important	Très important	-Permet d'aboutir à une politique d'acquisition cohérente par rapport à une politique globale - Aboutit à une collection à ciblée dont le nombre est gérable
Snapshot	Sélection automatique	Assez élevé (mais il n'est pas possible d'opérer un suivi de nombreuses versions des sites archivés)	Très important	Très important au début du projet (conception du robot)	-Permet d'aboutir à une collection représentative du Web. - conserve en partie les propriétés de navigabilité du

					Web conservé.
Le dépôt de sites	Sélection manuelle	Relativement faible	Faible	Très important	- Système intéressant pour l'acquisition de bases de données et du Web invisible.

Les modes de conservation

Mode de conservation	**Cette solution nécessite une transformation des données conservées ?**	**Permet-elle de conserver tous les aspects du site Web conservé (contenu, forme, fonctionnalités) ?**	**Quel est le niveau de faisabilité de cette solution ?**
Conservation analogique (sur papier ou sur microformes)	OUI	NON	Important
Le musée technique	NON	Difficile de le savoir	Faible et risqué
La migration des données	OUI	Difficile de le savoir	Important
L'émulation	NON	OUI	Très faible à l'heure actuelle

Les métadonnées

Nom du système de référencement	**Toutes les métadonnées nécessaires sont-elles présentes dans le système ?**	**Certaines métadonnées peuvent-elles être générées automatiquement ?**	**Le personnel de bibliothèque peut-il se former rapidement à ce système ?**
Unimarc	NON	NON	OUI
Dublin Core	NON	OUI	OUI
XML	OUI	OUI	NON (formation plus longue)

Annexe 3 : Essais de propositions d'archivage pour la Bibliothèque municipale de Lyon

Dans cette annexe, vous allez trouver un ensemble de propositions regroupées en trois scénarios distincts faits pour la documentation régionale de la Bibliothèque municipale de Lyon. Ces propositions ont été conçues à partir d'une commande de la bibliothèque en vue de l'archivage de sites Web à intérêt régional. Notons qu'il ne faut pas considérer ces scénarios comme achevés, mais plutôt comme des pistes de réflexion à affiner pour un projet concret.

Par ailleurs, ces propositions ont été conçues dans un cadre particulier et pour un établissement particulier. A ce titre, même si certaines propositions peuvent être reprises ailleurs, on ne saurait les appliquer dans d'autres bibliothèques sans une étude préalable fondamentale pour définir, notamment, les objectifs et le contexte technique et humain d'un tel projet.

Ce travail a été effectué sur une période très limitée (Trois mois) correspondant à un stage. De ce fait, il est certainement très incomplet et nécessiterait sans aucun doute des investigations plus longues.

Cette annexe s'articule donc autour de quatre grandes parties : la première permet de resituer le contexte dans lequel l'étude a été demandée et définit les objectifs et les contraintes de la bibliothèque pour un projet d'archivage. Quant aux trois dernières, chacune correspond en fait à un scénario particulier.

Annexe 3-1 : Contexte et objectifs : Vers un archivage des sites Web d'intérêt régional

Le département de la documentation régionale de la Bibliothèque municipale de Lyon a mis en place depuis l'année 2002, un annuaire rassemblant des sites Web traitant de la région Rhône-Alpes et des huit départements qui la composent. L'idée de commander une étude sur l'archivage de sites Web d'intérêt régional est venue de la réflexion conjointe d'Yvette Weber, responsable de la documentation régionale et du dépôt légal à la bibliothèque, de la direction générale de la bibliothèque ainsi que de Sylvie Pillet, stagiaire Enssib ayant participé à la conception de l'annuaire. Très rapidement, le caractère incomplet de l'annuaire au regard des missions du service de la documentation régionale est apparu à l'ensemble de ces protagonistes. En effet, ce service a pour double mission d'offrir au public des informations actuelles sur la région –ce qui correspond tout à fait aux fonctions de l'annuaire- mais aussi de protéger et conserver certains documents qu'elle propose. Or, bien entendu, l'annuaire ne permet pas au service d'accomplir, dans le domaine du Web, sa mission patrimoniale.

Il s'agissait donc d'imaginer ce que serait l'archivage de sites Web dans le contexte particulier de la Bibliothèque municipale de Lyon et de la région Rhône-Alpes.

1. L'intérêt de l'archivage

Nous avons déjà développé les avantages que pouvaient représenter l'archivage de sites Web d'intérêt régional. Sachant que l'Internet devient un moyen important de diffusion d'idées et de contenus informationnels aussi bien au sein de grandes institutions, qu'au sein de plus petites, au sein d'associations comme en celui de groupes plus informels, il serait dommage de ne pas au moins conserver quelques traces de ces sites Web. De nombreux usagers viennent à la Bibliothèque de Lyon pour ses collections de périodiques anciens. Certains quotidiens étaient le fait de grands groupes éditoriaux, mais également de petits groupes politiques, associatifs ou religieux. Ces mêmes groupes utilisent aujourd'hui Internet pour se faire connaître. Parfois, la création d'un site Web n'est qu'un moyen de varier les supports de communication : C'est le cas par exemple de certaines grandes municipalités qui éditent souvent des brochures ou des journaux sur papier. A l'inverse, le site Web est le seul média utilisé et alors, la perte du site Web du fait de sa non conservation sera une perte irremplaçable.

