archivage du web français et dépôt légal des publications electroniques

Mehdi Gharsallah

Mehdi Gharsallah
Laboratoire Paragraphe
Université Paris 8
97 rue des marguerites
92160 Antony
01 46 61 93 77 / 06 84 13 06 50

Dépôt légal des publications électroniques et préservation patrimoniale du web français

La production numérique en ligne française représente environ 1,4 % de la production mondiale. Ces données, établies par le RIPE (Réseaux IP Européens) n'ont pour objectif que de donner un ordre de grandeur. Cela étant, on estime la production quotidienne française à 100 000 nouvelles pages web tous les jours dont on sait que près de la moitié disparaîtra en moins d'un mois. En comparant ces chiffres avec le nombre de 300 000 sites français dont 150 000 en *.fr*, on se rend compte que nous perdons environ 1% de la production de documents en ligne toutes les deux semaines. Ce sont au moins 500 000 pages qui disparaissent tous les ans.
Comment qualifier ces chiffres ? Est-ce beaucoup ? Acceptable ? Insupportable ?
Le projet de sauvegarde permettant de stopper ce processus de perte n'est pas impossible à mettre en place et permettrait de juguler la perte irréversible de données.

A l'instar des livres, des phonogrammes, des films sur support photochimique, des vidéos, des logiciels, des bases de données, de la radio, de la télévision ; le World Wide Web est sur le point d'être conservé en France. Dernier-né des représentants de la mémoire de notre société, l'envie, la nécessité de le conserver se font sentir partout dans le monde. La France, fière d'avoir été le berceau de l'invention du dépôt légal, s'est donné pour objectif l'extension de celui-ci aux publications électroniques distantes.
Il reste néanmoins important de bien dissocier l'extension du dépôt légal aux publications électroniques distantes de la préservation patrimoniale du web. En effet, si la cause de cette dernière est tout aussi noble, les difficultés de mise en oeuvre, tant d'un point de vue technique que juridique ou même théorique, se situent du côté du dépôt légal.
Le dépôt légal rassemble tous les documents français destinés à une diffusion publique. Par la conservation systématique d'au moins un exemplaire de tout document publié, il participe à la création de la mémoire de la nation, mémoire à transmettre aux générations futures sans préjuger de ce qui les intéressera. Par ailleurs, la diffusion des documents quels qu'ils soient est favorisée par leur signalement dans de grandes bases de données telle que la Bibliographie Nationale Française. Enfin, le dépôt légal peut constituer un élément de preuve dans la protection du droit d'auteur. Il permet de dater et d'authentifier une publication et, par conséquent, de créer un précédent en cas d'utilisation illégale.
Le dépôt n'est pas toujours obligatoire. De nombreuses formes de dépôt existent, certaines restent à inventer. Le développement des relations de particulier à particulier sur Internet (Peer to peer), laisse imaginer des potentialités de dépôt collaboratif fondé sur ce mode d'échange.
Le dépôt volontaire par exemple est préconisé dans de nombreux pays notamment pour les œuvres autres que littéraires. Il est possible aussi d'imaginer un dépôt communautaire notamment pour le Web. La communauté des internautes peut rapporter les publications rencontrées au cours des navigations à l'organisme dépositaire. Cette approche est déjà expérimentée par des moteurs de recherche qui indexent les sites désignés par des particuliers. C'est le cas dans *l'Open Directory Project*.
Aujourd'hui, le dépôt légal est régi par la loi du 20 juin 1992 et le décret du 31 décembre 1993 qui a encore élargi le champ des documents concernés par l'obligation de dépôt.
Sont concernés : "les documents imprimés, graphiques, photographiques, sonores, audiovisuels, multimédias, quel que soit leur procédé technique de production, d'édition ou de diffusion" mais aussi "les progiciels, les bases de données, les systèmes experts et les autres supports de l'intelligence artificielle dès lors qu'ils sont mis à disposition du public par la diffusion d'un support matériel, quelle que soit la nature de ce support."
Mais pérenniser ces informations et ces supports, c'est aussi en quelques sorte en altérer l'intégrité. Comment garantir l'intégrité d'un document dont on modifie le support et les caractéristiques ?
Le web est plus qu'un hypertexte. C'est un hypermédia. Ce réseau est celui sur lequel la diversité d'éléments est la plus grande. Un site web, c'est-à-dire un ensemble de pages web liées sémantiquement, contient souvent du texte, des images, parfois de la vidéo, des fichiers en streaming au format flash ou real, des fichiers en téléchargement (exécutables, documents, fichiers compressés, etc.).
Si l'on compare ceci aux autres supports connus, la nouveauté ne semble pas flagrante. En effet, un journal ou un film contiennent aussi de nombreux médias différents. La nuance se situe dans la nature hétérogène du format de ces

