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

Diplôme de conservateur de bibliothèques

RAPPORT DE STAGE

L'archivage des sites Internet

Julien Masanès

Sous la direction de
Catherine Lupovici
Directrice du Département
de la bibliothèque numérique
de la BnF

2000

LE PROBLÈME DE L'ARCHIVAGE DES RESSOURCES NUMÉRIQUES EN LIGNE

1.1 Le développement des ressources en ligne

Une étude récente parue dans la revue scientifique *Nature*[1] proposait une estimation de la taille du web en février 1999 d'approximativement 800 millions de pages HTML contenant quelques 6 tera-octet d'information textuelle[2]. Les mêmes auteurs avaient proposé une estimation[3] pour décembre 1997 de 320 millions de pages, soit une augmentation de 150% en 14 mois.

Encore cette estimation ne tient-elle compte que de ce qui est publié au format HTML. Or un nombre important de documents mis en ligne sur Internet le sont dans des formats que l'on pourrait qualifier de formats de publication, mis en forme pour une impression comme les formats propriétaires d'Adobe, Acrobat et Poscript ou les formats de traitement de texte comme Word ou Worperfect. C'est notamment le cas des articles, rapports de recherche, cours et mémoires divers que mettent en ligne les universités et les institutions de recherche.

Internet est déjà devenu un instrument privilégié de publication pour la littérature grise car les coûts de diffusion sont extrêmement réduits, particulièrement pour des organisations qui disposent déjà du matériel nécessaire (serveur, accès au réseau) comme les universités et les laboratoires de recherche. L'IFREMER publie par exemple un certain nombre de rapports de recherche qui ne sont disponibles que sur son site Internet.

Le circuit de diffusion des articles scientifiques lui-même est en train d'être bouleversé par l'apparition des serveurs de pré-publications, dans lesquels les chercheurs versent leurs articles avant même qu'ils ne soient relus par les comités de relecture des revues scientifiques. Cela permet d'écourter le délai séparant la rédaction de la diffusion (6 mois pour un article, 3 mois pour des lettres). Certaines revues ont elle-même décidé de ne publier que sur Internet, renonçant à la publication papier et l'on peut prévoir qu'un nombre croissant de revues vont faire de même dans les années à venir.

---

[1] S. LAWRENCE, C. LEE GILES, '' Accessibility of onfirmation on the web'', dans *Nature*, vol . 400, 1999, p.107-109.

[2] La quantité d'information textuelle est obtenue en supprimant les balises des fichiers HTML. Ne reste alors que le texte de la page. Le texte étant codé en ASCII, un octet correspond à un caractère. Six tera-octets correspondent donc à six mille milliards de caractères soit une quantité de texte comparable à celle que contiennent les grandes bibliothèques nationales.

[3] S. LAWRENCE, C. LEE GILES, dans *Science*, vol. 280, 1998, p. 98-100.

L'exemple des publications scientifiques montre qu'Internet n'est pas seulement un nouveau canal pour la publication de documents qui l'étaient déjà, comme les articles. C'est aussi un moyen de publier toutes sortes de documents qui ne l'étaient pas auparavant ou alors seulement dans des limites très réduites (cours, rapports de recherche, notes etc.). L'extension du champ de ce qui est publié se double d'une réduction du filtrage qu'opérait le circuit éditorial classique.

Cela est encore plus marqué lorsque l'on observe ce qui ne fait même pas l'objet d'une mise en forme de publication dans des formats spécifiques et que l'on trouve sous la forme simple de fichiers HTML, langage de balise dominant actuellement. Ce 'tout-venant' de la publication sur Internet est extrêmement hétérogène allant de pages de navigation lourdement chargée en bannières, images et boutons de décorations à des pages contenant principalement du texte et dont certaines mesurent plusieurs centaines de Kilo-octets. Nombres de ces pages ne correspondent à aucun type de publication existant auparavant, que l'on considère leur auteur, leur contenu, leur durée de vie ou leur forme. A bien des égards, ces pages relèvent plus d'une nouvelle forme d'expression que de la publication au sens classique hérité de l'imprimé.

1.2 La responsabilité de mémoire

Doit-on pour autant considérer que cette masse imposante de textes, d'images, de sons et bientôt de vidéo, ne relève pas de la responsabilité de mémoire qui est celle des bibliothèques nationales ? Peut-on se contenter de ne chercher à archiver que les versions en ligne des publications classiques ? Poser la question c'est déjà mettre en évidence qu'une démarche qui se limiterait à de telles publications passerait à côté de cette profonde mutation de la diffusion de l'écrit que représente Internet. Une bonne part des matériaux des futurs historiens de la culture, des mœurs, de la politique, de l'économie est constituée de ces documents multiformes et éphémères. Il est absolument indispensable d'en garder une trace, au moins fragmentaire et d'en assurer la conservation à long terme.

Qui doit assumer ce rôle ? Cette conservation ne saurait être confiée à une sorte de musée des techniques, conservant les matériels informatiques et quelques échantillons des productions actuelles. Les logiciels, qui sont pourtant des outils, relèvent d'un dépôt légal informatique et font l'objet d'une conservation à la BnF. A fortiori, les ressources numériques qui, au-delà de leur forme particulière sont avant tout le véhicule d'un contenu documentaire, mériteraient-elles d'être conservées dans leur ensemble.

Les archives qui sont organisées autour d'une institution ou d'une entreprise ne semblent pas appropriées non plus devant le caractère universel de la production de ces ressources.