La région Rhône-Alpes est très dynamique sur le Web. Après l'Ile de France, elle est la région la plus riche en sites Web et en internautes. Cette richesse constitue un intérêt important pour la constitution d'un fonds puisque nous avons dores et déjà l'assurance que celui-ci sera riche. Mais en même temps, cette même richesse constituera une difficulté

importante pour l'archivage puisque très vite, la bibliothèque sera confrontée à une masse relativement importante de sites Web à traiter et archiver.
L'archivage représenterait également un intérêt important pour la bibliothèque de Lyon. Cet archivage continuerait le travail novateur opéré à la bibliothèque en matière de documents numériques. En effet, on ne saurait oublier qu'en 1995, la Bibliothèque de la Part-Dieu fut la première bibliothèque municipale française à proposer à ces usagers une connexion à Internet. Comme nous l'avons dit précédemment, l'archivage des sites Web permettrait en outre à la bibliothèque d'effectuer toutes ses missions premières au niveau du Web,et notamment ses missions patrimoniales.
Enfin, en l'état actuel, le projet de dépôt légal de sites Web développé par la BnF n'aboutirait pas, pour des questions de droit, à une diffusion sur le Net des sites Web archivés. Ne serait-il pas appréciable qu'un accès et une mise en valeur du Web local soit alors proposée en local, dans la région Rhône-Alpes ?

2. Les objectifs

Il s'agit ici à la fois de définir les objectifs de l'archivage de sites Web mais également d'envisager une définition du périmètre du Web concerné par ce projet.
Le but n'est pas de concevoir une politique d'acquisition complètement nouvelle pour les sites Web. Même si l'Internet comporte des spécificités qu'il faudra prendre en considération, l'objectif est *in fine* de reproduire pour les sites Web la politique d'acquisition développée par le service de la documentation régionale de la bibliothèque.
Dans le cadre de la bibliothèque de Lyon, l'intérêt régional recouvre l'ensemble des documents et témoignages qui concernent la région Rhône-Alpes et ses huit départements. L'exhaustivité, surtout pour les imprimés, est l'objectif visé par le service. A ce titre, tous les documents, pour peu qu'ils concernent la région ou l'un des ses huit départements, peut faire l'objet d'une acquisition. La situation de la bibliothèque, pôle associé de la BnF au titre du dépôt légal, participe de ce souci d'exhaustivité. A l'intérieur de ce principe de territorialité, tus les thèmes sont représentés de l'histoire à l'économie, en passant par l'urbanisme ou la sociologie. Tous les niveaux sont également représentés, de l'ouvrage grand public aux monographies et périodiques scientifiques.
Au niveau de l'archivage des sites Web, il s'agirait donc de conserver tous les sites Web qui traitent de la région ou des départements rhônalpins. Du fait de la variabilité des sites Web, il sera impossible de viser à l'exhaustivité puisque la bibliothèque ne pourra, du moins en l'état actuel, enregistrer toutes les versions de chaque site Web concerné. Du point de vue des mentions de responsabilités, le service acquière certes des documents imprimés dans la région, mais également d'autres émanant d'institutions nationales. C'est

le cas par exemple des rapports de l'INSEE qui, lorsqu'ils concernent la région, sont acquis. De la même façon pour les sites Web, la localisation du site sur un serveur local ne devrait pas être une condition d'acquisition. De la même façon, si un grand site d'envergure nationale consacre quelques pages à la région, ces quelques pages devraient faire l'objet d'une acquisition et d'une conservation. Ce fait montre que le niveau de granularité envisagé pour l'archivage pourra être relativement fin puisque ne pouvant concerné non pas le site lui-même mais l'une de ses parties.

Pour l'usager, il s'agirait de donner accès sur le long terme à toutes les composantes des sites Web : tous les contenus, la forme, les fonctionnalités. Nous verrons par la suite comment les contraintes de conservation conduiront la bibliothèque à pondérer chacun de ces aspects, mais du moins pour le moment, on peut dire que l'objectif est d'offrir à l'usager un fonds d'archive de grande qualité ce qui, nous l'avons vu, entraîne de grandes difficultés pour la conservation.

Le principal critère d'acquisition sera donc un critère territorial. La question de l'objet à archiver demeure alors. A priori, les forums de discussion ne feront pas l'objet d'un archivage. Tout d'abord parce que ces forums renferment souvent des données personnelles et que leur intérêt régional est souvent relativement faible. Par ailleurs, le suivi de ce type de document, du fait de leur haut niveau de variabilité serait des plus complexes. Les annuaires de sites Web locaux (comme Alpavista par exemple) ne devraient pas faire l'objet non plus d'un archivage. Ces annuaires sont davantage des outils que des sites Web et leur contenu informationnel a une durée de vie relativement limitée.

D'un point de vue général, voici, dans ses grandes lignes à quoi pourrait ressembler une ébauche de définition du périmètre du Web concerné par l'acquisition. La première partie du tableau présente succinctement les grands principes présidant à l'acquisition des sites Web à partir de l'analyse de leur contenu. La seconde partie indique des décisions possibles d'acquisition par rapport au type de site Web analysé.

Principe général d'acquisition	**Caractéristique**
Langue du site	Français
Thèmes traités	Tous les thèmes
Niveau	Indifférent
Localisation du site	Indifférente
Niveau de granularité	Variable
Critère d'acquisition	Traiter de la région Rhône-Alpes ou de l'un de ses départements ou zone géographique

	(ville, village, zone rurale, zone de montagne…)		
Type de site	**Acquisition**	**Pas d'acquisition**	**Réserves**
Forums de discussion		✓	
Sites Web d'accès payant		✓	A analyser au cas par cas
Revues électroniques	✓		Sauf si abonnement payant
Annuaires régionaux		✓	
Moteurs de recherche régionaux		✓	
Sites Web personnels		✓	Sauf s'ils correspondent au critère d'acquisition (c'est le cas, par exemple de certains sites de petites municipalités de la région)
Sites commerciaux (.com)	✓		Sauf s'il s'agit de sites publicitaires. Pour être acquis ce type de sites doit comporter des informations sur l'entreprise dont parle le site et pas seulement sur les produits qu'elle propose.
Sites Web traitant de l'Internet (.net)	✓		A la condition qu'il corresponde bien au critère d'acquisition
Bases de données	✓		A la condition qu'elle corresponde bien au critère d'acquisition.