éléments, c'est-à-dire que dans un journal, le texte et les images sont sur le même support papier et, tant au moment de la lecture, de l'archivage, de la conservation et de la restitution, c'est le support qui agit. Si les fichiers sources qui servent à la fabrication d'un journal devaient être conservés, ce serait bien plus compliqué. Or c'est exactement ce qui se passe sur le web. Tout est dynamique. Les pages n'existent pas en tant que telles. Elles sont reconstituées par le navigateur au moment de la lecture. En dehors de ces moments là, les éléments constitutifs d'une page « dorment » chacun à un endroit différent, qui, par ailleurs peut se trouver dans le même répertoire, sur la machine d'à côté ou à des milliers de kilomètres sans que cela ne fasse aucune différence.
Les archivistes, peu habitués à cette approche hypermediatique des choses ont éprouvé le besoin de dresser une typologie des objets du web. Cette typologie, utile pour mieux connaître le contenu réel des sites web pose aussi une question. Faut-il archiver les sites web dans leur ensemble ou faut-il collecter les éléments de manière indépendante et avec un niveau de priorité différent ? Il semble qu'aujourd'hui les équipes en charge de ces projets dans les bibliothèques nationales aient pris en compte le fait que sur un site web, la structure a une valeur patrimoniale aussi importante que les contenus textes qu'il comporte.
Trois écoles ont émergé au moment des premières réflexions sur la sauvegarde du patrimoine numérique en ligne.
- D'une part, les projets qui ne visaient uniquement l'archivage des « documents en ligne », c'est-à-dire dans la plupart des cas des textes en HTML, ou au format traitement de texte (*Microsoft Word* pour la plupart) ou encore des fichiers *Adobe Acrobat* au format PDF.
- D'autre part, les équipes qui privilégiaient le web comme une création à part entière et qui proposaient un archivage statique des pages en les passant par exemple sur microfilm.
- Enfin, les projets prenant en compte la nature hypermediatique et dynamique du web qui proposait avec toutes les difficultés que cela comporte, un archivage le plus proche possible de ce qui est réellement en ligne.
Il est clair que la dernière approche, bien que présentant le plus de difficultés techniques et théoriques, est la plus adaptée puisque c'est celle qui dénature le moins possible le support collecté.
Quelques problèmes très spécifiques :
Les fichiers en *streaming* :
Le *streaming* désigne le transfert de fichiers en flux, c'est-à-dire que l'utilisateur commence à voir le fichier alors qu'il est encore en train de le télécharger. Les plus courants sont les fichiers *Flash*, *Real* et *Windows Media Player*. Le premier concerne les animations, les deux derniers concernent la vidéo ou le son. La particularité de ces fichiers est qu'ils nécessitent des *plug-ins* (petits programmes que l'on ajoute à son navigateur). Pour ces fichiers, le problème d'un éventuel archivage se décompose en plusieurs parties.
Tout d'abord, dans le cas d'une collecte automatisée, le robot qui va archiver les sites devra posséder la faculté de reconnaître ces fichiers et de pouvoir les lire, donc d'avoir le logiciel adéquat.
Ensuite l'archive devra comporter le fichier ainsi que le programme permettant de le lire quel que soit le standard au moment de la consultation. Par exemple, si un fichier est archivé au format *Flash 4.0*, en vigueur aujourd'hui sans pour autant faire l'objet d'une normalisation, il devra pouvoir être lu dans dix ans que nous en soyons au format *Flash 40.0* ou que ce logiciel ou cette version n'existe plus, ce qui est fort probable.
L'archivage est encore plus ardu lorsqu'il s'agit de fichiers dont le logiciel de lecture est peu répandu. Par exemple, le programme *Chime* qui permet de lire les fichiers du site « molécules interactives » n'existe pratiquement que sur ce site en France.

Une politique cohérente d'archivage doit gérer les problèmes liés aux fichiers multimédias, nécessitant des plug-ins et ne faisant pas l'objet d'une normalisation.
Le suivi des plug-ins, la décision de les télécharger ou non doit être prise au moment de la captation par d'autres moyens que par des robots.
L'idéal serait donc de n'utiliser pour la publication en ligne que des technologies normalisées : HTML, VRML, les images Gif et JPG. Or aujourd'hui, c'est loin d'être le cas. Même si les développeurs multimédias s'efforcent de n'utiliser que des standards, ces derniers sont des standards de faits et non le résultat de normes. Il faudrait donc que l'organisme dépositaire télécharge le plug-in en plus du fichier concerné. Deux problèmes majeurs se posent alors. Premièrement, il est fort probable que l'organisme dépositaire soit missionné pour sauvegarder exclusivement les fichiers de données et non les logiciels pour les lire.
Deuxièmement, les machines et les systèmes pour faire fonctionner ces logiciels n'existeront plus dans peu de temps.

La marche à suivre est par conséquent complexe et difficilement automatisable.
Le choix le plus simple serait de ne capter que les fichiers standardisés. Cela impliquerait une perte non négligeable de contenus. Il faudrait donc télécharger en plus les logiciels adéquats pour lire les formats de fichiers exotiques. Outre les problèmes de droits engendrés par cette démarche, un tri doit être fait pour éviter de télécharger un nombre

incalculable de fois le même *plug-in*. En revanche, il est vital pour l'archive de comporter les logiciels tels que *Chime* non seulement téléchargé mais installé. Le web étant constitué d'une somme d'exception de ce genre, le casse tête est sans fin. Des choix seront faits par l'organisme en charge du projet. Espérons qu'ils soient judicieux.

La diffusion en direct

De plus en plus souvent les sites Internet proposent du « live ». Les données sélectionnées par le visiteur du site sont diffusées en direct par le serveur. Pour archiver ces données *en temps réel,* quelques problèmes spécifiques se posent.

Le cas le plus simple est celui de la webcam. La webcam est le moyen le plus ancien de délivrer des images quasi instantanément sur le web. Ces caméras, placées dans divers endroits de la planète, envoient une image toutes les 30 secondes ou toutes les minutes. L'image s'affiche sous un format classique (gif ou jpeg) dans le navigateur et se «rafraîchit » à une fréquence qui implique que ce système se rapproche plus de la projection de diapositives que du film. L'archivage en continu des webcams pose deux problèmes.

Le premier est de connaître précisément la fréquence de rafraîchissement de l'image et de caler le robot sur cette fréquence exactement.

Le deuxième est que l'image affichée porte toujours le même nom de fichier. Il ne faut donc pas écraser le précédent fichier à chaque fois que l'on sauvegarde une image. Ceci dit, on peut se poser la question de l'intérêt d'archiver les webcams en continu. En effet, le fait que la fréquence d'affichage de ces caméras soit très basse, montre bien que la valeur informative de ces images n'est pas dans le flux, ni dans le mouvement. Par exemple, l'intérêt de la webcam de la Place de l'Etoile, outre la vue, n'est pas de voir les voitures rouler, mais de savoir par exemple s'il y a beaucoup de trafic. Dans ce genre de cas, l'archivage par échantillonnage pourrait présenter un intérêt. Cette notion chère aux théoriciens de l'information et au traitement du signal consiste à relever de manière régulière un fragment représentatif d'un signal continu.

Dans le cas de l'archivage des webcams, l'échantillonnage consisterait à déterminer une fréquence pertinente de captation des images. L'ensemble des images collectées devant être représentatif de la diffusion complète. La fréquence est évidemment à déterminer au cas par cas et, par conséquent non automatisable dans l'état actuel des connaissances en reconnaissance d'image.

Les webcams offrent des images en direct ou en léger différé, mais ni son, ni images animées. C'est pourquoi, le logiciel *Real* occupe aujourd'hui une place de quasi-monopole dans le secteur de la retransmission audiovisuelle en direct sur Internet. Les inconvénients de *Real,* en dehors de son prix, est qu'il nécessite un serveur particulier et qu'il ne se laisse pas archiver facilement. En effet, pour sauvegarder du *Real* (diffusé en direct) il faut aller directement sur le serveur. Sur ce serveur, deux possibilités existent, soit les diffuseurs ont eux-mêmes créé une archive temporaire et, dans ce cas, il est possible avec leur autorisation de la récupérer. Soit ils n'archivent rien de leurs diffusions et dans ce cas, aucune captation n'est possible.

