Vers la conservation des sites web régionaux

Ce sont souvent les objets les plus courants, ceux dont la banalité est si manifeste qu'ils en deviennent invisibles, qui, sur le long terme, deviennent rares et précieux : les tracts, l'exemplaire du journal… Les sites web font partie de ce type de document : communs et éphémères.

Magali Haettiger

haettiger@aol.com

Cette comparaison entre les sites web et la presse est loin d'être aussi fortuite qu'il n'y paraît à première vue. Le développement de la presse, dans la seconde moitié du XIXe siècle, a permis à de nombreux groupes plus ou moins formalisés et plus ou moins importants de s'exprimer. Aujourd'hui, ces groupes, ces associations diffusent souvent leurs idées par le Net, moins coûteux que l'imprimé. Comme pour les tracts, le web est aussi le vecteur d'opinions et de prises de position. Comme l'exemplaire du journal qui une fois lu est généralement jeté, le site web n'a pas pour but d'être conservé. Un site web évolue, change de localisation, disparaît et n'a finalement de sens pour l'internaute que le temps de sa consultation. La presse n'est pas produite pour être conservée, mais nul n'aurait l'idée de regretter que les bibliothèques aient détourné ce principe et inscrit parmi leurs missions sa conservation. Mais, alors que les bibliothèques françaises peuvent s'enorgueillir de leurs collections de quotidiens, notamment locaux, le web local demeure absent de leurs missions de conservation.

Pour le moment, en France, seule la Bibliothèque nationale de France s'occupe de l'archivage de sites web dans l'objectif de mettre en place, à court terme, un véritable dépôt légal du web français. Il semble donc que les sites web régionaux ne seront pas véritablement mis en valeur en tant que tels : *a priori,* les sites webs rhônalpin, breton ou picard n'existeront qu'en tant que partie du web national.

La question de la conservation du web régional s'impose donc. Cet article n'a pas pour but de traiter précisément tous les problèmes que pose l'archivage de sites. Il n'a pas non plus pour objectif de présenter un projet universel, valable pour toutes les BMVR (bibliothèques municipales à vocation régionale) et bibliothèques municipales. Il s'agit plutôt d'évaluer dans quelle mesure la conservation et la valorisation des sites web d'intérêt régional peuvent être envisagées à une échelle locale compte tenu des problèmes et difficultés qu'engendre ce type de projet.

L'acquisition de sites web

Deux aspects de l'archivage de sites web sont aujourd'hui hautement problématiques : l'acquisition des sites et leur conservation proprement dite.

L'acquisition des sites web n'apparaît pas à première vue comme l'action la plus complexe du processus de conservation. Les sites web sont avant tout des documents numériques, ce qui leur confère une reproductibilité infinie et facilite les procédures de capture, lesquelles peuvent être opérées à distance. En outre, les sites web sont souvent gratuits.

* Cet article est tiré du mémoire de DCB, « L'archivage des sites web d'intérêt régional » (direction : Élisabeth Noël), soutenu en 2003.

***Magali Haettiger**, titulaire d'une maîtrise d'histoire et d'un DEA d'ethnologie, est actuellement élève-conservateur à l'Enssib.*

Le site web : un objet complexe

L'acquisition de sites web, dans le cadre d'un projet d'archivage, implique l'élaboration et l'explicitation d'une véritable politique de sélection des sites. Ces critères de sélection sont parfois des plus classiques (le contenu, l'auteur, la localisation…). D'autres, également nécessaires, sont spécifiques au cas précis du web.

Il faut dire qu'en tant que document, le site web est un objet particulièrement complexe. En premier lieu, c'est un objet instable qui évolue, connaît des versions successives. De ce fait, même si une majorité d'entre eux ne subit aucune modification, on ne saurait considérer un site comme un objet achevé. L'instabilité du web est particulièrement frappante dans le cas des sites web dynamiques, dont la forme évolue en fonction de l'internaute. Le caractère mouvant d'un site web nous interroge sur la définition même du site web en tant qu'objet à conserver. En effet, une politique de conservation de ces sites ne devrait pas se contenter de la capture et de l'archivage d'une ou quelques versions d'un même site, mais bel et bien de toutes les versions successives de celui-ci. Or, avec le Net, nous nous trouvons dans un contexte d'autopublication. À ce titre, les modifications d'un site ne sont que rarement signalées à l'internaute. Le suivi de chaque version d'un même site web s'apparente alors à un lourd travail de veille documentaire.