Certes les bibliothèques nationales ne sont habituées à traiter que le matériel publié. Elles ont habituellement à faire à des acteurs clairement identifiés et responsables de leurs publications, les éditeurs. En parallèle les imprimeurs doivent également fournir des exemplaires de tous les travaux qu'ils réalisent. Les formes de publications, bien que variées sont limitées et la publication est fixe une fois réalisée.

Rien de cela n'est valable pour Internet. Nous l'avons vu, la réduction des frais de publication[4] et la suppression du filtre éditorial permettent une explosion du volume de textes publiés alors que l'identification des acteurs se fait plus difficile voire impossible dans certains cas.

Du point de vue de la forme, la technique des langages à balise permet de constituer des documents protéiformes incluant non seulement des images mais aussi des sons, des animations, qui peuvent être composés dynamiquement avec d'autres documents. Les liens hypertextes posent différemment le problème de l'individuation d'un document. L'évolution des documents qui peuvent être modifiés à tout moment est bien sûr contradictoire avec la nécessité de conserver ces documents (quelles versions conserver, à quelle périodicité refaire une copie ?).

Tout cela est nouveau pour les bibliothèques nationales. Elles devront s'adapter sur le plan technique et organisationnel pour y faire face. Cela fait maintenant plus de cinq ans que la généralisation du protocole HTTP a permis d'impulser ce nouveau mode de diffusion du savoir et les bibliothèques nationales sont maintenant à la croisée des chemins. Certes il leur faudra s'adapter en profondeur pour prendre en charge cette nouvelle fonction mais elles sont les structures qui sont le mieux à même de le faire. D'autre part, si le basculement de l'édition a toutes les chances de se poursuivre vers le 'en ligne' et même de s'accélérer avec l'arrivée des livres électroniques, les bibliothèques nationales risquent de se voir réduites à terme à un rôle marginal si elles se cantonnent à la conservation des documents imprimés.

[4] La plupart des fournisseurs d'accès à Internet, même les fournisseurs gratuits, propose de mettre en ligne gratuitement plusieurs dizaines de Mega-octets par abonnement.

2 ETAT DE L'ART À LA BNF

L'archivage des ressources numériques en ligne est pour l'instant à la BnF à l'état de projet. La réflexion se mène dans le Département de la bibliothèque numérique.

2.1 Le Département de la bibliothèque numérique

Le Département de la bibliothèque numérique fait partie de la Direction des services et réseaux crée à la fin de l'année 1998. Cette direction regroupe plusieurs autres services transversaux comme l'Agence de bibliographie nationale, le Département du dépôt légal ou le Département des systèmes d'information. Elle est, avec la Direction des collections et la Direction de l'administration et du personnel l'une des trois directions de la Bnf.

Ce département de la bibliothèque numérique regroupe les services qui traitent des documents numériques au sein de l'établissement, à une exception près. C'est un autre service, dépendant du Département de l'Audiovisuel qui est dépositaire des collections de documents multimédias d'une part et des progiciels, banques de données et systèmes experts d'autres part. Ces documents font l'objet d'un dépôt légal[5] depuis le 1 janvier 1994 que gère la BnF.

Les trois services du Département de la bibliothèque numérique sont le service de la Numérisation dans lequel travaillent 26 équivalents temps plein, le Service de fourniture de documents à distance (une personne chargée de projet) et le service de coordination Internet dans lequel travaillent 7 équivalents temps plein. C'est à ce dernier service qu'incombe la gestion du site web de la BnF, de l'Intranet et la réflexion sur la mise en place d'un servie d'archivage des ressources électroniques en ligne. C'est à ce service que j'ai été rattaché durant mon stage.

La place du Département de la bibliothèque numérique au sein de la BnF est originale et correspond à une phase de transition durant laquelle les documents numériques ne peuvent être traités directement par les départements des collections. Durant cette phase, la nécessité d'un service transversal s'est fait sentir pour réaliser des travaux exploratoires, acquérir des compétences techniques, valider des matériels et des procédures et servir de support technique pour tous les services qui seront concernés par les documents numériques et principalement les départements des collections.

---

[5] La loi sur le dépôt légal informatique date du 20 juin 1992. Le décret d'application est du 31 décembre 1993.

L'exemple de la numérisation des collections, projet le plus avancé puisque prioritaire depuis la création de l'établissement est une illustration de ce type de transition. Le service de la numérisation a mené un important travail de développement dont un des fruits est le Poste d'Accès à la Bibliothèque Numérique. Par ailleurs le service a eu en charge, outre tous les aspects techniques de la numérisation, la constitution des collections de textes numérisés. C'est le service de la numérisation qui a réalisé la classification et les quelques 200 textes d'accompagnement proposant des cheminements de lecture, des présentations chronologiques et thématiques des textes mis en accès dans Gallica. Dans cette phase, le service de la numérisation a effectué un travail de politique documentaire, en liaison avec les départements des collections, mais sous sa responsabilité.

Les documents iconographiques ont été, eux, sélectionnés dès le départ par le département de l'audiovisuel. Le service de numérisation se contentant d'organiser les lots de production et le contrôle de la qualité.

Maintenant que les développements ont atteint une certaine maturité, des changements dans la répartition des tâches vont s'opérer. Le travail de politique documentaire va, à terme, revenir sous la responsabilité des départements des collections qui sont d'ors et déjà impliqués dans le choix des futurs collections. Le travail d'exploration et de développement est en grande partie réalisé. Reste le travail de support technique organisant la production et le contrôle de la qualité des collections, d'éventuels rafraîchissements ou des migrations de support à l'avenir.