Ce tableau est bien sûr loin d'être complet, mais il peut servir de base de réflexion, en l'attente d'une analyse plus poussée du Web local.

3. Les données à prendre en considération

Les contraintes d'ordre budgétaire et de personnel sont à prendre en considération pour un tel projet. Toutefois, nous ne nous situons pas encore à ce niveau d'élaboration d'un projet. Il s'agit plutôt de fournir à la bibliothèque de Lyon des pistes possibles pour la mise en place d'un projet. Cela ne signifie pas que les données budgétaires ou de personnel n'ont pas été prises en considération pour l'élaboration des trois scénarios possibles. Ainsi, l'idée d'un catalogage des sites en Unimarc a-t-elle été éliminée d'office compte-tenu du nombre important de personnes qu'elle nécessitait.

Pour élaborer ces scénarios nous avons également tenu compte de contraintes externes à l'établissement. C'est ainsi que l'un des trois scénarios envisage une collaboration avec la BnF. Tous les scénarios s'appuient par ailleurs sur la norme OAIS d'organisation des archives électroniques.

Par ailleurs, il nous est apparu important d'intégrer dans ce projet d'archivage plus que les seuls sites Web. En effet, la bibliothèque de la Part-Dieu s'est lancée, depuis plusieurs années déjà, dans une vaste politique de numérisation de fonds anciens et fragiles. Ces images numérisées sont à l'heure actuelle stockée sur des CD. Il nous a semblé que la conservation de ces CD devait être envisagée sous peine de devoir réitérer dans quelques années les mêmes opérations de numérisation, à la fois coûteuses et risquées pour les collections. Ces CD devraient donc entrer dans le dispositif de conservation conçu pour les sites Web. Ils seraient référencés succinctement, entrés dans une base de donnée pour les migrations, vérifiés et subiraient des migrations successives pour lutter contre leur vieillissement.

Enfin, nous avons tenu compte non seulement de l'existant mais encore des projets à long terme de l'établissement. La bibliothèque serait intéressée par l'acquisition d'un outil de recherche performant. Nous n'entrerons pas dans les détails puisqu'il s'agit d'un projet en cours, mais disons que ce projet d'acquisition a orienté notre choix d'un référencement basé sur le XML, langage justement utilisé par l'outil en question. Le choix du XML, dans le cas de l'acquisition de cet outil, permettrait d'intégrer les sites Web archivés dans l'ensemble plus vaste des collections de la bibliothèque.

Annexe 3-2 scénario 1 : Une approche sélective de l'archivage des sites Web

Présentation générale du scénario

Dans ce scénario, il s'agit d'appliquer un mode de sélection manuel des sites Web à archiver. Il ne serait pour autant pas raisonnable de mettre en place un service spécifique pour l'acquisition des sites Web. Il nous est donc apparu plus intéressant de s'appuyer sur l'existant, en l'occurrence, l'annuaire des sites Web rhônalpins de la documentation régionale. Cet annuaire a été mis en place en 2001 et rassemble plus de 600 sites Web. Ce premier scénario consisterait donc à enregistrer une à deux fois par an les sites Web sélectionnés dans l'annuaire. Notons cependant que certains sites présentés dans l'annuaire pourront ne pas être enregistrés : c'est le cas par exemple des sites ne présentant que des images filmées de la ville par une Webcam, ou encore de ceux présentant l'état de la circulation automobile, ou encore la météo. En effet, la conservation de ces sites n'aurait certainement du sens que si nous pouvions les enregistrer quotidiennement, ce qui ne serait pas possible.

Au niveau de la conservation, un double enregistrement serait effectué, un premier sur cassettes DLT et un second sur serveur. Le premier enregistrement fera office de copie de conservation. Les sites y seront stockés dans leur format d'origine. Il s'agira également de conserver de la même façon, les programmes permettant la lecture et la visualisation de ces formats (Word, Acrobat reader…). Dans le cas où des émulateurs seraient développés de façon satisfaisante par exemple, il serait alors possible d'utiliser ces copies sur DLT. Par contre, les cassettes DLT subiraient bien entendu des migrations de façon à rafraîchir périodiquement les supports. Parallèlement à cette copie sur cassettes, les sites Web seraient enregistrés et stockés sur un serveur dédié. Sur ce serveur, les sites Web subiraient les transformations successives selon le rythme d'obsolescence des formats dans lesquels ils ont été enregistrés. En outre, les sites Web subiraient dès leur copie sur le serveur une première migration puisqu'ils ne seraient pas enregistrés en HTML mais en XML de façon à faciliter leur traitement par un moteur de recherche XML dont nous reparlerons plus précisément. Par contre, chaque transformation des sites Web