Un site est également un objet éphémère : d'après une étude de l'OCLC, la durée de vie moyenne d'un site n'excéderait pas six semaines. Quant à l'authentification des sites, elle s'avère parfois fastidieuse. En effet, si l'auteur n'est pas toujours clairement identifié, il n'existe, par ailleurs, aucun système d'identification unique pour les sites web, qui serait un équivalent de l'ISBN. Il s'avère donc parfois difficile d'identifier rapidement que le site web que l'on consulte est bien la nouvelle version d'un autre site ayant changé de localisation (URL).

Le site web est donc un document mouvant, éphémère et faiblement « normé » : autant de caractéristiques dont il faut tenir compte pour une politique d'acquisition. Notons que ces difficultés sont déjà ressenties par de nombreuses bibliothèques et ce, en dehors de tout projet de conservation. Les annuaires de sites et les signets font de plus en plus partie des outils et services que proposent les établissements. Le développement de ces outils nécessite l'élaboration de critères de sélection qui expose déjà les bibliothécaires aux difficultés de l'évaluation des sites et de la veille documentaire.

> Le site web est un document mouvant, éphémère et faiblement « normé »

Outre la prise en considération des caractéristiques générales du web (mouvant, éphémère et peu normé), l'archivage de sites passe obligatoirement par une définition plus fondamentale de l'objet site web lui-même. Le web recouvre en fait plusieurs objets bien différents. À ce titre, qu'est-ce qui, parmi les informations circulant sur le Net, doit être considéré comme un site web ? Prenons l'exemple d'un site se prolongeant sur un forum de discussion : l'objet « site web » que l'on souhaite archiver intègre-t-il ce forum de discussion ou doit-on fixer sa limite en amont ? La plupart des sites web établissent des liens externes vers d'autres sites. Cette ouverture du site vers d'autres fait partie intégrante des fonctionnalités de l'Internet. À ce titre, un site web n'est jamais seulement un élément du Net puisqu'il participe de sa navigabilité. À la fois simple élément du Net, il en est également un vecteur : espace et chemin. Le choix de ces liens hypertextuels externes fait partie du site en question et de la démarche de son auteur. De ce fait, il semblerait logique de conserver à la fois le site lui-même et l'ensemble des sites vers lesquels il renvoie. Or la pertinence de ces sites « externes » en regard des critères de sélection de la bibliothèque n'est pas toujours attestée : faut-il alors conserver l'ensemble de ces sites ? Ne considérer que le site lui-même et non les liens vers lesquels il pointe ?

Enfin, un site web ne se définit pas uniquement par son contenu informationnel. La forme du site, les choix esthétiques opérés par son créateur sont également des données importantes à conserver sur le long terme dans la mesure où ces caractéristiques pourront être le reflet d'une norme esthétique, ou au contraire d'une certaine originalité. Les fonctionnalités de recherche qu'offrent certains sites web peuvent également être pertinentes à conserver. L'objectif de tout archivage de site consisterait donc à conserver toute l'intégrité du site – contenu, forme et fonctionnalités –, ce qui nécessite d'intervenir sur toutes les couches techniques du document numérique site web et, donc, à un haut niveau d'abstraction et d'exigence.

Une politique d'acquisition

La mise en place d'un projet de conservation de sites web induit l'énonciation, au préalable, d'une véritable politique d'acquisition. Cette politique s'articule essentiellement sur deux niveaux.

Il s'agit d'abord de circonscrire la part du web que l'on souhaite conserver. Pour la plupart des projets développés à l'heure actuelle, le principal critère de sélection de sites est d'ordre territorial, puisqu'il s'agit de mettre en place un dépôt légal du web national. Or, le réseau échappe à

la notion de territorialité. Qu'est-ce qu'un web national ? L'ensemble des sites hébergés sur des serveurs localisés sur le territoire national ? L'ensemble des sites dont le nom de domaine comprend une indication nationale (.fr par exemple) ? Les sites qui, quelle que soit leur localisation « physique », traitent du territoire national ? Les sites web dont l'auteur est un ressortissant du pays ? Dans ce cas, comment en être certain ? Les points de vue sur cette question diffèrent d'un pays à l'autre : alors que la Bibliothèque nationale d'Australie conserve les sites web qui traitent du pays et sont créés par un Australien, la Suède archive également des sites sur les pays considérés comme étrangers. Cette délimitation territoriale du web est davantage problématique dans le cas d'une circonscription régionale du web. En effet, s'il est ardu de définir la part nationale du web, cette tâche est encore plus difficile lorsque l'on souhaite délimiter le web régional, ne serait-ce que parce qu'il n'existe pas de nom de domaine régional qui permettrait d'identifier un socle minimal pertinent de sites à archiver.