Pour le traitement et l'archivage des documents numériques en lignes, le Département de la bibliothèque numérique commence à jouer le même rôle de service transversal et de développement. Mais dans ce cas la volonté politique semble moins nettement affirmée et il reste en partie à convaincre la direction de la nécessité d'un tel service.

Voilà ce qui a été réalisé à ce jour.

2.2 Le projet de collecte des ressources électroniques en ligne

Au cours de l'année 1998, deux informaticiens de la section du dépôt légal des documents informatiques, Yves Robert et Philippe Steuer, se sont vus confié une réflexion sur 'le dépôt légal des ressources électroniques en lignes'. Lors de la réorganisation de la bibliothèque en trois directions, ils ont été rattachés au tout nouveau Département de la bibliothèque numérique. L'orientation de leur travail a été redéfinie en abandonnant le strict cadre du dépôt légal pour mener une réflexion sur la départementalisation des procédures

de sélection des ressources, principalement celles disponibles sur le web [6]. Le niveau de granularité privilégié est celui du site, ce qui offre l'avantage de réunir un ensemble de ressources (parfois des milliers de fichiers) qui émanent dans la plupart des cas d'une même structure (université, association, institution, entreprise).

Sur le plan technique la réalisation d'une plate-forme opérationnelle de test a été achevée et a servi durant mon stage, permettant de capturer un échantillon de 23 sites.

Le bilan de ce travail de deux années a été présenté lors d'une réunion regroupant différents secteurs concernés et particulièrement des représentants des services des collections le 3 décembre 1999.

Ce travail a permis de définir les différents maillons et les différents acteurs d'une chaîne d'archivage des ressources en ligne, principalement les ressources disponibles sur le web. Les différents maillons sont la sélection, la collecte, l'archivage, la description et la communication. Les différents acteurs impliqués seraient les Départements des collections pour la sélection, la communication et la description, le Département des systèmes d'information pour la collecte, l'archivage et la communication. Le service de coordination Internet serait pour sa part concerné par la collecte et l'archivage.

Du point de vue technique, une identification précise des problèmes particuliers posés par la collecte et l'archivage de ressources en ligne a été réalisée et des outils et des procédures pour les régler ont été mis au point.

Voici les principaux.

2.2.1 Problèmes liés à la collecte

Deux modalités peuvent être envisagées pour la collecte des ressources en lignes. Soit ces ressources font l'objet d'un dépôt de la part du fournisseur (qui n'est pas forcément clairement identifié dans le cas d'un site web), que ce dépôt se fasse dans le cadre d'un dépôt légal ou dans le cadre d'une négociation individuelle. Soit la ressource fait l'objet d'une capture à distance.

Cette dernière solution, si elle permet d'envisager une collecte massive en l'absence d'un cadre juridique de dépôt légal, ne va pas sans poser des problèmes et d'abord des problèmes techniques. Il est en effet impossible de capturer des parties de certains sites. C'est notamment le cas lors que le site ne sert en fait que de passerelle vers une base de donnée, comme un catalogue de bibliothèque par exemple. Dans ce cas, la capture ne

[6] Y. ROBERT, PH. STEUER. ''Collecte des ressources électroniques en ligne : propositions de travail'', rapport interne du 9/3/99.

rapatriera que le formulaire de requête et non le contenu de la base. Ce qui peut n'être que secondaire dans le cas d'un catalogue de bibliothèque peut s'avérer très gênant dans certains cas comme dans celui des serveurs de pré-publications qui stockent leurs articles dans des bases qu'il est impossible de capturer[7]. C'est ici le contenu même qui échappe à la capture. De tels cas, relativement rares, pourraient faire l'objet de négociations avec le fournisseur.

Un autre cas est celui où le serveur fournit des pages dynamiques, c'est-à-dire des pages qui sont générées par la navigation ou les requêtes envoyées au serveur et n'existent pas à l'état de fichier fixe. Il est également impossible de capturer de telles pages, en tout cas en totalité. L'image du site qui sera rapatriée en local ne sera alors qu'une image tronquée. L'utilisation de pages dynamiques est cependant peut répandue pour l'instant et se limite aux sites commerciaux et aux moteurs de recherche.

A ces réserves près, la capture est une technique de collecte qui semble indiquée pour les ressources en ligne qui sont par définition accessibles à distance. Les outils nécessaires sont relativement nombreux et performants. Nous y reviendrons.

2.2.2 Problèmes liés à l'archivage et à la communication

Archiver un site Internet, c'est archiver un ensemble de fichiers. La plupart sont actuellement au format HTML, d'autres sont des fichiers d'images (JPEG, GIF etc.), de publication (.pdf, .doc) ou encore de sons, de vidéo etc. Tous ces fichiers sont reliés entre eux par des liens hypertextes qui assurent la navigabilité du site. Même si les contours d'une future communication de ces ressources en ligne sont encore vagues, il semble évident que le travail à l'échelle du site impliquera à un moment ou un autre d'offrir au moins la possibilité de naviguer 'en local' sur les sites archivés. Or cela ne va pas sans poser quelques problèmes. En effet un certain nombre de liens internes au site sont des liens absolus. Cela signifie que l'adresse à laquelle ils renvoient n'est pas l'adresse locale mais l'adresse d'origine du site. Dans ce cas une navigation 'locale' est impossible.

Pour résoudre ce problème, deux solutions existent.

La première consiste à modifier automatiquement tous ces liens absolus pour les transformer en liens relatifs. Cela présente l'inconvénient, outre de nécessiter un traitement assez lourd, d'altérer les pages HTML. La copie archivée ne correspond plus à l'originale.