sera conservée sur un cédérom, pour une durée variable à déterminer. Cette copie de chaque transformation devra permettre de pallier les difficultés que l'on pourrait rencontrer lors d'une migration d'un format à un autre. En cas de migration défectueuse, il serait alors possible d'utiliser une copie du site précédant la migration. En résumé, trois copies seront effectuées : la première sur DLT qui ne subira aucune transformation de formats et qui sera accompagnée des copies des programmes et utilitaires de visualisation (système d'exploitation, programmes…), la deuxième sur serveur subira des transformations de formats dont une première en XML. C'est ce deuxième enregistrement sur serveur que le public consultera grâce à un moteur de recherche. La troisième copie, contrairement aux deux autres, ne devra pas être conservée indéfiniment. Sur cette troisième copie, sur cédérom, seront stockées les versions précédant directement une migration de format et feront office de copie de secours en cas de migration défectueuse.
La gestion des migrations sera organisée comme une gestion des risques. Nous verrons plus tard de quelle façon elle prendra forme. Le référencement des sites Web contiendra à la fois des informations bibliographiques mais également des informations sur la structure informatique des fichiers, les formats, les transformations effectuées… Ces informations seront notamment conservées dans une base de données dont l'accès sera réservé aux personnels en charge des archives des sites Web. Cette base de donnée devra permettre de retrouver et identifier le plus rapidement possible les fichiers comportant le même format de façon à lister, dans le cas où un fichier serait susceptible de devenir obsolète, les fichiers en danger et de procéder à leur conversion en un nouveau format cible. En outre, cette base de donnée devra comporter des informations les plus complètes possibles sur les formats en général. A ce titre, il serait intéressant d'envisager une connexion directe avec la BnF ou avec tout consortium s'occupant d'échanger des informations sur les formats de façon à constituer une base de données actualisée sur les formats et logiciels informatiques.

Au niveau du référencement, les sites Web seront traités en XML selon la même DTD que celle utilisée à la BnF de façon à faciliter, le cas échéant des échanges et des récupérations de « notices ». Au niveau de la description bibliographique des sites Web, les notices entrées dans l'annuaire seront utilisées. Le numéro d'identification du site Web sera formé de l'adresse URL du site et de son numéro d'entrée dans l'annuaire (indiqué dans la notice).

L'accès aux archives se fera donc par le biais d'un serveur qui stockera les sites Web archivés en XML. La recherche de sites se fera par le biais d'un moteur de recherche. Toutes les versions d'un même site seront liées par un lien hypertexte et devront comporter le même identifiant. Idéalement, la recherche devrait pouvoir être effectuée de plusieurs façons :

- En entrant le nom ou l'adresse du site recherché ce qui permettrait de choisir entre plusieurs versions d'un même site. Ce mode de recherche s'apparenterait à celui développé par la Waybackmachine où chaque version d'un même site est classée par année d'enregistrement.
- Par une recherche plus classique par le biais du moteur de recherche
- En naviguant dans les archives. Dans ce cas, par exemple, vous pouvez choisir de naviguer dans les archives de 2005. Par contre, la navigation réelle ne sera effective que dans la mesure où les liens hypertextes présents sur les sites renverront à d'autres sites eux aussi archivés.

Modes opératoires généraux

1. Enregistrement et traitements

1. Paramétrer « l'aspirateur » de sites Web

Dans cette première phase, il s'agit tout simplement d'entrer les adresses des sites Web qui seront enregistrés et conservés. Ce que l'on appelle « aspirateur » est tout simplement un programme informatique permettant de récupérer les sites Web. Certains logiciels sont disponibles gratuitement. Notons cependant que le logiciel choisi devra être capable d'absorber un grand nombre de sites Web.

Dans l'idéal, il serait intéressant de coupler l'aspirateur de sites Web avec le logiciel de gestion de l'annuaire des sites ASIR PRO. Il serait ainsi possible dès que l'on entre un site Web dans l'annuaire et que l'on crée sa notice, d'entrer cet ensemble dans l'aspirateur. Ceci éviterait que l'on entre deux fois le même site Web, d'abord dans l'annuaire puis dans l'aspirateur de sites.

2. Aspiration des sites et enregistrement sur le serveur

Cette seconde opération effectuée une à deux fois par an consiste tout simplement à aspirer et enregistrer les sites dont on a entré la liste dans l'aspirateur.

Outre les sites Web, cette opération permet également de récupérer des informations sur les sites à partir du MIME. Ces informations concernent le poids des fichiers informatiques, les formats, les langages des fichiers et sont donc capitales.

3. Test de l'enregistrement

Cette opération consiste tout simplement à vérifier que tous les sites Web ont bien été enregistrés, qu'ils sont accessibles et complets. Cette phase de vérification sera effectuée automatiquement par un logiciel.

4. Récupération et transformation des métadonnées

A ce stade des opérations, il s'agit de transformer deux types de données en métadonnées : les éléments des notices entrés dans ASIR PRO et les éléments du MIME récupérés lors de l'opération d'aspiration. La BnF utilise notamment un logiciel permettant de transformer les informations du MIME en métadonnées XML. Par contre, concernant les éléments des notices, il s'agira de trouver voire éventuellement de développer le même type d'outil mais permettant de transformer des données écrites en MySQL en métadonnées XML. L'adjonction de ces deux éléments (données des notices et données du MIME) formera ainsi les métadonnées des sites Web. Ces métadonnées devront être encapsulées dans chaque site Web correspondant.

5. Enregistrement des archives sous DLT

Les sites Web et leurs métadonnées encapsulées sont alors enregistrés sous cassettes DLT.

6. Test et traitement des cassettes DLT

Chaque cassette DLT est alors testée là encore pour vérifier que tous les sites et leurs métadonnées y sont présents et accessibles. Chaque cassette est alors estampillée. Une étiquette lui est adjointe correspondant à un numéro d'identification (numéro d'inventaire), la date d'enregistrement, le poids en octets des informations contenues, la nature des informations (sites Web,

programmes…), une indication de sa localisation. Les cassettes sont alors rangées dans le silo.
Ces informations sont également enregistrées dans la base de données de gestion des migrations : le numéro d'inventaire de chaque cassette, la date d'enregistrement, le poids en octets, le type d'informations, la localisation mais également la liste des sites Web ou des programmes contenus.