Ce premier niveau, très général, de délimitation correspond somme toute à des critères classiques dans une politique d'acquisition (contenu, auteur, localisation…). Au second niveau, il s'agit de circonscrire l'objet « site web » en soi, c'est-à-dire donner une définition claire de l'unité de conservation que l'on va traiter : va-t-on conserver les sites vers lesquels renvoie le site web que l'on souhaite conserver ? À partir de quel moment considère-t-on que l'on a affaire à une nouvelle version d'un site [1] ? Combien de captures de sites espère-t-on effectuer par an ? Considère-t-on les *newsgroups* et les forums comme des sites web ? À quel niveau d'exigence et de complexité technique veut-on se placer pour ce projet ?…

Les procédures d'acquisition

Après la politique d'acquisition, se pose alors le problème de l'acquisition proprement dite des sites. Il existe deux grands types de procédures d'acquisition de sites web.

L'acquisition manuelle

La sélection manuelle et le dépôt représentent les deux grands types de procédures manuelles d'acquisition de sites.

La sélection manuelle est employée par les bibliothèques nationales d'Australie et du Canada. Dans ce cas de figure, ce sont les bibliothécaires qui sélectionnent les sites pertinents dont la capture est ensuite effectuée à l'aide d'un logiciel.

Le dépôt consiste à demander ou contraindre les créateurs d'envoyer une version de leurs sites web. Cette procédure ne saurait être utilisée comme seul moyen d'acquisition. Tout d'abord parce que, nous l'avons vu, Internet est un espace d'auto-édition ce qui démultiplie le nombre « d'éditeurs » et donc d'interlocuteurs potentiels de la bibliothèque. D'ailleurs, le dépôt n'est utilisé que par la Bibliothèque nationale de France, en complément de ses procédures d'acquisition automatiques et pour des sites bien particuliers.

Les acquisitions manuelles sont bien entendu relativement coûteuses en personnel.

L'acquisition automatique

Un site web, nous l'avons dit, est avant tout un document électronique et, de ce fait, son traitement automatisé est envisageable. La sélection et l'acquisition automatique de sites web sont alors effectuées par un robot formé d'un *crawler* qui va parcourir le web et rapatrier dans une base de données les URL des sites pertinents à archiver et d'un *harvester* ou logiciel de collecte qui va acquérir physiquement les sites correspondant aux URL sélectionnées. Ce système permet d'acquérir une masse conséquente de sites web et de conserver, pour l'usager, la navigabilité du web archivé. Le produit de cette collecte ou *snapshot* peut être comparé à une véritable photographie du web. En outre, contrairement à l'acquisition manuelle, les critères de sélection y sont moins discriminants et, de ce fait, on aboutit à un web archivé beaucoup plus représentatif du web de l'époque.

> L'acquisition automatique de sites est une opération techniquement lourde qui sollicite énormément les réseaux

Toutefois, l'acquisition automatique de sites comporte également plusieurs limites majeures. Tout d'abord, l'une des grandes difficultés de ce type d'acquisition est de traduire en algorithmes de pertinence les critères de sélection de la politique d'acquisition. Par ailleurs, l'acquisition automatique ne permet pas d'obtenir les sites web invisibles ou *deep web*. Le web invisible recouvre l'ensemble des sites qui ne peuvent

1. La question du repérage d'une nouvelle version d'un site est relativement problématique. Tout d'abord, comme nous l'avons vu, la nouvelle version d'un site peut avoir changé d'adresse URL par rapport à la version précédente : dans ce cas, et en l'absence d'un identifiant unique, on aura tendance à considérer que cette nouvelle version est en fait un nouveau site et le lien entre versions ne sera pas restitué et donc ne sera pas transparent pour l'usager. On peut, par ailleurs, envisager qu'un repérage automatique des changements opérés sur un site puisse être effectué par un outil de veille. Cependant, il ne faut pas oublier que d'un point de vue informatique la modification d'un seul caractère (une virgule, par exemple) constitue une nouvelle version. Or ce type de changement sur un seul caractère ne devrait pas normalement donner lieu à une nouvelle capture du site en tant que nouvelle version. Là encore, dans le cas d'un emploi de ce type d'outil de veille, il s'agira de paramétrer correctement celui-ci et donc de définir clairement un seuil à partir duquel on considère qu'une modification est considérée comme suffisamment importante pour donner lieu à une nouvelle capture du site en tant que nouvelle version.

Les problèmes techniques de la conservation du web

Un site web est avant tout un document numérique et, à ce titre, pose des problèmes de conservation. Ces problèmes sont de deux types : la conservation des supports et celui de l'obsolescence technique.