Une autre solution consiste à monter un serveur virtuel qui déroutera les liens réels et permettra de naviguer sur le site comme si l'on était réellement sur Internet. Cela est

[7] C'est le cas par exemple du serveur de pré-publications en mathématique Grenoblewww-mathdoc.ujf-grenoble.fr.

possible sur un serveur comme Apache mais nécessite quelques manipulations techniques. Si l'on se place dans le cadre d'un archivage à long terme, ce genre de manipulations, maîtrisées aujourd'hui, pourraient poser problème à une époque où les outils que nous utilisons auront disparu depuis longtemps.

2.2.3 Conclusions du projet

Le projet permet de se faire une idée assez claire des problèmes techniques posés par la collecte et l'archivage des sites Internet. Un certain nombre a été réglé pratiquement à une petite échelle, comme celui de la navigabilité en local des sites collectés. Avec des moyens limités (quatre PC, un graveur de CD-ROM), il a été possible de capturer une trentaine de sites, de les archiver sur disques durs et pour certains sur CD. Mais vu les moyens à la fois humains et matériels limités mis en œuvre, il ne pouvait s'agir que d'une plate-forme de test à petite échelle.

3 LES OBJECTIFS DU STAGE

3.1 Etude d'un échantillon de sites

Pour permettre aux personnes concernées dans la bibliothèque d'avoir une idée plus précise sur ce qui est disponible sur Internet, de la masse documentaire que cela représente et des types de formats utilisés, Catherine Lupovici m'a demandé de réaliser une étude sur un échantillon de sites capturés grâce à la plate-forme de test mise au point au service de coordination Internet par Yves Robert et Philippe Steuer. Cette étude était conçue dès l'origine comme devant donner des éléments précis pour convaincre de la nécessité d'un service d'archivage et pour aider à la réflexion sur la démarche à adopter pour mettre en place un tel service.

3.2 Réflexion sur les métadonnées pour la préservation à long terme des documents numériques

Une réflexion a lieu actuellement à l'échelle européenne sur la préservation à long terme des documents numériques dans le cadre du projet Nedlib. Ce projet vise à produire des recommandations pour les bibliothèques numériques concernant les différents aspects de la préservation à long terme (organisation, stratégie, formats, métadonnées) et à tester un certain nombre de logiciels et de matériels pouvant réaliser, au moins en partie, les fonctions d'un archivage à long terme. Ce travail commun doit permettre d'avancer sur un certain nombre de questions que pose l'archivage des documents en ligne et de mettre en commun les expériences des participants.

La BnF est partie prenante de ce travail et est notamment responsable du WorkPackage 4 qui concerne les métadonnées nécessaires pour assurer une préservation à long terme. Catherine Lupovici qui m'avait convié à participer à la réunion de Nedlib qui se tenait à Paris la dernière semaine d'août m'a proposé de participer avec Elisabeth Freyre et elle-même à la rédaction d'un document de travail concernant les métadonnées pour la préservation à long terme.

4 ÉTUDE D'UN ECHANTILLON DE SITES WEB

4.1 Constitution de l'échantillon

Le but de l'étude n'était pas de donner une image de l'ensemble des sites mais des sites présentant un intérêt documentaire réel. Le choix a donc été effectué à partir d'une liste de sites déjà sélectionnés, la liste des signets de la BnF. Au vu des délais qu'imposait la durée de mon stage il a fallu limiter le nombre de sites sélectionnés à 23 ce qui est peu comparé aux quelques 60 000 noms de domaines répertoriés à l'AFNIC[8]. A ces 23 sites, j'ai ajouté 8 revues en ligne (voir liste en dernière page des annexes).

4.2 Les traitements effectués

Le but de cette étude était d'obtenir des données quantitatives sur les publications que contiennent ces sites, de définir une typologie des formats utilisés, d'étudier les métadonnées et le nombre de liens hypertextes contenus dans les sites.

Le point de vue adopté n'est pas un point de vue purement technique comme celui adopté par exemple pour le Web Watch Project[9]. Notre souci était de tenter de cerner quels sont les types de publications, les formats utilisés et leur contenu. Cependant, vu le nombre de fichiers (plus de 140 000 pour l'ensemble) il était hors de question de procéder à une étude manuelle de chaque fichier. J'ai donc du définir une série de traitements automatisables qu'Yves Robert a eu la patience d'implémenter sous environnement UNIX ou NT selon les cas.

Ces outils automatiques permettent de se faire une idée du contenu d'un site, au moins autant que ce qu'un simple coup d'œil sur la tranche d'un livre nous permet de savoir sur la taille du contenu de ce livre.

Or si la taille totale d'un site est une information intéressante elle n'est pas en elle-même suffisante. En effet, un site peut contenir beaucoup d'image et peu de texte ou l'inverse, il peut contenir des pages HTML très décorées mais contenant peu de texte. J'ai cherché à tenter de 'filtrer' les fichiers qui ont un contenu documentaire réel et de mettre à l'écart les fichiers contenant principalement des balises, des images ou des boutons de décoration et

---

[8] L'AFNIC est l'organisation qui répartit les noms de domaines IP en France.