7. Sur le serveur, transformation des sites archivés en XML

Au niveau du serveur les sites Web sont alors convertis en XML. Les logiciels capables de transformer du HTML en XML existent mais il faut préciser qu'ils sont, pour le moment difficiles à utiliser.

8. Test de l'archive sur serveur

Il s'agit alors de vérifier que la transformation en XML s'est effectuée dans de bonnes conditions, qu'aucun site ne manque, que chaque site comporte bien ses métadonnées…

9. Etablir les liens entre les différentes versions d'un même site

Il s'agit alors d'établir des liens hypertextes entre les versions successives d'un même site Web. Cette opération devra vraisemblablement être effectuée manuellement.

2. Gestion des migrations

Deux cas de figure peuvent se présenter :

1. Dans le cas d'un simple rafraîchissement de supports

Ce cas de figure concerne les cassettes DLT, mais aussi les cédéroms que la bibliothèque produit dans le cadre de sa politique de numérisation.
Avant tout, il s'agit de préciser qu'à l'instar des cassettes DLT, il serait important que des informations sur les cédéroms produits par la bibliothèque soient entrés dans la base de donnée de gestion des migrations. Chaque cédérom aurait donc un numéro d'identification, une localisation, la date d'enregistrement, le type d'informations et la liste de ces informations.
Le responsable de la base de données pourrait alors savoir quels cédéroms ont plus de cinq ans et engager un processus de rafraîchissement de supports, ou

quelles cassettes ont plus de dix ans… Il s'agirait alors de fixer une période de validité des supports en fonction de leur condition d'archivage. Passée cette période de validité, le contenu des cassettes ou cédéroms concernés serait alors enregistré sur de nouveaux supports. Une fois la migration faite, un test devra être effectué. Si la migration a réussi il faudra alors modifier la date d'enregistrement au niveau de la base de données et se débarrasser de l'ancien support.

2. Dans le cas d'une migration plus complexe avec conversion de fichiers en un nouveau format

Ce type de migration doit permettre de pallier l'obsolescence des formats et l'impossibilité d'accéder aux fichiers concernés. La solution consiste alors à convertir le fichier en danger en un nouveau format, ou dans un nouveau langage ou code. Cette opération est risquée pour le fichier et il s'agit donc de gérer ce risque.

1. En premier lieu il faut être certain que le format « f », le code « c » ou le langage « l » sont touchés par un risque important d'obsolescence. Si par exemple la nouvelle version d'un logiciel est compatible avec sa version précédente, on ne saurait considérer que le risque d'obsolescence est important. A ce stade, on se rend aisément compte de l'importance pour la personne en charge des archives d'obtenir un excellent niveau d'information. C'est à ce titre qu'un échange d'informations sur les formats et langages apparaît fondamental sinon au niveau international du moins au niveau français.

2. Une fois que le risque d'obsolescence d'un format est attesté, il s'agit d'évaluer le nombre de fichiers concernés.

3. Les deux opérations précédentes permettent d'évaluer le risque général encouru par le fonds. Il s'agit alors d'effectuer sur CD une copie de sauvegarde de tous les fichiers concernés par ce risque d'obsolescence.

4. Il s'agit ensuite de choisir un nouveau format ou un nouveau langage cibles dans lesquels les fichiers seront convertis. Le choix du format cible doit s'appuyer sur des informations fiables sur celui-ci et sur son niveau de compatibilité avec le format précédent. A ce niveau, l'expérience d'autres établissements peut être une donnée riche en enseignements. Le format cible doit répondre à plusieurs critères et sa sélection doit pouvoir prendre la forme d'une grille. Par exemple, on peut imaginer la grille suivante :

Grille de sélection d'un nouveau format

Informations générales

Nom du format :

Description du format :

Fonctions :

Langages :

Equipement requis :

L'utilisation de ce format est-elle compatible avec l'équipement de la bibliothèque ?

OUI NON (choix à éliminer alors)

Informations sur le type de format

Ce format est :

Propriétaire fermé **- 2**

Propriétaire ouvert **- 1**

Non propriétaire **+ 1**

Est une norme **+ 2**

Niveau de compatibilité du format

La conversion de fichiers en format « F » en ce format cible entraîne des pertes :

Au niveau des fonctionnalités **- 2**

(navigabilité, mode de recherches… à détailler)

Au niveau de l'apparence formelle **- 3**

(taille des images ou des caractères, qualité des images, qualité du son, altération des animations, altération des couleurs, élimination des images…)

Au niveau de la compréhension et du contenu informationnel du fichier **- 4**

(perte d'une partie de la ponctuation, perte des caractères accentués, perte de phrases…)

Note globale :

Décision finale :

Format retenu format non retenu

Cette grille est bien entendu très basique. Dans le cas qui nous occupe, il faudrait envisager une grille plus détaillée. Ainsi pour le niveau de compatibilité, il faudrait pouvoir attribuer une note plus précise déclinée en fonction du type de document touché par la migration. Il est évident, par exemple, que dans le cas d'images une altération de caractéristiques formelles (couleur, taille…) peut être plus problématique que dans le cas d'un texte simple. Inversement, pour une base de données, la perte de fonctionnalités de recherche et de contenu informationnel serait plus dommageable que pour une image.

Il est donc vraisemblable qu'une grille de sélection basique sera élaborée mais qu'elle ne servira que de modèle général, des grilles de sélection plus précises devant être élaborées au cas par cas. L'important toutefois est que la grille de sélection permette d'effectuer un choix raisonné entre plusieurs formats cibles en attribuant une note ou un pourcentage de compatibilité.