Les bases de données et les sites dynamiques

Pour être archivées, les bases de données doivent surmonter des problèmes techniques et juridiques délicats. Ces bases peuvent en effet être directement liées à des pages web créant ce qu'on appelle des sites dynamiques.

La particularité de ces sites est de proposer aux utilisateurs des pages entièrement personnalisées. Ces pages sont constituées en fonction des requêtes de l'utilisateur à partir d'une base de données d'éléments. Pour le robot qui parcourt et archive les pages, il peut éventuellement être possible d'archiver une version mais il lui sera de toutes façons impossible de visiter toutes les combinaisons de pages possibles.

C'est l'une des différences majeures entre les hypertextes et les bases de données. Il est possible de dresser la carte d'un hypertexte aussi complexe soit-il alors que seul le modèle d'une base peut être dessiné.

Cette particularité est mise en évidence par les 5ème et 6ème principes du *Rhizome* établis par Deleuze et Guattari.

« […]Principe de cartographie et de décalcomanie : un rhizome n'est justiciable d'aucun modèle structural ou génératif. Il est étranger à toute idée d'axe génétique, comme de structure profonde.[…]

Gilles Deleuze et Félix Guattari ; *Rhizome*, Paris, France, Les Éditions de Minuit, 1976.

A priori, la seule méthode pour archiver ces sites dont le nombre ne fait qu'augmenter est de se faire « livrer » par ses concepteurs la base de données ainsi que les moyens d'y accéder.

Ubiquité des documents et notion d'original

L 'une des particularités des documents sur le Web est qu'ils peuvent être « édités » ou publiés, c'est-à-dire, proposés au public en ligne, par plusieurs personnes ou plusieurs sites.

En dehors des questions liées aux droits d'auteurs, il est tout à fait possible et même fréquent sur Internet de dupliquer un document. Ces pratiques sont liées à la nature numérique du média et au support en réseau maillé.

L'archivage de ces documents peut soulever certaines questions. La préservation patrimoniale d'une part et le dépôt légal d'autre part sont fondés sur les notions d'auteur et d'original.

Le musée du Louvre pourrait-il conserver la *Joconde*, toutes ses copies, tous les tableaux qui prennent la *Joconde* comme base d'inspiration ?

C'est à cette difficulté que sont confrontés les organismes de préservation. De plus, dans l'univers des médias numériques, il n'existe pas de différence entre un original et sa copie ; à tel point que ces notions perdent leur sens.

Un propriétaire de site peut, s'il est peu scrupuleux ou s'il en a l'autorisation, copier le contenu d'un document et le réintégrer dans son propre site, comme s'il publiait un livre uniquement en compilant des photocopies d'autres livres.

Une autre méthode consisterait à « appeler » l'adresse du document de manière à l'insérer dans son site non pas de manière physique, mais de manière logique. Les caractéristiques de délocalisation des publications électroniques et de virtualité des pages web permettent cela.

Le problème qui se pose alors est que le document en question va être archivé deux fois. Cela ne serait pas grave si ce n'est que les archivistes ne pourront pas dire lequel de ces documents est « intègre » puisque la notion d'original aura disparu. Le problème se pose ainsi : lorsque deux textes similaires sont collectés sur deux sites différents, comment établir lequel est l'original ?

Pour les livres, le problème ne se pose plus puisqu'un livre ne peut être copié que lorsqu'il est publié et, par conséquent déposé.

A terme, il en sera peut-être de même pour les sites web. Ils seront peut-être soumis au dépôt avant mise en ligne.

Cette initiative n'est pas forcément souhaitable dans la mesure où elle réduirait de manière considérable la facilité de publication ainsi que l'anonymat des auteurs. Cela engendrerait de manière certaine une baisse de production des pages personnelles ou les laisserait en dehors de la préservation patrimoniale.

Ce constat est un argument majeur pour la séparation définitive des démarches de préservation patrimoniale de celles du dépôt légal.

Dans l'état actuel des choses et tant que la tâche d'archivage sera en retard par rapport à la production, le dépôt légal des publications électroniques distantes remplira son rôle de préservation patrimoniale mais aura des difficultés à assurer de manière certaine la propriété intellectuelle d'une publication.

La notion de frontière

Enfin, il est important d'aborder une des principales caractéristiques du World Wide Web. Le Web est mondial et ouvert.

La structure hypertextuelle du Web ne permet pas de délimiter clairement un site. Un site web peut être constitué simultanément d'éléments provenant de toute la planète. La notion de frontière est absurde et contre nature lorsque

l'on parle du Web. Néanmoins il faudra trouver les critères qui permettront de définir et de sélectionner un site français d'un autre.
En effet, le dépôt légal ne peut s'appliquer que sur un domaine restreint et clairement défini tel que la production nationale. Dans l'état actuel de la législation concernant le dépôt, il est impossible pour un pays comme la France de prétendre déposer une publication étrangère ou dont la provenance n'est pas identifiée.
De la même manière, il est difficile de déposer une œuvre complète, un site Web par exemple, sans que la provenance des éléments constitutifs de cette œuvre ait été identifiée.

Les particularités du web en font un objet qui ne se laisse pas collecter facilement. Des questions théoriques et techniques se posent et font de cette entreprise une des plus complexe jamais posée aux archivistes, documentalistes et bibliothécaires. Cependant plusieurs approchent existent et certains pays ayant commencé cette tâche montrent le chemin à suivre.

Panorama mondial de l'archivage du web
Il s'agit ici de montrer à travers cinq projets différents la multiplicité des approches possibles : intégrale (Internet Archive), exhaustive (Kulturar W3), sélective (BNQ), automatisée ou manuelle. Chacune d'entre elles apporte son lot d'avantages et d'inconvénients.
L'approche intégrale, idéale dans une optique de préservation patrimoniale, archive sans se soucier de la provenance, de l'intégrité et de la qualité. En revanche, elle ne permet pas la mise en place d'un dépôt légal et ne s'inscrit pas dans un cadre légal ou simplement déontologique.
L'approche exhaustive, souvent automatisée, permet de recueillir un corpus important à moindre frais mais ne permet qu'une sélection sommaire, pas ou peu d'indexation et bloque souvent sur des questions techniques. La qualité même de l'archive est mise en question.
L'approche sélective, souvent manuelle, offre une archive de grande qualité, une sélection et une indexation manuelle des contenus, des autorisations… En revanche ses faiblesses se situent au niveau du prix et de la lenteur relative de ces démarches.