[9] B. KELLY, I. PEACOCK, (page consultée le 10 décembre 1999). Web Watching UK Web Communities : Final report for the Web Watch Project, [en ligne]. URL :http://www.ukoln.ac.uk/web-focus/webwatch/reports/final/. Ce rapport donne des indications sur le profil des pages HTML, les serveurs utilisés, les fichiers d'exclusions des robots.

peu de texte. Pour cela, il a fallu enlever toutes les balises des fichiers HTML et calculer leur taille en caractère. Au-delà d'un certain seuil (plus de 5000 caractères[10]) les fichiers HTML ont été comptabilisés comme des fichiers 'riches' en texte, au même titre que les fichiers au format pdf, doc, ps, et txt.

Pour les fichiers images, principalement deux formats ont été utilisés sur les sites que nous avons capturés : gif et jpeg. Chacun correspond la plupart du temps à une utilisation spécifique : les fichiers gifs sont utilisés pour les décorations des pages HTML, le format jpeg est utilisé le plus souvent pour la diffusion d'images.

La figure 1 résume la chaîne de ces traitements. Le point de départ est constitué par l'ensemble des fichiers capturés pour un site. La liste des fichiers avec leur taille et leur type MIME permet de classer les fichiers par format et de calculer pour chaque type de format la taille totale sur le site. On saura ainsi que pour le site de l'enssib c'est 122 Mo de fichiers au format pdf qui ont été capturés.

Le traitement des fichiers HTML fut plus long. La recherche des balises de métadonnées, des liens hypertextes et finalement le nettoyage de toutes les balises pour calculer la taille en texte sont des opérations qui ont pu prendre plusieurs jours pour les plus gros sites.

4.3 Les résultats

Les résultats de cette étude font apparaître en premier lieu une forte disparité de la taille totale des sites (voir figure 2 en annexe). Les sites de l'INRIA et du Ministère de la Culture ont une taille très nettement supérieure aux autres (plus de 1.5 Go pour chacun).

Le tableau suivant, récapitulant le nombre et la taille des fichiers par format fait apparaître une nette prédominance des fichiers HTML en nombre de fichiers. Cependant, si l'on regarde les tailles, les formats de publications et notamment les formats d'Adobe (pdf et ps) se révèlent être très répandus et totaliser presque la même taille (610 Mo contre 644) de fichier que le HTML.

[10] Je suis conscient du caractère relativement arbitraire de ce seuil. Il a été déterminé après observation sur quelques dizaines d'exemples de fichiers HTML qui avait un contenu réel.

Type de fichier	HTML	PDF	DOC	RTF	TXT	PS	Images JPEG	ZIP
Nbre de fichiers des sites (et des revues)	92607 (4055)	924 (2)	263	119 (1)	171	568	16761 (2388)	23
Taille des fichiers, en Mo, des sites (et revues)	645 (27)	276 (0,8)	27	4,3	1	337	1995 (28)	2,2

Si l'on effectue le tri que nous avons décrit sur les fichiers HTML, il est possible de calculer le total des fichiers 'riches' du site. En comparant ce total avec l'ensemble des fichiers (hors images), il est possible de construire un indice (sur 10) qui indique la 'densité' des fichiers riches par rapport à l'ensemble. La taille totale du site est ainsi pondérée par la taille des fichiers images et celle de tous les fichiers ayant un faible contenu textuel (voir figure 3 en annexe). Le site de l'enssib pourtant presque 10 fois plus petit que celui du ministère de la culture s'avère avoir un contenu textuel comparable. Cela est du au fait que le site Culture contient beaucoup d'images et a un faible indice de densité, contrairement au site de l'enssib qui a un très fort indice de densité textuelle (9/10).

Il est possible à partir de ces données et des données similaires concernant les fichiers images de déterminer l'orientation d'un site, vers l'image ou au contraire vers le texte par exemple.

La figure 4 propose une représentation graphique de l'orientation des sites. Ici, le cœur de cible est constitué par les sites ayant une orientation 'image' alors qu'à la périphérie on trouve les sites contenant principalement du texte.

L'étude fait apparaître une très faible utilisation des métadonnées. Seuls 4 sites sur 23 utilisent les métadonnées HTML et un seul, celui de l'ADBS a inclus des métadonnées Dublin Core dans ses fichiers HTML. Cela pose évidemment un problème pour un traitement de ces fichiers. Vu leur nombre, il serait utile si ce n'est indispensable que les producteurs incluent eux-mêmes un certain nombre de renseignements minimaux, ne serait-ce que l'auteur et le titre du document.

Enfin le nombre moyen de lien par page HTML est une information importante qui permet d'estimer l'information 'd'aiguillage' que peut renfermer le site (voir figure 5 en annexe).

5 LES MÉTADONNÉES POUR LA PRÉSERVATION À LONG TERME

Nous avons vu que sur l'échantillon de sites de notre étude, peu de responsables de site incluent des métadonnées, même les plus élémentaires (comme l'auteur, le titre du document et des mots-clés facilement implémentable avec HTML).

Cela interdit presque pour une bibliothèque numérique de travailler à l'échelle du fichier car les fichiers sont, nous l'avons vu très nombreux et leur traitement individuel serait très lourd (rien que les 23 sites de notre échantillon comprennent presque trois fois plus de fichiers que le dépôt légal ne reçoit de monographies chaque année).

L'arrivée du langage de balise XML réglera peut-être ce problème en perfectionnant le système des métadonnées et en permettant qu'il fasse l'objet d'un traitement automatique. Mais cela dépendra toujours de la volonté des auteurs en dernière instance. Le fait que des moteurs travaillant sur les métadonnées voient le jour pour proposer un référencement de l'information plus précis que ce que permettent les moteurs d'indexation actuels, sera sans doute une incitation forte pour les auteurs qui voudront voir leurs pages référencées.