La durée de vie des supports

La capacité de stockage des supports électroniques est de plus en plus importante : le CD commercialisé dès 1983 avait une capacité de stockage de 680 Mo ; le DVD, entré sur le marché en 1995, peut contenir 4,7 Go. Inversement, alors que cette capacité de stockage tend à s'accroître, la durée de vie de ces mêmes supports est des plus limitée. Il est toutefois très difficile d'obtenir dans ce domaine des données fiables. Alors que les constructeurs annoncent pour leurs CD une longévité de 75 à 200 ans, les laboratoires de recherche indépendants annoncent une durée de vie moyenne, pour ce même support, de 5 à 25 ans. Loin de l'optimisme des constructeurs, la durée de vie des supports électroniques ne dépasse que rarement 10 ans.
Par ailleurs, comme pour tout type de support, les conditions environnementales de conservation sont importantes à prendre en considération. Selon une étude de la Digital Preservation Coalition*, un cédérom conservé à une température de 10 °C peut obtenir une espérance de vie de 30 ans. Ce même CD, conservé dans un environnement de 28 °C, voit cette même espérance limitée à 3 mois maximum.

L'obsolescence technique

Contrairement au livre, pour lequel le support et le contenu sont intrinsèquement liés, la structure d'un document numérique comporte plusieurs niveaux qui induisent une séparation entre le support et le contenu informationnel. Chaque niveau nécessite, pour être traité par la machine et accessible par l'utilisateur, la médiation d'équipements mais aussi de logiciels. Ces médiations successives conditionnent la consultation du document numérique par l'utilisateur. Cette dépendance technique qui constitue pour celui-ci une contrainte parfois importante, devient une difficulté majeure dans le cadre d'une action de conservation de ce type de document. En effet, on estime que le cycle de validité des logiciels et des périphériques est de l'ordre de 2 à 5 ans. La compatibilité ascendante entre les versions successives d'un logiciel n'est pas toujours assurée.
On peut alors tout à fait imaginer que l'on ait archivé un site web sur un support en bon état et dans d'excellentes conditions de conservation et être dans l'impossibilité d'y accéder et de le consulter à cause d'un format obsolète.

Les solutions techniques de conservation

Pour le moment, la migration apparaît comme le seul moyen d'assurer la conservation de documents numériques sur le long terme. D'un point de vue général, la migration consiste à effectuer une transformation plus ou moins importante du document à conserver, en suivant l'évolution des techniques. Il est possible de distinguer globalement deux types de migrations :
• La migration de support qui consiste à copier le document sur un nouveau support de stockage. Ce type de migration ne présente pas de difficulté majeure dans la mesure où le document lui-même n'est pas véritablement transformé, le train de bits n'étant pas altéré.
• Celle qui consiste à modifier le format ou le codage de données du document. Il s'agirait donc, dès qu'un format est amené à disparaître, de convertir le fichier en danger d'obsolescence dans un nouveau format cible. Cette forme de migration est plus complexe puisqu'elle atteint les couches plus abstraites du document. Cette transformation peut, le cas échéant, entraîner des modifications plus ou moins profondes au niveau des fonctionnalités offertes par le site web, au niveau de sa forme mais aussi au niveau du contenu du site. La migration doit se concevoir comme une véritable gestion des risques, l'objectif étant de choisir un format cible (si possible standard) qui portera peu atteinte au document de départ. La migration nécessite donc un excellent niveau d'information sur l'évolution des formats et des renseignements techniques sur les documents d'une grande précision (les métadonnées). L'information technique devient alors, dans le cadre des migrations, un enjeu vital pour la pérennité des documents.

La conservation des sites web et de tout document électronique oscille entre deux risques majeurs : d'une part le risque de perdre définitivement la possibilité d'accéder à un fonds si l'on ne lutte pas contre le risque d'obsolescence technique et, d'autre part, le risque de transformer irrémédiablement et profondément le document que l'on souhaite conserver.
Si la migration est le seul moyen actuel d'assurer la pérennité d'un document électronique, d'autres solutions sont en cours d'expérimentation, comme par exemple l'émulation. Un émulateur est un dispositif logiciel permettant d'exécuter sur un certain type d'ordinateur des instructions écrites pour un autre type d'ordinateur. Il serait donc imaginable d'accéder aux sites web dans leur forme originale par le biais d'un émulateur capable d'exécuter des instructions écrites dans des langages obsolètes, pour du matériel obsolète. Cette solution offrirait l'avantage de conserver l'intégrité de l'objet archivé et de permettre, sur le long terme, d'accéder à toutes les composantes d'origine de celui-ci (forme, contenu et fonctionnalités). L'émulation ne doit pas apparaître comme une solution à court terme mais véritablement comme une perspective intéressante pour l'archivage électronique.