Internet Archive

Dans les années 80, Brewster Kahle est étudiant au MIT (Massachusetts Institute of Technology) et il collabore à la création de *supercomputers* à la *Thinking Machines Corporation.* Dès 1989, il fonde WAIS (Wide Area Information Service) qui est l'un des premiers services publics d'Internet, une sorte de gigantesque base de données en réseaux. Il avait d'ores et déjà pressenti l'intérêt commercial de la croissance fulgurante d'Internet et il revendra WAIS Inc. à AOL (America On Line) en juin 1995.
En mars 1996, il fonde *Internet Archive,* un projet de recherche visant à archiver le web mondial. Ce projet se transformera vite en une société basée à San Francisco. Cette société mettra sur le marché un outil commercial appelé *Alexa* dès Juillet 97.
Alexa est un logiciel composé de plusieurs modules :
une partie client qui s'affiche sur le navigateur Internet.
une partie robot qui « butine » le web, rapatrie, indexe et rend disponible un nombre colossal de pages. Cette partie du logiciel s'appelle la *WayBack Machine.* Cet outil donne des indications sur la fréquentation, le renouvellement, le nombre de liens, les versions précédentes du site ainsi que les sites liés ou les sites donnant des informations sur des sujets approchants. Il permet surtout d'accéder aux versions passées des sites archivés par Internet Archive.
Même si la pérennité de ces archives n'est pas assurée, si l'indexation est sommaire (par date uniquement). Même si les articles payants sur le site actuel sous la rubrique « archives » sont ici gratuits, la valeur patrimoniale d'un tel travail est indéniable. L'approche d'Internet Archive consiste à démarrer les projets dès qu'ils sont techniquement possibles et à analyser les impacts ensuite. Cette manière de faire, très différente de celle pratiquée en France, porte ses fruits dans la course contre le temps que représente l'archivage du web. En effet, sur ce dossier en particulier, la France a une politique d'attente qui lui porte préjudice. Le souhait de ne commencer à archiver que lorsque le cadre législatif sera prêt est une mauvaise technique. Ainsi, le corpus de départ de l'archive française sera issu d'Internet Archive.
L'approche d'Internet Archive est intéressante sur plusieurs points. Tout d'abord son aspect international correspond

bien à la logique du web qui n'a ni frontière ni nationalité.
Mais l'aspect international de cette approche ne permet pas aujourd'hui de dépôt légal. L'accès gratuit à certains fichiers peut porter préjudice aux sociétés souhaitant vendre leurs archives, et l'absence de demande d'autorisation ne respecte pas les droits moraux des auteurs.

Le dépôt légal n'est évidemment pas possible avec Internet Archive. L'approche internationale, l'absence de vérification des publications collectées, le manque de stabilité de l'association, le manque de garantie concernant la pérennité de l'archive font que cette opération serait impossible. Seul l'aspect patrimonial est intéressant.

L'archive en elle-même ressemble plus à un amas de documents classés par date et par URL qu'à une base documentaire digne de ce nom. Aucune indexation du contenu n'est faite. Personne n'est aujourd'hui capable de donner le nombre de sites collectés. Il semblerait que les membres d'Internet Archive ne souhaitent communiquer que sur la taille de l'archive en giga octets plutôt que sur le nombre de fichiers contenus.
Enfin, la pérennisation des sites n'est pas prévue et aucune réflexion n'a été menée sur le long terme.

Cet avant-goût de ce que pourrait être une archive patrimoniale du web français laisse songeur.
La capacité d'analyse offerte par un tel corpus est sans fin pour des chercheurs souhaitant étudier l'évolution d'un média aussi volatil que le Web.

Bibliothèque Nationale du Québec

Certainement l'une des institutions ayant le plus communiqué sur son projet, la Bibliothèque Nationale du Québec (BNQ) a ouvert la voie aux institutions dans le domaine de l'archivage des publications électroniques émanant du web.
Son approche extrêmement sélective a conduit la BNQ à prévoir une préservation à long terme non pas pour les sites web eux-mêmes mais pour les documents textuels qu'ils pourraient contenir, considérés comme importants et n'existant pas sur d'autres supports.
Le projet se découpe en trois phases :

Phase I : février - septembre 2001
Dépôt d'environ 1 000 titres signalés par une vingtaine de ministères et organismes gouvernementaux dans la Banque des publications gouvernementales accessibles par Internet. Cette Banque est diffusée par le MRCI (Ministère des Relations avec les Citoyens et de l'Immigration) sur le portail du gouvernement du Québec.

Phase II : mars - décembre 2002
Dépôt rétrospectif de l'ensemble des publications assujetties au dépôt légal diffusées sur les sites Web de quatre ministères invités (Culture et Communication, Education, Finances et Ressources Naturelles). On estime le corpus à 3 000 titres.

Phase III : janvier 2003
Dépôt de l'ensemble des titres diffusés par les ministères et organismes gouvernementaux (250 ministères et organismes, près de 50 000 titres).

La politique générale est donc de préserver en priorité les documents des sites gouvernementaux, essentiellement au format *Adobe Acrobat (.pdf)* ou traitement de texte. Le choix de fichiers aux formats propriétaires non normalisés est étrange venant de la part de documentalistes.

PANDORA et la Bibliothèque Nationale d'Australie

En juin 1996, la Bibliothèque Nationale d'Australie met en place le projet PANDORA (Preserving and Accessing Networked Documentary Ressources of Australia). De juin 1996 à fin 1997, ses chercheurs développent une politique et des procédures de sélection, de capture et d'archivage pour l'accès à long terme des publications électroniques australiennes. Dès la fin 1997, le concept est déposé et une première archive d'environ 229 documents est créée.
A l'heure actuelle, PANDORA a conçu grâce au SCOAP (Selection Committee on Online Australian Publications), un processus de travail ainsi qu'une liste des problèmes rencontrés et des spécifications techniques.