Si l'on se place maintenant dans la perspective d'un archivage à long terme, d'autres informations seront nécessaires pour permettre d'en garantir l'accès dans 20 ou100 ans. A ce moment, l'équipement informatique aura profondément changé, les logiciels et les formats qui nous sont familiers aujourd'hui auront disparus depuis longtemps. Les supports sur lesquels les données auront été sauvées ne seront peut-être même plus lisibles tout simplement faute de lecteur en état de marche.

Cela pose une série de problèmes à la communauté des bibliothèques auquel le projet Nedlib, entre autre, tente de faire face. Pour mener une réflexion globale et en coordination avec les autres professionnels confrontés à ces problèmes, notamment ceux chargés de la conservation des données scientifiques, l'équipe de Nedlib a décidé d'adopter le modèle fonctionnel OAIS (Open Archival Information System) proposé par le Consultative Committee for Space Data Systems[11]. Ce modèle permet de décrire l'organisation fonctionnelle d'un réservoir de documents numériques et s'applique donc aux bibliothèques moyennant un certain nombre d'aménagements en cours de discussion.

A la base du problème de l'archivage à long terme se trouve le fait que l'information numérique ne soit pas accessible directement. Elle ne peut l'être que grâce à une série de

[11] CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS, (page consultée le 10 décembre 1999). Reference Model for an Open Archival information System (OAIS), [en ligne]. URL :ftp://nssdc.gsfc.nasa.gov/pub/sfdu/isoas/us12/CCSDS-650.0-W-3.pdf.

dispositifs matériels et logiciels dont l'utilisateur ne soupçonne souvent pas la complexité ni même parfois l'existence. Or ces dispositifs évoluent, disparaissent relativement rapidement et si l'on y prend garde, un document numérique devient vite inaccessible.

La première des choses à faire est de repérer toutes ces dépendances de la manière la plus systématique possible. Il m'a semblé que pour cela l'utilisation du modèle en couche d'information était le plus approprié. Ce modèle est proposé en annexe[12] du modèle OAIS pour déterminer l'information nécessaire à l'utilisation future des logiciels.

Dans ce cadre, on peut considérer que ce problème de base revient à assurer la transition entre la couche binaire et l'utilisateur final. Cela se fait au travers du passage entre différentes couches d'information (couche physique, couche binaire, couche structurée couche objet et couche application). A chacun de ces passages, des dispositifs matériels ou logiciels sont requis, avec des spécifications de standards particuliers.

Un premier travail pour une bibliothèque numérique chargée de l'archivage à long terme sera de repérer tous ces passages et de constituer pour chacun une métadonnée qui sera jointe au fichier. On ne peut attendre des auteurs qu'ils se chargent de fournir de telles métadonnées dont l'intérêt ne leur semblera sans doute pas évident. Mais cela ne devrait pas poser de problème dans la mesure où ces métadonnées seront, pour les documents en ligne au moins, relativement facile à générer automatiquement.

A partir de ces métadonnées qui repèrent un certain nombre de standards, la bibliothèque numérique doit assurer la disponibilité des matériels et logiciels qui permettent de les utiliser, leur émulation ou la migration des fichiers dans d'autres standards. Une bibliothèque numérique d'archivage à long terme se doit donc d'assurer une cohérence globale (support/dispositif de lecture/ fichiers/logiciels) ce qui représente un travail nouveau, nécessite des informations et des outils nouveaux. Les métadonnées ainsi repérées sont le point de départ qui permet d'organiser cette cohérence et représentent donc une information cruciale qui doit être soigneusement recueillie.

Cette démarche a été validée lors de la réunion du groupe de travail qui s'est tenue à Berne les 6 et 7 décembre et le rapport sera rédigé pour le mois de janvier.

[12] Op. cit., p.129.

6 CONCLUSION

L'archivage des documents numériques en ligne pose une série de problèmes que le Département de la bibliothèque numérique est chargé d'identifier pour qu'une réflexion soit menée au niveau de l'ensemble de l'établissement et particulièrement des départements des collections. Il faut à la fois convaincre de la nécessité de prendre ce travail en charge et explorer les outils et les procédures qui permettraient de le faire.

Dans ce cadre, mon stage a consisté à fournir des éléments quantitatifs et une typologie documentaire du contenu de sites Web pour aider à la réflexion sur le niveau de granularité auquel travailler et sur le type de fichiers qu'un service d'archivage serait amené à traiter.

Parallèlement, j'ai été associé à la réflexion sur les métadonnées sur la préservation à long terme des documents numériques, métadonnées dont la création sera un élément important pour permettre un accès futur à ces documents.

Ces métadonnées comme les formats de publication étudiés sur l'échantillon de sites web, sont des éléments nouveaux que les bibliothèques ne sont pas habituées à traiter. Les choix qu'elles seront amener à faire concernant ces nouveaux éléments (niveau de granularité, choix des fichiers, constitution des métadonnées) seront important pour la conservation de ces documents, et même cruciaux pour les établissements chargés de leur conservation à long terme.

En France, c'est naturellement la BnF à qui incombe cette responsabilité. Comme nous l'avons vu, c'est à un département transversal, qui a déjà jouer un rôle exploratoire et de développement pour la constitution des collections numérisées que cette tâche a été confiée. Reste que le niveau de priorité accordé aux collections numérisées en son temps n'est pas encore accordé à l'archivage des documents en ligne. Une partie de travail du Département de la bibliothèque numérique consiste donc encore à convaincre de la nécessité de mettre les moyens nécessaires pour faire face à cette tâche.