Il est possible également d'envisager de tester la compatibilité des formats cibles avec un échantillon de quelques fichiers

5. Une fois identifié le format le plus adapté, il s'agit alors de convertir l'ensemble des fichiers concernés.
6. Les fichiers sont alors testés de façon à vérifier que la migration a bien été effectué et n'a pas entraîné de pertes supplémentaires. Si la migration a échoué, il est possible de procéder à une nouvelle opération de migration en se servant des copies de secours (phase 3)

Equipements et moyens nécessaires

L'équipement informatique requis consiste :

- En un serveur stable (Unix) avec d'importantes capacités de stockage
- Au moins un graveur de CD
- Un lecteur/enregistreur de cassettes DLT
- Et bien sûr une connexion haut débit à Internet (câbles)

Les logiciels nécessaires sont :

- Un logiciel d'aspiration de sites Web
- Un logiciel de conversion du HTML en XML
- Un logiciel de conversion du MIME en métadonnées XML
- Un logiciel de conversion de MySQL en métadonnées XML
- Une base de données (MySQL)
- Un moteur de recherche XML

Au niveau du matériel, certains de ces équipements ont déjà été acquis par la bibliothèque. Du point de vue des logiciels, certains sont disponibles gratuitement, c'est le cas par exemple de certains « aspirateurs » de sites Web mais également du logiciel de conversion en XML. Quant au moteur de recherche XML, l'achat d'un métamoteur XML pourrait faire office de moteur.

Du point de vue du personnel, le recrutement d'un informaticien s'avérerait sans aucun doute nécessaire aussi bien pour la mise en place de l'architecture globale du système informatique que pour la maintenance de celui-ci ou les phases d'aspiration des sites Web. Par ailleurs, étant donné le niveau de compétence requis en informatique pour la tenue de la base de données de gestion des migrations, le conservateur, responsable de cette collection de sites Web archivé, devra être aidé d'un informaticien.

Avantages et inconvénients

Le grand avantage de ce scénario est qu'il s'appuie sur un outil et des pratiques d'acquisitions déjà existants, puisqu'il se fonde sur l'annuaire des sites Web régionaux. De ce fait, le niveau de faisabilité de celui-ci est relativement important.

Toutefois, ce scénario comporte également de nombreux inconvénients. Paradoxalement, le premier inconvénient vient justement du fait qu'il s'appuie sur l'annuaire. En effet, l'annuaire n'a pas été conçu dans l'optique d'un archivage. La mise en place de passerelles entre le logiciel ASIR PRO et le système d'archivage risque d'être problématique. D'autre part, l'annuaire n'est pas pour le moment complet. De nombreux sites Web régionaux manquent et

celui-ci comporte essentiellement des sites Web d'institutions régionales ou nationales. A ce titre, même si nous ne l'avons pas mentionné dans les moyens nécessaires, le recrutement d'un assistant supplémentaire notamment pour l'acquisition de sites Web serait un avantage précieux pour cette entreprise. L'annuaire, par ailleurs, ne comporte pas certains sites au contenu problématique (les sites de sectes ou de groupes extrémistes). Il est tout à faite normal que ces sites ne figurent pas dans un annuaire mais il est dommage qu'ils ne figurent pas dans les archives. Il serait donc intéressant que ces sites soient conservés tout en limitant leur consultation ce qui est tout à fait faisable à partir des métadonnées (limitation d'accès) ou en les conservant à part sur cédéroms. Enfin, cette collection de sites Web sera incomplète à plus d'un titre, tout d'abord parce que l'annuaire ne l'est pas, mais également parce que les périodes d'enregistrement seront limitées (au maximum deux enregistrements par an). Il apparaît en effet peu raisonnable de solliciter davantage le réseau en effectuant plus de deux enregistrements par an.

Annexe 3-3 Scénario 2 : Une automatisation de l'acquisition

Présentation générale du scénario

Il s'agirait ici de pallier l'incomplétude des collections dans le scénario précédent, par une sélection automatique et systématique de sites Web. En d'autres termes, la bibliothèque effectuerait une fois par an un snapshot du Web régional. Les critères de sélection manuels seraient donc remplacés par des algorithmes de pertinence.
Au niveau de la conservation et des métadonnées, nous retrouverions les mêmes systèmes que dans le scénario précédent avec une base de données de gestion des migrations, trois types de copies et des métadonnées en XML. Par contre, contrairement au précédent scénario, les métadonnées de description n'utiliseraient pas les données des notices de l'annuaire, mais les balises HTML du site Web lui-même.
Quant à l'identifiant du site, il sera formé par l'adresse URL et le titre du site Web.

Modes opératoires

1. Le paramétrage du robot
   La bibliothèque devrait donc se doter d'un moteur capable de rechercher les sites pertinents à archiver et de les aspirer sous la forme d'un snapshot. Se pose alors le problème du paramétrage des algorithmes de pertinence de façon à circonscrire la partie du Web présentant un intérêt pour la région Rhônalpine.

2. La prise du snapshot
   Une fois circonscrite la partie du Web jugée pertinente, un snapshot sera effectué. Il permettra de récupérer non seulement les sites Web, mais encore des données liées au MIME.