Le processus d'archivage est le suivant :
- Evaluation de la publication de manière à déterminer sa structure, ses particularités, etc.,
- Obtention de la permission de l'éditeur d'archiver sa publication,
- Catalogage de la publication sur la base de données de la Bibliothèque Nationale Australienne de manière à s'assurer que cette publication est accessible, notamment par la création d'un hyperlien vers elle,
- Envoi d'une requête pour archiver la publication dans « l' *Archive Management System* ». Cette action lance le robot *Harvest* (logiciel de collecte de fichiers en ligne) pour qu'il récupère les fichiers sur le web. Parfois c'est *WebZip* (autre logiciel de sauvegarde de site qui a la particularité de compresser les archives qu'il produit) qui est utilisé.
- Comparaison entre l'archive récupérée du web et la source en ligne de manière à vérifier que toutes les pages et tous les liens ont été sauvegardés correctement,
- Renvoi d'un rapport de vérification signifiant si la copie est conforme ou s'il faut corriger des erreurs,
- Construction d'une page d'entrée pour la publication archivée et l'allocation d'une PURL (Persistent Uniform Ressource Locator) qui est plus stable qu'une URL classique car elle permet de suivre les changements d'adresses,
- Vérification périodique de la collecte et comparaison entre les pages en ligne et l'archive, surtout pour les publications périodiques.

Tel que décrit, le travail de PANDORA n'est que partiellement automatisé. Les réalisateurs de ce projet tentent ainsi de conserver un équilibre entre rapidité de collecte, pertinence et qualité de l'information archivée. Leurs critères de sélection concernant la qualité des publications sont d'ailleurs assez sévères. Leur argument est le suivant : contrairement à l'imprimerie, les éditeurs en ligne ne font pas de sélection, elle doit donc être faite au niveau de la collecte.

KulturarW3 (Suède)

Le projet KulturarW3 est sans doute le projet semi-sélectif automatisé le plus abouti. En effet, dans le panorama mondial des projets d'archivage de données en ligne, le projet suédois est apparemment le seul qui ait choisi une approche entièrement automatisée.
Commencé en septembre 96 au moment où la Bibliothèque Royale venait de recruter un ingénieur, Johan Palmkvist, le projet KulturarW3 a bénéficié d'une aide financière initiale de 3 millions de couronnes suédoises (~366 000 EUR) pour établir la première étude de définition concernant les différentes méthodes de collecte, de préservation et de mise à disposition des documents en ligne suédois.

Les Suédois ont choisi le principe d'une approche exhaustive et ont effectué 10 collectes du web depuis janvier 97 dont trois sont maintenant partiellement accessibles en ligne. Le robot utilisé est un robot d'indexation modifié en robot d'archivage.
Les domaines exploré sont *.se*, *.com*, *.net*, *.org*, et *.nu*. La taille du web suédois représentait, en 2000, 5 millions de pages réparties sur 31 000 sites (25000 en *.se* et 6000 sur les autres domaines). Avec les images, les sons etc., la taille de la base s'élève à environ 9,7 millions de fichiers soit 200 Giga bites. Les ingénieurs du KulturarW3 précisent à ce sujet que les problèmes ne viennent pas de la taille de l'archive mais du nombre important de fichiers hétérogènes et de liens à gérer.

Depuis le 8 mai 2002, un décret autorise la Bibliothèque Royale à acquérir, préserver, et rendre accessible tout document publié sur le web suédois.

Cité en exemple par toutes les équipes internationales, le projet KulturarW3 montre le chemin à suivre vers la préservation patrimoniale des contenus en ligne mais aussi vers le dépôt légal des publications. Leur volonté de collaboration avec les autres états est une aubaine pour les

CoBRA+ (International)

Ce dernier projet est certainement le plus proche de la problématique de l'archivage des données électroniques nationales en France. En effet, l'équipe de Kaisa Kaunonen, chargée du projet à la Bibliothèque Nationale de Finlande, travaille à la confection d'un robot « Harvester » en cours de test. Si ce robot s'avère efficace, il sera employé par tous les partenaires de CoBRA+.
La Bibliothèque Nationale de France a pu contribuer à l'examen des spécifications et au test du robot de collecte développé dans le cadre de NEDLIB. Enfin NEDLIB apporte un modèle fonctionnel de système de dépôt, d'archivage et de communication de documents électroniques qui sert de base à la réflexion pour l'intégration des différentes composantes existantes dans un seul modèle d'archivage.

Modèle Fonctionnel de NEDLIB (dSEP)
Phase 1 : Sélection
L'institution de dépôt définit ses propres critères de sélection dans le cadre des lois de dépôt légal nationales. Le modèle fonctionnel de NEDLIB permet d'accepter toute forme de publications électroniques (disquettes, cd-rom, Bases de données, sites web…) et est suffisamment ouvert pour intégrer de futurs nouveaux formats.
Phase 2 : Acquisition
Bien que le processus d'acquisition soit dépendant du système automatisé existant dans l'institution de dépôt, dSEP s'interface avec ce système pour permettre ultérieurement une vérification de l'acquisition.
Phase 3 : Capture
Durant cette phase, la publication électronique est transférée du système de publication de l'éditeur vers le système de dépôt de l'institution. Les publications sont accueillies sur des supports tels que disquettes ou cd-rom ou bien via un réseau.
Phase 3.1 : Authentification
La source de la publication est vérifiée par l'intermédiaire d'un registre des éditeurs.
Phase 3.2 : Contrôle qualité
La publication est vérifiée de manière à s 'assurer qu'elle est complète et qu'elle ne comporte aucun virus.
Phase 3.3 : *deposit package*
La publication ainsi que les informations concernant son support (configuration système) sont sauvegardés sous la forme d'un ensemble qui contient la publication ainsi que tous les éléments nécessaires à sa lecture (logiciels…)
Phase 4 : Enregistrement
Les informations concernant la publication et son mode de lecture sont enregistrées dans le système.
Phase 5 : Vérification
Cette phase contrôle que la publication est toujours accessible dans le système d'un point de vue physique et logique.
Phase 6 : Description
Indexation de la publication dans le catalogue général de l'institution de dépôt.
Phase 7 : Stockage
La publication est stockée sur support. Ces supports sont régulièrement renouvelés.
Phase 8 : Préservation
Lorsque le risque de ne plus pouvoir accéder à la publication devient important, plusieurs méthodes sont mise en place pour assurer son accès.
Phase 8.1 spécifications des formats de fichiers
Pour les documents codés dans des formats standards tels que HTML ou XML, faire évoluer les documents avec ces normes.
Phase 8.2 migration
Transférer les publications codées dans des formats propriétaires vers des formats normalisés.
Phase 8.3 émulation
Cette phase implique le développement de logiciels capables de simuler le système d'origine de la publication.
Phase 9 : Mise à disposition
Phase 10 : Monitoring
Le système est régulièrement testé pour valider son bon fonctionnement , notamment l'accès aux plus ancienne publications.