7 ANNEXES

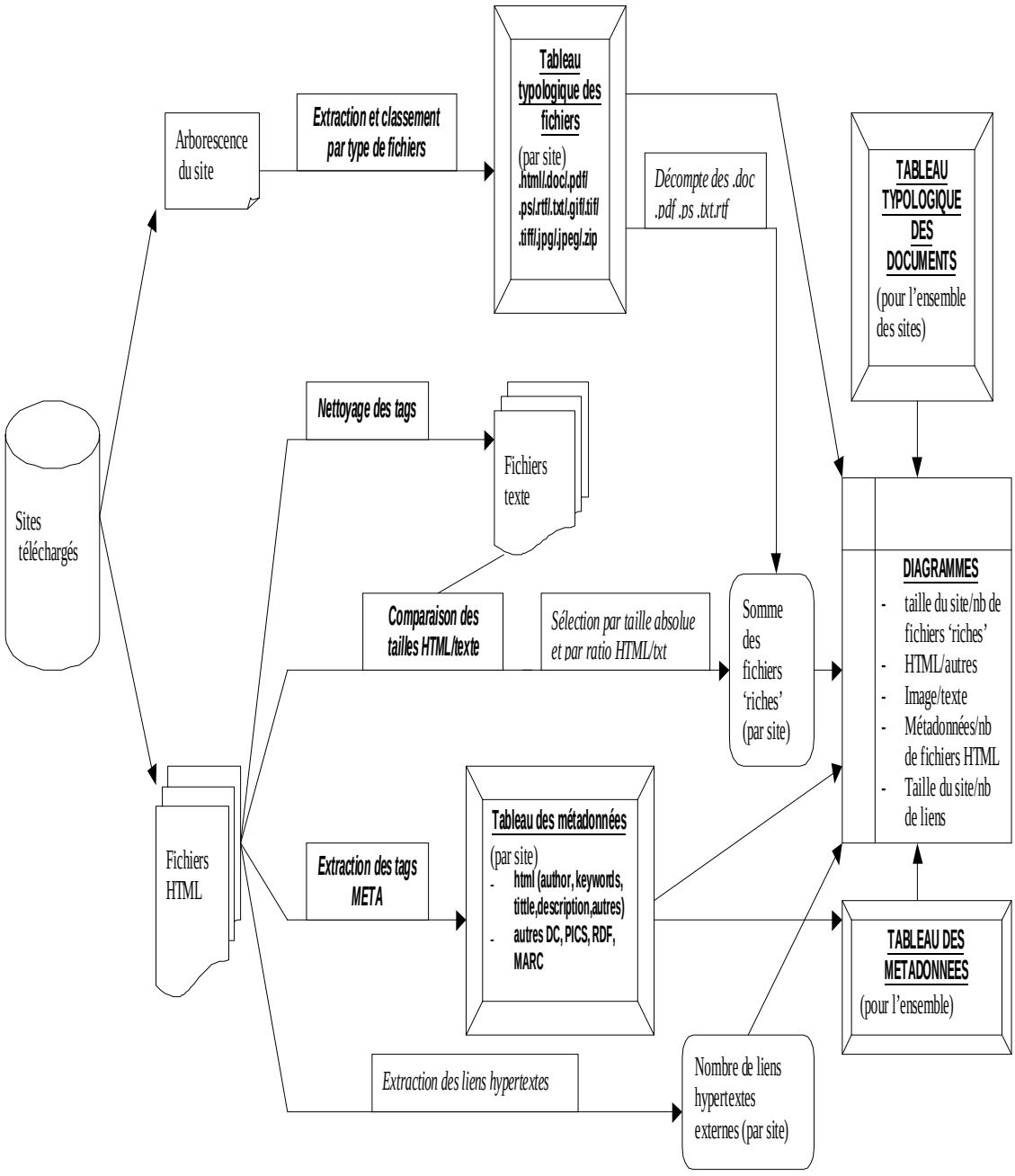

Figure 1 : schéma du traitement des sites

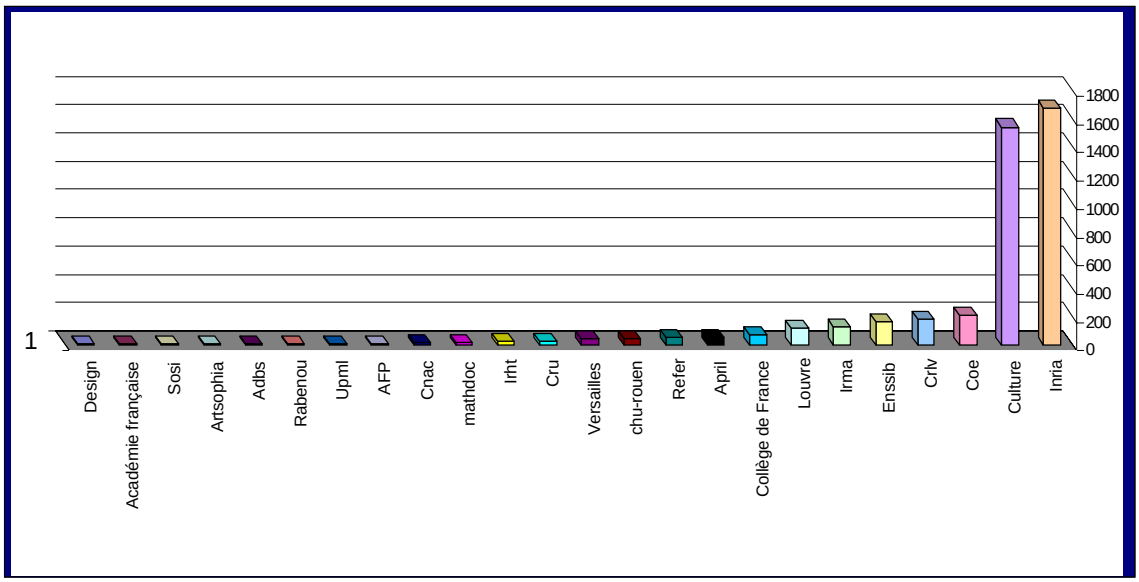

Figure 2 : taille des sites en Mo

NOM DU SITE	Taille texte en Mo	Indice Texte	Taille totale
Inria	582	5	1670
Enssib	133	9	164
Culture	99	3	1540
Coe	92	4	211
Irma	54	5	126
chu-rouen	23	8	43
mathdoc	13	8	18
Refer	12	4	51
Rabenou	7	9	8
April	7	5	54
Adbs	5	8	8
Upml	5	7	8
Irht	4	3	21
Versailles	3	2	42
Louvre	3	1	114
Design	2	9	3
Crlv	2	1	182
Cnac	2	3	12
Collège de France	1	1	75
Académie française	1	6	3
AFP	1	2	9
Cru	1	1	22
Sosi	1	3	5
Artsophia	1	3	6

Figure 3 : tableau indiquant la taille des fichiers texte riches (en Mo) et l'indice de densité textuelle (sur 10)

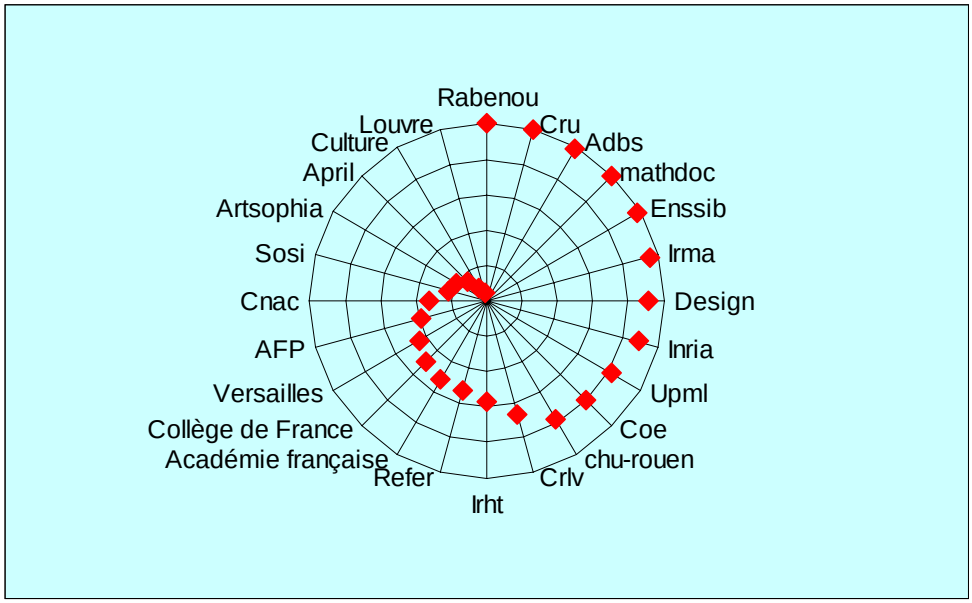

Figure 4 : représentation graphique de l'orientation (Image/texte) des sites. Ici le coeur de cible est constitué par les sites orientés image.

NOM DU SITE	Nombre de liens par page HTML
AFP	321,5
chu-rouen	12,3
Adbs	9,7
Cru	7,4
Coe	5,5
Irma	4,9
Rabenou	4,2
Versailles	3,6
Enssib	2,3
Design	2,0
Inria	1,9
April	1,8
Sosi	1,0
Refer	0,9