* Digital Preservation Coalition. Media and formats. *Digital Preservation coalition's website* (en ligne) : http://www.dpconline.org/graphics/medfor/media.html [Dernière consultation le 16 mars 2003]

être traités par des moteurs de recherche. La proportion du web invisible par rapport au web global est très difficile à évaluer. Certains n'hésitent pas à dire qu'il représenterait 40 % du web, ce qui paraît largement surévalué. Les bases de données en font partie. Or certaines de ces bases offrent un contenu informationnel sans équivalent. Seule la Bibliothèque nationale de France s'est lancée dans l'archivage de ces bases de données en complétant le mode opératoire d'acquisition automatique par une procédure de dépôt manuel des bases de données non acquises par le robot.

Enfin, l'acquisition automatique de sites est une opération techniquement lourde qui sollicite énormément les réseaux. Ainsi, le nombre de *snapshots* que peut effectuer un établissement par an est relativement limité. À titre d'exemple, la Bibliothèque royale de Suède a effectué

deux *snapshots* par an depuis 1997. De ce fait, il est impossible d'envisager la capture de toutes les versions de chaque site acquis. Sans parler de saupoudrage, ce mode d'acquisition est plus extensif qu'intensif puisqu'il permet d'obtenir une surface importante du web pertinent mais sans réelle profondeur.

D'une façon ou d'une autre, que le choix se porte sur l'acquisition manuelle ou sur l'acquisition automatique même complétée par une procédure de dépôt, l'exhaustivité, du moins pour le moment, n'est pas pensable au niveau de l'Internet.

Outre les difficultés liées à l'acquisition des sites web, la conservation proprement dite de ceux-ci, à l'instar de n'importe quel document électronique, est problématique. En effet, l'archivage des sites web s'expose au double problème de la fragilité des supports électroniques d'archivage et de l'obsolescence technique (voir encadré page précédente). Mais, l'archivage des sites web ne se heurte pas uniquement à des obstacles d'ordre technique : la question juridique y est également centrale (voir encadré ci-contre). Enfin, un projet de conservation de sites ne peut se faire sans un investissement financier minimal en matériel mais également en personnel. Dans un contexte général de réduction des ressources des bibliothèques, la mise en place d'une politique d'archivage de sites web régionaux n'est-elle pas prématurée ?

Vers la conservation du web régional ?

La réflexion qui s'engage aujourd'hui sur l'archivage des sites web dépasse le cadre de la Bibliothèque nationale de France et commence à devenir un sujet d'interrogations au sein de grandes bibliothèques municipales.

Il faut dire que la conservation de sites web présentant un intérêt régional serait intéressante à plus d'un titre. Avant tout, nombreux sont les sites web régionaux qui présentent un véritable intérêt : à court terme parce qu'ils comportent un contenu informationnel sur la région, mais également à long terme parce que, de par leur contenu, mais aussi leur forme, ils sont le reflet d'idées, de mentalités et de normes formelles. La création d'un site web est un moyen relativement peu coûteux de diffuser une information. De ce fait, elle permet à un village ou une association n'ayant pas les moyens de produire un bulletin municipal ou associatif de s'exprimer. Se développe ainsi sur Internet toute une documentation qui n'a pas d'équivalent sur support papier.

La réflexion qui s'engage aujourd'hui sur l'archivage des sites web dépasse le cadre de la Bibliothèque nationale de France et commence à devenir un sujet d'interrogations au sein de grandes bibliothèques municipales

Pour le moment, les bibliothèques proposent le plus souvent une sélection de sites web sous la forme d'annuaires ou de signets. En l'absence d'une politique de conservation, elles se retrouvent en situation d'incomplétude, comme si elles proposaient des documents en libre accès qu'elles jetaient au fur et à mesure. Cette situation est inconfortable pour ces établissements dans la mesure où ils n'assument pas, du point de vue d'Internet, l'ensemble de leurs missions, mais également parce qu'ils ne maîtrisent pas véritablement l'objet qu'ils mettent à la disposition du public. La question du public est évidemment centrale ici. Cependant, étant donné que l'on se place du point de vue de la conservation, l'intérêt que représentent ces sites ne se conjugue pas au présent et nul ne saurait affirmer ce que le public de demain recherchera dans les sites web conservés,

Les questions juridiques liées à l'archivage du web

D'un point de vue général, l'archivage des sites web français s'effectue à la BnF dans l'esprit d'un dépôt légal du web mais en l'absence d'une loi. L'article 10 du projet de loi sur la « Société de l'information », présentée en Conseil des ministres le 13 juin 2001, avait pour but de compléter la loi du 20 juin 1992 sur le dépôt légal. Dans le cadre de ce projet de loi, seule la BnF aurait la responsabilité de la collecte et de la conservation.

Cette loi clarifierait la question de l'acquisition des sites web et permettrait de légaliser l'obligation de dépôt pour les bases de données et les sites web invisibles. Toutefois, ce projet ne résout pas de nombreuses difficultés légales intimement liées aux problèmes techniques que pose la conservation des sites.