La situation française

La Bibliothèque Nationale de France et l'INA mènent depuis 1998 des études sur la question. Au départ, la BNF a eu du mal à trouver la direction vers laquelle elle souhaitait s'orienter et a commencé par sélectionner les sites sur le critère de la densité textuelle : elle considérait alors qu'un site contenant beaucoup de texte avait plus de valeur patrimoniale qu'un site plus graphique. Au fur et à mesure des concertations et des études, le jugement de la BNF a évolué vers une approche plus intéressante.
Observant les projets des différents pays précurseurs, la BNF a conclu que les approches sélectives ou exhaustives seules n'étaient pas satisfaisantes.
Le projet actuel repose donc sur le mariage de ces deux approches.
Cette approche intégrative se fonde sur le développement d'un outil central, un moteur, qui permettra un repérage exhaustif des sites français. La constitution d'une archive d'un échantillon représentatif de ces sites sera réalisé.
L'échantillonnage sera effectué en fonction d'un critère de notoriété des sites, c'est-à-dire qu'il sera fondé sur le nombre de liens pointant vers ses pages. Ce critère de notoriété est aujourd'hui utilisé par les moteurs de recherches, notamment *Google*. Bien que discutable, ce critère présente de nombreux avantages. Il permet grâce à un critère relativement objectif de mesurer le caractère incontournable d'un site.
Autrement dit, dans un domaine donné, plus un site va être cité par ses pairs, plus le moteur va considérer que c'est une référence du domaine. Il deviendra en quelque sorte leader d'opinion d'une communauté et devra être présenté comme prioritaire à l'archivage par rapport aux autres sites de la même constellation. D'autre part les contenus, sont souvent issus des sites majeurs et repris par les plus petits. Il est donc fondamental de conserver les sites majeurs pour préserver les versions originales de ces contenus.
Cet échantillon sera signalé aux services du dépôt légal en mettant en exergue les sites nécessitant un traitement individualisé.
Pour les sites ayant été repérés mais ne pouvant faire l'objet d'une aspiration à distance, il sera demandé aux éditeurs de fournir leurs contenus par dépôt volontaire.
Cette approche offre l'avantage de marier l'efficacité d'un repérage large et systématique avec la précision et le suivi d'une collecte manuelle.

Evolutions de la loi

C'est le 12 novembre 2003 qu'a été adopté en conseil des ministres le projet de loi relatif au droit d'auteur et aux droits voisins dans la société de l'information. Il entraîne une modification du deuxième alinéa de l'article 1er de la loi n° 92-546 du 20 juin 1992 relative au dépôt légal qui est remplacé par les deux alinéas suivants : « *Les logiciels et les bases de données sont soumis à l'obligation de dépôt légal dès lors qu'ils sont mis à disposition d'un public par la diffusion d'un support matériel quelle que soit la nature de ce support* » et « *sont également soumis au dépôt légal les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication publique en ligne* ». Sous la conduite du ministère de la Culture et de la Communication, ce texte est une adaptation de la directive européenne 2001/29 du 22 mai 2001 relative à l'harmonisation de certains aspects du droit d'auteur qui vise à rapprocher les législations des Etats membres en matière de propriété littéraire et artistique en prenant en compte l'impact des technologies de l'information et de la communication. Ce projet de loi se décompose en cinq parties qui visent chacune des points particuliers.
D'un point de vue archivistique et patrimonial, c'est le titre IV concernant le dépôt légal qui marque un véritable tournant. En effet, ce texte charge la Bibliothèque nationale de France (BNF) et l'Institut national de l'audiovisuel (Ina) de mettre en place un dépôt légal des publications électroniques distantes. Modifiant ainsi la loi de 1992, ce projet donne enfin le cadre juridique nécessaire à la sauvegarde du patrimoine numérique français. En effet, depuis 1995, aucune institution n'acceptait en dépôt les sites web. Ainsi tout le web français de la première génération a disparu sans laisser de traces. Bien sûr, d'un point de vue juridique, l'événement est majeur mais c'est aussi une véritable révolution pour les auteurs de pages web qui vont certes avoir l'obligation de dépôt (mais la majorité du corpus devrait être collectée de manière automatique), mais aussi la reconnaissance de leur travail d'auteur en garantissant la date de dépôt ainsi que l'originalité de l'œuvre. Le web français va donc être soumis au dépôt légal et, par conséquent, préservé. Mais quel web français ? Pour Jean-Michel Rodes, directeur de l'Inathèque, « *un site français est un site dont l'adresse postale du propriétaire est sur le territoire. Cela représente environ 300 000 sites. 150 000 en .fr et à peu près autant en .com, .org et .net* ». 70 % des sites devraient être collectés de manière automatique par l'Ina comme par la BNF. Les 30 % restant seront collectés manuellement ou soumis au dépôt volontaire de leurs propriétaires. Les prototypes d'aspirateurs de sites ont d'ores et déjà réglé le problème des sites dynamiques mais pas encore ceux se trouvant derrière des formulaires ("deep web") ou avec de grosses bases de données. Les sites dynamiques peuvent donc être aspirés sous forme statique mais la base de données qui sert à générer les pages ne saurait, pour le moment, être archivée. L'Ina et la BNF vont donc travailler de concert sur une liste commune de sites, se partageant le travail en fonction des contenus. Ainsi, dans une logique de continuité des collections, l'Ina sera en charge de l'archivage des flux et devra sélectionner, collecter et conserver les sons et images animées du web national. La BNF se chargera des éléments documentaires plus classiques (texte, images fixes…). La difficulté concernera les sites réellement mixtes qui proposent toute sorte de contenus et qu'aucune des deux institutions ne peut revendiquer comme attribuables à sa collection existante. Ce partage des responsabilités peut apparaître comme surprenant de prime abord mais l'enjeu est tel qu'il en va de la survie de l'institution à qui on ne confierait pas cette tâche. Si la BNF archive les télévisions et les radios en ligne, il est fort probable que l'Ina devienne un musée. Cette réflexion pose d'ailleurs la question du rôle des Archives nationales dans ce nouveau contexte. Il semblerait qu'un autre projet contraigne les institutions publiques à déposer leurs sites mais aussi leurs intranets aux Archives de France. L'approche choisie pour la sélection et la collecte est exhaustive et semi-automatisée. Cela signifie qu'aucune sélection qualitative n'est effectuée *a priori*. Tous les sites français identifiés vont donc être aspirés lorsque la technique le permettra et collectés manuellement si l'automatisation est impossible. Cette approche n'exclut pas la mise en place de règles de priorité des éléments à archiver. En effet, puisqu'il est impossible d'archiver le web français en continu, certains contenus, jugés plus urgents à archiver, seront traités en premier. « *A l'instar du moteur de recherche Google, nous retenons le critère de notoriété des sites comme élément de gestion des priorités. Il reste tout de même à éclaircir les questions de granularité :Google traite la page alors que nous traiterons les sites* », déclare Julien Masanès, en charge du programme à la BNF aux côtés de Catherine Lupovici, directrice du département de la numérisation. La loi devrait voir le jour dans le courant de l'année 2004 bien que le calendrier risque d'être retardé par l'intervention exceptionnelle de la Cnil (Commission nationale de l'informatique et des libertés). En effet, pour la première fois, la Cnil sera consultée pour une loi concernant le droit d'auteur et le dépôt légal notamment à cause de la conservation de données personnelles concernant les propriétaires de sites. Par ailleurs, l'impact de la législation sur les aspects patrimoniaux n'est pas sans conséquences. Si l'Ina et la BNF bénéficient d'une exception leur permettant de collecter des publications sans autorisation de leurs auteurs et de les diffuser, ce sera uniquement sur place et pour des chercheurs accrédités. Alors que Jean-Michel Rodes affirme qu'« *il faut construire l'archivage du web à l'image de celui-ci* », il est regrettable que 99 % de contenus diffusés librement soient soumis à des dispositions restrictives dues au 1 % de contenus à vocation commerciale. Les mises