Artsophia	0,9
Upml	0,8
Irht	0,6
Cnac	0,3
Académie française	0,3
Culture	0,2
Louvre	0,0
Collège de France	0,0
Crlv	0,0

Figure 5 : nombre de liens hypertextes par site.

Liste des sites et de leur URL

Académie française	www.academie-francaise.org
Adbs	www.adbs.fr
AFP	www.afp.com
April	www.april.org
Artsophia	www.artsophia.com
chu-rouen	www.chu-rouen.fr
Cnac	www.cnac-gp.fr/musee
Coe	culture.coe.fr
Collège de France	www.college-de-France.fr
Crlv	www.crlv.org
Cru	listes.cru.fr
Culture	www.culture.fr
Design	www.design.fr
Enssib	www.enssib.fr
Inria	www.inria.fr
Irht	irht.cnrs-orleans.fr
Irma	www-irma.u-strasbg.fr
Louvre	www.louvre.fr
mathdoc	Grenoblewww-mathdoc.ujf-grenoble.fr
Rabenou	www.rabenou.org
Refer	www.refer.org
Sosi	www.sosi.cnrs.fr/AFRHC/
Upml	www.upml.fr
Versailles	www.chateauversailles.fr

Liste des revues en ligne et de leur URL

artmag	www.artmag.com
lmda	www.lmda.net
livresse	perso.wanadoo.fr/livresse
acores	www.micronet.fr/~acores
anacoluthe	www.anacoluthe.com
webmatin	www.webmatin.com
zazieweb	www.zazieweb.com
Solaris	www.info.unicaen.fr/bnum/jelec/Solaris