En effet, la mise en œuvre d'un plan de conservation des sites peut entraîner la modification de ceux-ci. Dans le cas d'une migration, un établissement peut être amené à transformer le format de certains fichiers, risquant ainsi de porter atteinte à l'intégrité du site. Or, notamment depuis un jugement du 9 février 1998 du tribunal de commerce de Paris, des pages web se sont vu attribuer la qualité d'œuvre protégeable au titre des droits d'auteur. Sachant que les droits moraux protègent l'auteur et son œuvre de toute dénaturation, les procédures de conservation mises en place par la bibliothèque et la transformation des sites (migrations) ne peuvent-elles alors être considérées comme une atteinte aux droits d'auteur ?

De la même façon, la BnF, grâce à une procédure de dépôt, va conserver des sites web invisibles. Or, l'accès à ces sites web est souvent restreint, protégé par un mot de passe et même parfois payant. Ces sites seront donc intégrés dans l'ensemble des archives du web français mais, pour autant, ils seront encore soumis à certaines restrictions d'accès et de divulgation liées au droit d'auteur. Ce faisant, la BnF n'aura peut-être pas la possibilité de diffuser ces sites archivés *via* son site web.

ni quels types de sites seront intéressants. Par contre, il paraît vraisemblable que ces usagers rechercheront au moins le même confort et les mêmes potentialités de recherche et de navigation que ceux offerts par le web aujourd'hui.

Les paramètres à prendre en considération pour un projet

Si de nombreux bibliothécaires sont aujourd'hui conscients de l'apport que pourrait représenter pour leur fonds local une collection de sites web régionaux archivés, la mise en pratique d'un tel projet doit faire l'objet d'une véritable réflexion, surtout lorsque l'on prend en considération le coût que risque d'engendrer un tel projet.

En premier lieu, il s'agit donc de savoir si tous les types d'établissements peuvent l'envisager aujourd'hui. À l'heure actuelle et aux vues des difficultés de l'archivage de sites, ce type de projet ne peut être pris en charge que par des établissements importants ayant placé la question du patrimoine local parmi leurs axes politiques fondamentaux. La connaissance du web local constitue également un préalable important. Enfin, la maîtrise de son outil informatique et la présence, dans la bibliothèque, d'un service dédié uniquement à l'informatique sont une nécessité incontournable.

Outre l'établissement lui-même, la qualité du web local est une donnée importante à envisager. En effet, l'archivage de sites nécessitant un investissement minimal relativement important, le web local doit être jugé suffisamment intéressant aussi bien en quantité qu'en qualité (variété des contenus, des auteurs, des formes, qualité des sites…).

D'un point de vue financier, il est, à l'heure actuelle, très difficile d'évaluer précisément le coût d'un projet de conservation du web. Le mode d'acquisition, le nombre de sites concernés, le type d'accès que l'on souhaite proposer aux usagers sont autant de données qui interviennent dans ce calcul. Pour le moment, d'un point de vue matériel, il n'existe pas de système d'archivage produit à grande échelle : tout projet est unique et nécessite l'harmonisation de briques informatiques très diverses. Il existe de nombreux logiciels de collecte gratuits (HTTRACK, par exemple). Cependant, la mise en conformité de ces logiciels gratuits avec le système global de la bibliothèque, leur installation et paramétrage représentent une charge de travail importante pour les informaticiens. Ainsi ne peut-on parler que d'une gratuité très relative. Si l'investissement de base peut s'avérer coûteux, l'entretien de ce fonds d'archives le sera également du fait des migrations successives qu'il faudra gérer pour en assurer la pérennité. Donc, d'un point de vue général, il faut considérer que l'archivage sera une opération onéreuse. Or, comment faire accepter à la tutelle financière qu'Internet, mode de communication relativement peu coûteux, permettant d'accéder à des informations souvent gratuites, nécessite des budgets conséquents pour en assurer la conservation et que ces frais seront en totalité pris en charge par des fonds publics ? La disproportion de prix existant entre le simple accès au document et ce même document conservé risque d'être un sujet difficile de négociation avec la tutelle.

Comment faire accepter à la tutelle financière qu'Internet, mode de communication relativement peu coûteux, permettant d'accéder à des informations souvent gratuites, nécessite des budgets conséquents pour en assurer la conservation ?

Vers la conception d'un projet

Ces paramètres étant énoncés, l'élaboration d'un projet d'archivage de sites web nécessite en premier lieu de définir précisément une politique d'acquisition de sites adaptée. Cette politique d'acquisition doit avant tout se fonder sur la politique globale de l'établissement, tous types de documents confondus. En d'autres termes, il s'agit d'appliquer au domaine de l'Internet la définition que l'établissement donne de la notion « d'intérêt régional ». En dehors de ces critères généraux, cette politique d'acquisition doit également s'appuyer sur des critères particuliers au web et liés à une définition de l'objet site web, comme dit précédemment. Le but est donc d'établir une grille de sélection, adaptée à la politique générale de l'établissement, de façon à circonscrire, du moins théoriquement, la portion du web concernée par l'archivage.