3. Test du snapshot
   Il s'agit alors de vérifier que l'enregistrement a bien été effectué, que tous les sites sélectionnés sont présents. Cette vérification permettra en outre d'évaluer la part des sites qui n'a pu faire l'objet d'un enregistrement (Web invisible). Cette partie invisible du Web peut faire l'objet d'un dépôt mais il

faut savoir le nombre de personnes nécessaire pour ce type d'opération serait très important puisque le traitement de chaque site concerné nécessite au moins une journée de travail pour une personne qualifiée en bibliothéconomie et en informatique. De ce fait, nous ne traiterons pas cette possibilité d'acquisition de sites qui serait certainement trop lourde.

4. Production des métadonnées

Par le biais d'un logiciel, les informations contenues dans le MIME et celles contenues dans le code source HTML du site sont transformées en métadonnées XML et encapsulées dans chaque site Web.

5. Enregistrement sur DLT

Le snapshot est alors enregistré sur cassettes DLT

6. Test et traitement des cassettes DLT

Les cassettes sont alors testées et traitées (voir phase 6 du mode opératoire du scénario n°1)

7. Au niveau du serveur, les sites Web sont alors convertis en XML

8. Test des archives au niveau du serveur

9. Etablissement de liens entre les versions successives d'un même site

On considère alors qu'un site Web est la version suivante d'un autre, lorsque son titre et éventuellement son URL sont semblables.

Nota Bene : Le mode opératoire pour les migrations est exactement identique à celui développé dans le premier scénario

Equipements et moyens nécessaires

Du point de vue du matériel informatique, il faut envisager :

- Un serveur Unix
- Au moins un graveur de cédéroms
- Un lecteur/enregistreur de cassettes DLT
- Une connexion extrêmement puissante à Internet
- Un robot collecteur pour effectuer les snapshots

Les logiciels nécessaires sont :

- Un logiciel de conversion du HTML en XML
- Un logiciel de conversion du MIME en métadonnées XML
- Une base de données (MySQL)
- Un moteur de recherche XML

Le personnel nécessaire pour ce scénario est quasiment incalculable puisque le développement et le paramétrage du robot nécessiteront une importante équipe d'ingénieurs en informatique. La collaboration avec une équipe de recherche en informatique est envisageable.

Avantages et inconvénients

Ce scénario permettrait une acquisition extensive du Web régional et la constitution d'une collection véritablement représentative du Web régional.

Cependant, les inconvénients de ce scénario sont si important qu'il apparaît difficilement réalisable. En effet, le paramétrage d'un robot et la circonscription d'une portion du Web d'intérêt régional seraient particulièrement difficiles. En effet, lorsque l'on souhaite repérer les sites Web français, nous disposons au moins d'un nom de domaine particulier marquant un ancrage territorial au niveau français (.fr). Or il n'existe aucun nom de domaine pour les régions françaises. Le paramétrage d'un robot pour la région Rhône-Alpes nécessiterait donc d'envisager des algorithmes de pertinence plus pointus et plus complexes. D'autre part, une telle entreprise solliciterait considérablement le réseau informatique de la bibliothèque. Le coût d'une telle opération serait quasiment prohibitif. Enfin, si ce scénario parvient à pallier les manques du scénario précédent, il n'empêche que nous demeurons ici au niveau d'une acquisition extensive des sites Web. Le suivi plus fin de quelques sites est impossible à effectuer. Il sera impossible de conserver les sites Web apparus entre deux snapshots ainsi que les versions successives d'un même site.

Annexe 3-4 Scénario 3 : Une collaboration avec la BnF

Présentation générale

Ce scénario se fonde sur une collaboration étroite entre la BnF et la Bibliothèque municipale de Lyon. Dans le cadre de cette collaboration, la Bibliothèque de Lyon recevrait régulièrement de la BnF une portion du snapshot national correspondant au Web rhônalpin. En contre-partie, la bibliothèque de Lyon devrait indiquer à la BnF les sites Web de la région qui auraient échappé au snapshot et notamment les sites et les pages Web apparaissant ponctuellement dans la région à l'occasion de grands événements culturels ou politiques.

La bibliothèque de Lyon s'occuperait donc de la conservation et de la mise en valeur de cette portion du snapshot national. Parallèlement à cette tâche, il serait alors tout à fait envisageable que la bibliothèque effectue un suivi plus fin de certains sites Web particulièrement intéressants. Il s'agirait alors d'enregistrer plus fréquemment que lors des snapshots certains sites Web. Ces sites seraient versés au snapshot rhônalpin mais également transmis à la BnF de façon à compléter ses collections.

Concernant la conservation, nous retrouverions les trois copies déjà présentées dans le premier scénario. En outre, les sites Web suivis par la Bibliothèque de Lyon seraient aussi enregistrés sur cédéroms avant d'être versés dans le serveur. En dehors du cas particulier de ces sites Web, ce scénario impliquerait la même architecture que dans le scénario n° 1 avec une base de données de gestion des migrations, la gestion des risques pour les migrations successives etc.

En ce qui concerne les métadonnées, la bibliothèque municipale récupérerait les métadonnées incluses dans le snapshot de la BnF. Les métadonnées des sites Web suivis par la Bibliothèque de Lyon seraient elles aussi copiées à partir de celles du snapshot (seule la date d'enregistrement du site serait alors modifiée).

Modes opératoires

Dans ce scénario, trois types de modes opératoires peuvent être distingués :

Le traitement du snapshot

Ce mode opératoire concerne le traitement et la récupération de la copie de la portion rhônalpine du Web national.

1. La définition de la zone concernée
   A ce stade, il serait nécessaire de définir avec la BnF la partie du Web dont la Bibliothèque de Lyon souhaiterait obtenir une copie. Les modalités de récupération seront également définies (Cassettes DLT ou par connexion sécurisée…).

2. La récupération de la portion de snapshot
   La BnF convertit les sites Web en XML. De ce fait, la Bibliothèque de Lyon n'aurait pas à se charger de cette tâche. Il s'agira alors pour la bibliothèque de tester la livraison métadonnées incluses.