en application de cette loi sont évidemment à suivre de très près, notamment pour toutes les questions relevant du droit d'auteur qui seront traitées dans nos colonnes prochainement. Notons tout de même le changement notoire de statut pour les agents de l'Etat. « *Dans la mesure strictement nécessaire à l'accomplissement d'une mission de service public, le droit d'exploitation d'une œuvre créée par un agent de l'Etat dans l'exercice de ses fonctions ou d'après les instructions reçues est, dès la création, cédé de plein droit à l'Etat.* » De plus, le chapitre III du projet transpose les articles 6 et 7 de la directive qui visent à lutter plus efficacement contre la contrefaçon. Le texte introduit donc des sanctions en cas de contournement d'une mesure technique efficace de protection d'une œuvre. Autant dire qu'en matière de contrefaçon, à l'instar de la sécurité informatique dans le commerce électronique, c'est l'obligation de moyens qui et prise en compte et non l'obligation de résultats. Ce texte extrêmemnt riche en nouveauté fera sans aucun doute l'objet de débats. La remise en question du droit de copie privée concernant les logiciels sera discutée. Notons tout de même une avancée majeure dans les exceptions aux droits d'auteurs concernant « *la représentation par des personnes morales en vue d'une consultation strictement personnelle de l'œuvre par des personnes atteintes d'une déficience motrice, psychique, auditive ou de vision[…] ».*

L'exemple du site du Premier Ministre

Conformément à la demande des Archives Nationales, le site du Premier Ministre, comme tous les autres sites gouvernementaux, en *.gouv,* est livré régulièrement sur support physique (en l'occurrence le disque compact). Cependant, conscient de l'aspect nécessaire mais insuffisant de cette démarche, le responsable du site, Monsieur Benoît Thieulin, a entrepris d'archiver les différentes versions de ce site et de les proposer au public.

Des données existantes :

La prévoyance de ses prédécesseurs qui n'avaient pas détruit les versions antérieures du site a permis d'accéder aux données sauvegardées.

En revanche, la multiplicité des techniques employées sur ces sites a rendu la tâche plus ardue. Le souci de Benoît Thieulin était d'une part de conserver le patrimoine que représentent ces sites mais aussi de le restituer au public.

Fréquence d'archivage :

Le site n'est pas archivé de manière régulière. La notion de version intéresse Benoît Thieulin, c'est-à-dire qu'il s'efforce de conserver une image exacte de la version d'un site avant évolution vers une autre formule ou un autre gouvernement. Il fonctionne avec la méthode de l'instantané (Snapshot).

Uniformisation des technologies :

Les premiers sites du gouvernement sont des sites statiques, n'utilisant que des technologies normées comme le HTML et les formats d'images Gif ou Jpeg.

Puis, pour des questions pratiques, les sites sont devenus dynamiques : pour créer les pages, il est nécessaire de faire appel à différents éléments dans une ou plusieurs bases de données. Dans le cas du site du Premier Ministre, la technologie propriétaire Cold Fusion avait été choisie. A ce sujet, Benoît Thieulin regrette d'ailleurs que les logiciels libres et ouverts n'aient pas été plus populaires à cette époque car, dit-il, s'il devait refaire un choix, ce serait vers ces solutions qu'il s'orienterait.

Pour l'accès aux archives, il fallait uniformiser les technologies. Le site a été standardisé. En effet, il a été demandé à un prestataire de convertir les sites dynamiques en sites statiques renforçant encore l'image de photographie instantanée. Conscient de l'aspect perfectible de ce choix, Benoît Thieulin a préféré, à l'époque, et en l'absence de recommandation sur le sujet, aller vers les technologies standardisées et, par conséquent pérennes.

Navigation par strates et URL déclinable

A l'adresse <http://www.archives.premier-ministre.gouv.fr>,

On accède à un menu de navigation permettant de voir les quatre versions du site de 1996 à 2002.

Bilan et perspectives

Ce travail de recherche n'a pas la prétention de répondre à toutes les questions posées par l'archivage du web français. Cela étant, au bout de ces trois années d'études, un certain nombre de points clefs émergent. Ce travail ouvre des perspectives de recherches dont les applications pourraient répondre aux besoins des programmes d'archivages qui se mettent en place.

La question fondamentale posée est la suivante : « pourquoi archiver le web ? ».
Répondre à cette question ne se fait pas sans la décomposer en problématiques élémentaires.
Le cadre théorique dans lequel s'inscrit l'archivage du web français est celui de la mémoire et plus particulièrement celui du « devoir de mémoire ».
Comprendre pourquoi on archive depuis de longs siècles c'est avant tout comprendre ce qu'est une archive.
Si l'on considère l'archive comme un support externe de mémoire, alors comprendre l'archive et son utilité revient à se poser la question de l'utilité de la mémoire.
Or la mémoire sert à reproduire des comportements qui ont favorisé l'adaptation.
Cette adaptation des comportements à des situations nouvelles pourrait, pour les cogniticiens, être assimilée à la définition de l'intelligence.
La mémoire permet donc de reproduire des comportements intelligents. Elle est source d'intelligence.
Ce raisonnement infère que l'archive, support externe de la mémoire d'une société, facilite la reproduction de comportements intelligents, adaptés.
Mais tout n'est pas archivé. Une sélection s'effectue selon le critère complexe de la valeur patrimoniale d'un document.
Archiver le web français est utile si on considère que ce support de communication et d'information possède une valeur patrimoniale d'une part dans son contenu mais aussi dans sa forme.
Les critères de mesure de la valeur éditoriale d'un support sont multiples. Pour le web les suivants ont été retenus :
Le web offre une grande richesse de type de contenu. Ses aspects hétérogènes et multimedia en font un support de texte, d'image, de son, de vidéo, de logiciel, d'information, d'art, de jeux etc.
Le nombre colossal de publications et d'auteurs lié à la facilité relative de publication, offre un reflet de la culture de notre société d'une justesse jamais atteinte par un support de publication. Associant les aspects fonctionnels, culturels, et politiques, le web couvre la quasi-totalité des sujets et centres d'intérêts de la population.