Le choix des modalités d'acquisition intervient ensuite et dépend des ambitions de la bibliothèque mais également de ses moyens en financement et en personnel. Ainsi, la sélection manuelle a l'inconvénient d'être partielle tout en mobilisant de façon permanente un nombre important de personnes. Il peut s'agir toutefois d'une option intéressante pour un établissement dont une partie au moins du personnel maîtrise déjà bien la sélection de sites pour un annuaire ou des signets. La sélection automatique peut s'avérer intéressante lorsqu'il s'agit de collecter une masse importante de sites, toutefois cette option est à déconseiller. Outre qu'elle nécessite un coût d'investissement conséquent, elle sollicite énormément les réseaux. Cette solution est également coûteuse en personnel

très spécialisé, surtout en début de projet. Le paramétrage du robot, les essais répétés et ensuite la maintenance informatique représenteraient des tâches lourdes pour les informaticiens de la bibliothèque.

Le choix peut se porter sur un mode de sélection automatique ou manuel mais, en dehors des inconvénients que nous avons déjà décrits, il faut préciser que l'établissement prend également le risque, en développant son projet de façon entièrement autonome, d'acquérir une grande majorité de sites web qui auront déjà fait l'objet d'une acquisition par la Bibliothèque nationale de France, dans le cadre du dépôt légal. De ce fait, ne faudrait-il pas envisager une collaboration entre la BnF et les grandes bibliothèques municipales intéressées pour l'archivage du web régional ? Pour le moment et à notre connaissance, cette collaboration ne fait pas l'objet d'un projet particulier, sinon sous la forme d'opérations ponctuelles. Ainsi, la BnF a-t-elle fait appel à une contribution des bibliothèques françaises pour le repérage de sites web électoraux locaux lors de la capture des sites liés aux dernières élections de 2002. Il serait certainement envisageable d'entériner, du moins avec quelques établissements volontaires, une coopération intéressante. Celle-ci permettrait à la BnF de bénéficier des compétences développées en région pour la veille sur le web local, et aux bibliothèques concernées d'alléger leur travail d'acquisition et de se concentrer ainsi sur les tâches de conservation et de valorisation auprès de leur public de ces archives locales. Il nous semble que plusieurs pistes de collaboration sont concevables, correspondant pour chacune à un certain niveau d'association et d'implication des établissements[2].

Au premier niveau, on pourrait imaginer que les bibliothèques envoient à la BnF des suggestions de sites web à archiver[3]. En échange, la BnF offrirait une interface de consultation sécurisée aux usagers des bibliothèques concernées. Il ne s'agirait donc pas de placer les archives sur Internet, mais bien de permettre, à un niveau local, une consultation au sein des bibliothèques en province.

Dans ce cas, les bibliothèques ne participent pas véritablement à la conservation proprement dite. Par contre, elles obtiennent un accès pour la consultation. Au niveau de la BnF, ce système, bien que lourd à gérer, peut permettre de profiter d'un repérage local plus fin. En revanche, pour les bibliothèques participantes, cette solution nécessite que le personnel ait une connaissance approfondie et suivie du web local, ce qui n'est pas toujours évident.

> L'archivage de sites web est encore à une étape pionnière de son développement

Au deuxième niveau, les bibliothèques concernées pourraient suggérer à la BnF certains sites à archiver comme dans le premier niveau. Cependant, au lieu de se contenter d'une interface de consultation, les bibliothèques recevraient de la BnF la partie du web qui les concerne sous la forme d'une cassette DLT ou d'un envoi en réseau. Cette partie du web correspondrait à ce que les bibliothèques définissent comme d'intérêt régional (ou autre). Il s'agirait donc pour elles de conserver la partie du web qu'elles auraient acquise et archivée si elles en avaient eu la possibilité (c'est-à-dire si elles avaient eu leur propre robot). La bibliothèque aurait alors en charge la conservation de cette partie locale du web en parallèle avec la BnF, qui conserverait bien entendu la responsabilité de l'archivage de la totalité du web français.

Cette solution offre de nombreux avantages puisqu'elle permet de mettre en place une conservation et une mise en valeur partagées. La BnF y gagnerait certainement une expertise locale qui lui permettrait de compléter efficacement ses collections. Quant aux bibliothèques participantes, elles obtiendraient la maîtrise d'une partie de ces collections ce qui, pour une valorisation à un niveau local, est fondamental. Par contre, la conservation de cette partie du web représenterait une charge lourde pour ces bibliothèques. C'est pourquoi, il paraît peu vraisemblable que toutes les bibliothèques puissent participer à un tel projet.