3. Enregistrement sous cassettes DLT
   Une fois testé, le snapshot sera alors enregistré sous cassettes DLT. Celles-ci feront office de copies de secours et seront traitées, comme vu précédemment dans le premier scénario et testées.

4. Enregistrement sur le serveur
   Le snapshot sera alors enregistré sur le serveur et testé.

5. Gestion des migrations
   Les migrations seront gérées comme nous l'avons vu pour le premier scénario

Mise en place d'une veille pour la BnF

Pour ce mode opératoire, la Bibliothèque de Lyon indiquera à la BnF certains sites Web ayant échappé au snapshot. Cette veille peut s'effectuer de deux façons :

- Ou bien par l'envoi d'un mèl à la BnF en vue d'indiquer la référence du site concerné

- Ou bien en passant par une interface sécurisée donnant accès au robot de la BnF et permettant à la Bibliothèque de Lyon de commander directement l'enregistrement du site repéré.

Quel que soit le système choisi, ce système de veille impliquera plusieurs phases :

1. La définition commune du type de sites Web concernés par cette veille
   Il s'agira pour la BnF et la Bibliothèque de Lyon de définir ensemble les critères de pertinence et de sélection de sites Web concernés par ce travail de veille.

2. L'indication à la BnF par un moyen ou un autre des sites Web concernés

Le suivi et l'enregistrement par la Bibliothèque de Lyon de quelques sites Web

Ici, c'est la bibliothèque de Lyon qui s'occupe de l'enregistrement de certains sites Web. Soient qu'ils n'entrent pas dans la définition commune des sites Web à sélectionner pour la BnF (voir mode opératoire précédent), soit que la BnF ne puisse enregistrer fréquemment des sites Web particulièrement intéressants.

1. Sélectionner quelques sites dont l'intérêt mériterait un suivi plus particulier
   Cette sélection doit s'appuyer sur une analyse fine du Web et sur des critères de sélection particuliers.

2. Analyse de la fréquence des mises à jour des sites sélectionnés
   Il s'agit alors d'évaluer la fréquence d'évolution des sites Web sélectionnés. Certains logiciels de veille (agents intelligents) peuvent alerter l'usager qui le demande dans le cas où le site qui l'intéresse aurait été modifié. Toutefois, cette tâche ne saurait être uniquement automatisée. En effet, d'un point de vue informatique, le rajout d'une virgule est un changement.

Il sera donc nécessaire qu'une personne évalue après le logiciel l'ampleur de la mise à jour opérée et l'intérêt ou non d'effectuer un nouvel enregistrement du site.

3. L'aspiration du site

Le site sera alors acquis grâce à un logiciel d'aspiration tel que nous l'avons décrit dans le premier scénario. Les informations issues du MIME seront également récupérées.

4. Production des métadonnées

Il s'agira alors de transformer les informations issues du MIME et les balises HTML en métadonnées XML. Les métadonnées seront alors encapsulées dans le site Web.

5. Enregistrement du site sur CD

Le site et ses métadonnées sont alors enregistrés sur un CD qui subira le même traitement que les cassettes DLT (voir scénario n°1).

6. Conversion en XML

Le site Web est converti en XML est enregistré dans le serveur hébergeant le snapshot. On établit alors un lien hypertexte avec la version précédente du site Web.

Le site est alors envoyé à la BnF si elle le souhaite.

Equipements et moyens nécessaires

Au niveau de l'équipement informatique, sont requis :

- Un serveur stable (Unix) avec d'importantes capacités de stockage
- Au moins un graveur de CD
- Un lecteur/enregistreur de cassettes DLT

En ce qui concerne les logiciels :

- Un logiciel d'aspiration de sites Web

- Un logiciel de conversion du HTML en XML
- Un logiciel de conversion du MIME en métadonnées XML
- Un agent de veille (indication des mises à jour subies par les sites Web sélectionnés)
- Une base de données (MySQL)
- Un moteur de recherche XML

En ce qui concerne le personnel, la participation du service informatique de la Bibliothèque de Lyon sera nécessaire pour la mise en place des outils et leur maintenance. Il serait également souhaitable que le responsable du service participe à l'enrichissement de la base de gestion des migrations ainsi qu'aux prises de décisions pour les migrations successives. Le recrutement d'un assistant serait également souhaitable pour le travail de veille sur les sites Web. Cet assistant devrait avoir des compétences en informatiques lui permettant de mener à bien les opérations concernant le suivi et l'enregistrement des sites que devrait effectuer la bibliothèque.

Avantages et inconvénients

Ce scénario offre de nombreux avantages. Tout d'abord, de tous les scénarios présentés ici, il est celui qui offre la collection la plus intéressante de sites Web à la fois représentative du Web local et permettant un suivi de quelques sites Web. La collaboration avec la BnF permet d'appréhender une approche plus complète de la collection. Par ailleurs, il permettrait d'éviter le gaspillage qui consisterait à refaire ce que la BnF fait déjà, gaspillage qui est inévitable dans les deux autres scénarios.

Par contre, l'inconnue de ce scénario est l'attitude de la BnF. Les modalités du contrat de collaboration ne sont en aucun cas définis pour le moment. La BnF peut demander un niveau de participation plus important. Ce scénario n'est qu'une ébauche de ce que pourrait être une collaboration intéressante mais la mise en place de celui-ci nécessiterait évidemment des négociations poussées entre les deux établissements. Toutefois, dans le cas où ces négociations aboutiraient, ce scénario serait certainement le plus intéressant des trois.