Le fondement de ce travail apparaît alors de manière claire. L'archivage du patrimoine est utile et fait progresser la société en augmentant sa culture.
Le web fait partie de ce patrimoine. Il doit par conséquent être conservé.

La notion de dépôt légal ne semble pas nécessaire à la mise en place d'un archivage patrimonial du web. Cela étant, le dépôt légal présente des enjeux patrimoniaux et juridiques qui en feraient une raison suffisante à la mise en place de la collecte et de la conservation du web français.

Néanmoins ce qui, au démarrage de cette étude, était une présomption devient aujourd'hui une certitude.
Le dépôt légal des publications électroniques en ligne doit impérativement être dissocié de la collecte et de la conservation patrimoniale de ces mêmes publications.
A l'exception des Pays-bas, où la bibliothèque nationale a constitué une collection nationale de publications en dépôt par voie d'accords avec les éditeurs, la plupart des pays ont recours à un instrument légal, sous une forme ou une autre, pour assurer l'exhaustivité de leur collection nationale constituée par voie de dépôt.
C'est justement cet aspect réglementaire qui fait la limite du dépôt légal dans le cas des publications électroniques distantes.
En effet, toutes les publications « hors la Loi » soit pour contenu raciste ou violent, soit pour non-respect du droit d'auteur, ne pourraient être déposées volontairement par leurs auteurs.
Loin de penser qu'il faut passer outre ces infractions, il ne peut être nié que l'aspect législatif du dépôt légal nui à l'exhaustivité de la collection.
Or l'exhaustivité de la collection fait partie des principes fondamentaux du dépôt légal.
De plus, la provenance de la publication fait aussi partie intégrante du dépôt légal. Cette provenance est parfois impossible à déterminer clairement.
Soyons clairs, dans le cas des publications électroniques distantes, le dépôt légal, ne pourrait dans l'état actuel des choses remplir son rôle d'exhaustivité et d'identification. Un dépôt légal à deux vitesses, réservé aux publications légales et identifiables est envisageable, mais il ne doit pas être la priorité dans la mise en place de l'archivage du web français.

Bibliographie

Académie Universelle des cultures ; *Pourquoi se souvenir ?* - Paris : Grasset, 1998

Ackerman, Mark S. ; *Collection maintenance in the digital library : proceedings of digital librairies*. - Austin, Texas, 1995

Bachimont, Bruno. *L'archive numérique: entre authenticité et interpretabilité.* In Archives, Vol. 31, N.1, 2000. P3-15

Baddeley, Alain ; *La mémoire humaine : théorie et pratique*. - Grenoble, PUG, 1993

Bearman, David ; *Reinventing Archives for Electronics Records : Alternative Service Delivery Options*, In Margaret Hedstrom, ed. Electronic Records Management Program Strategies, Archives and Museum Informatics Technical Report, N°18 - 1993

Brodie, Nancy ; *L'archivage des publications électroniques : le rôle de la Bibliothèque nationale du Canada*. - Ottawa : Nouvelles de la bibliothèque nationale Vol.29 n°10, 1997

Cambier, Jean ; *La mémoire*. - Paris : Le cavalier bleu, 2001

Cameron, Jasmine ; *National Collection of Autralian Electronic Publications*. - Cambera : 10th national Library Technicians' conference, 1999

Cameron Jasmine, Phillips Margaret E ; *Building national collections of Internet publications*, In world librairies on the information superhighway, 1999

Chabin, Marie-anne ; *Je pense donc j'archive : l'archive dans la société de l'information*. - Paris, l'Harmattan, 1999

Chilvers Alison, John Feather. *The management of digital data : a metadata approach*. - The Electronic Library, 1998

Commission Européenne ; *A study of issues faced by National Libraries in the field of deposit collections of electronic publications*. - Luxembourg, 1995

Conseil du Trésor, sous-secrétariat à l'inforoute gouvernemental et aux ressources informationnelles ; *Conserver les documents électroniques : comment et pourquoi ?* - Collection en ingénierie documentaire, 1999

Cole Timothy W., Kazmer Michelle M. ; *SGML as a Component of the Digital Library*. - Library Hi Tech, 1995

Deleuze, Gilles ; *Foucault*. - Paris, Minuit, 1986

Favier, Jean ; *Que sais-je : Les archives*. - Paris, PUF, 2001

Kattan, Emmanuel ; *Penser le devoir de mémoire*. - Paris, PUF, 2002

Ricoeur, Paul ; *La mémoire ; l'histoire, l'oubli*. - Paris, Seuil, 2000

Roubaud, Jacques, **Bernard, Maurice** ; *Quel avenir pour la mémoire ?* - Paris, Gallimard, 1997

Tadie, Jean-Yves et Marc ; *Le sens de la mémoire*. - Paris, Gallimard, 1999

Task Force on archiving of digital data ; *Report on preserving digital data*, 1996

Signets

Digital library SunSITE Collection and Preservation Policy. Berkeley Digital Library SunSITE. 1996

http://sunsite.berkeley.edu/Admin/

AFNIC
http://www.nic.fr/

DNS*
Un document très complet sur ce qu'est un DNS et ce qu'il contient.
http://www.eisti.fr/res/res/rfc1034/1034tm.htm

Safeguarding Australia's web resources: guidelines for creators and publishers
http://www.nla.gov.au/1/scoap/guidelines.html

Le Projet Open Directory
http://dmoz.org/

Search Engine Watch
Un site explicatif et comparatif concernant les moteurs de recherches majeurs.
http://searchenginewatch.com/

The Landfiel Group
Données concernant les normes et protocoles du Web.
http://www.landfield.com/index.html