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

Il s'agit donc d'appliquer un modèle déjà maîtrisé. Toutefois, ce niveau d'organisation, même s'il correspond au plus haut degré de participation et de collaboration des bibliothèques, n'est sans doute pas le plus intéressant. Tout d'abord, parce que les collections obtenues ne seraient pas forcément cohérentes. Ensuite, d'un point de vue matériel, la multiplication de robots sur le territoire et de procédures de collecte risque de surcharger le réseau national et né-

2. Il faut bien entendu préciser que ces pistes ne sont que théoriques et que ni la BnF ni les bibliothèques municipales ne se sont prononcées sur l'une quelconque d'entre elles.

3. À noter que ces suggestions ne concerneraient que les sites web qui n'ont pas fait l'objet d'une collecte et qui ne sont pas déjà archivés.

cessiterait donc un calendrier très contraignant de collectes. Ensuite, les difficultés de personnel qu'engendre en région la maîtrise du dépôt légal imprimeur seraient aggravées par l'acquisition d'un nouveau type de documents.

D'autres possibilités sont envisageables : par exemple, qu'au premier et deuxième niveaux, certaines bibliothèques s'occupent de la collecte régulière de certains sites web locaux particulièrement mouvants. Une bibliothèque peut souhaiter suivre plus finement certains sites web, les enregistrer plus souvent et en déposer une copie à la BnF par exemple.

On peut espérer que l'expérience acquise par les bibliothèques nationales en matière d'archivage électronique et le développement d'outils informatiques adaptés permettront d'envisager plus sereinement la mise en place, à un niveau local, de projets de ce type

En conclusion

L'archivage systématique des sites web d'intérêt régional par les bibliothèques municipales, même importantes, se heurte aujourd'hui aux mêmes contraintes et difficultés que pour les grands établissements nationaux en charge de ce type de projet. L'archivage de sites web est encore à une étape pionnière de son développement.

Faut-il alors considérer que les bibliothèques municipales ou les BMVR n'ont aucun rôle à jouer dans ce type de projet ? Bien au contraire : l'expertise de ces établissements au niveau de la documentation régionale, les connaissances sur la région et le web local, l'expérience acquise en matière de traitement et de valorisation des fonds locaux sont indispensables non seulement dans le cadre de l'archivage du web local mais également pour la mise en place d'un dépôt légal du web national.

Cependant, à l'heure actuelle, il semble risqué pour un établissement de se lancer seul dans ce type de projet, même localisé. On peut alors espérer que l'expérience acquise par les bibliothèques nationales en matière d'archivage électronique et le développement d'outils informatiques adaptés permettront d'envisager plus sereinement la mise en place, à un niveau local, de projets de ce type.

S'agit-il alors d'adopter une attitude attentiste ? Ne pourrait-on envisager une collaboration entre ces bibliothèques et la Bibliothèque nationale de France ? Une telle solution, nous l'avons vu, serait certainement intéressante mais pose bien entendu plusieurs problèmes, notamment en termes d'organisation et de modalités de fonctionnement d'une collaboration.

Mars 2003

BIBLIOGRAPHIE SOMMAIRE

ASCHENBRENNER, Andreas, « Long-term preservation of digital material: Building an Archive to preserve digital cultural heritage from the Internet » (Master thesis). En ligne : http://www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/
[Dernière consultation le 16 mars 2003.]

BULLOCK, Alison, « La conservation de l'information numérique : ses divers aspects et la situation actuelle », *Flash Réseau,* 1999, nº 60. En ligne : http://www.nlc-bnc.ca/9/1/p1-259-f.html
[Dernière consultation le 24 mars 2003.]

LUPOVICI, Catherine, « Les stratégies de gestion et de conservation préventive des documents électroniques », *Bulletin des bibliothèques de France.* 2000, t. 45, nº 4, p. 43-54. En ligne : http://bbf.enssib.fr/bbf/html/2000_45_4/2000-4-p43-lupovici.xml.asp
[Dernière consultation le 24 mars 2003.]

MASANES, Julien, « The BnF's project for web archiving », contribution for the European conference on digital libraries (ECDL) 2001: What's next for digital deposit libraries? Darmstadt, 8 septembre 2001. En ligne : http://bibnum.bnf.fr/ecdl/2001/france/sld001.htm
[Dernière consultation le 16 mars 2003.]

OCLC (Online Computer Library Center), web characterization. In *OCLC's website* : http://wcp.oclc.org
[Dernière consultation le 16 mars 2003.]