Date : 10/07/2008

L'archivage d'Internet, un défi pour les décideurs et les bibliothécaires : scénarios d'organisation et d'évaluation L'expérience du consortium IIPC et de la BnF

Gildas Illien, chef du service du dépôt légal numérique, Bibliothèque nationale de France

Meeting: **107. Managing libraries in a changing environment – legal, technical and organisational aspects**

Simultaneous Interpretation: English, Arabic, Chinese, French, German, Russian and Spanish

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

Introduction

Archiver le Web : un projet fou ? C'est en tout cas la réaction première et, somme toute, compréhensible, que l'on est en droit d'attendre de tout internaute normalement constitué. L'idée même de capturer et de conserver cet ensemble diffus, multiforme, massif, éphémère, et en mutation constante, ne serait-elle pas l'ultime invention d'une bibliothèque devenue absurde à force de vouloir tout collecter, tout conserver ?

Face à un tel défi, la tentation est grande de passer son chemin. Les bibliothécaires ont bien assez à faire avec la collecte du papier, leurs magasins sont saturés, leurs effectifs en baisse et ils ont déjà fait preuve de beaucoup de compréhension en s'abonnant aux ressources électroniques, en mettant leur catalogue en ligne ou en organisant le référencement des sites Web les plus fiables. Mieux encore, ils ont entrepris de numériser leurs collections et lancé de nouveaux services qui placent la bibliothèque à un simple clic des internautes trop paresseux ou trop éloignés pour prendre le chemin de la salle de référence. Faudrait-il à présent déployer des efforts coûteux pour tenter de capturer l'immensité du Web, ce puits sans fond, dont les contenus regorgent de pornographie, de publicité, et de milliards de récits de vies ordinaires pleins de fautes d'orthographe ?

Pourtant…en cherchant bien dans ses souvenirs vieux de cinq siècles (les bibliothécaires français vivent très longtemps), un bibliothécaire de la Bibliothèque nationale de France (BnF) se souvient d'autres déferlantes.
La naissance de l'imprimerie, qui a rendu la bibliothèque du Roi beaucoup trop étroite pour que le Savoir continue de tenir dans une seule pièce : il a fallu pousser les murs, construire, multiplier les inventaires et les catalogues ; l'avènement de la presse qui, au 19ème siècle, a brusquement obligé à passer à la vitesse supérieure, en recrutant des armées de magasiniers robustes capables de réceptionner chaque matin plusieurs tonnes de papier ;

l'invention du cinéma, de la radio, de la télévision, qui ont encore étendu le spectre de la bibliothèque et démultiplié ses flux et ses stocks, achevant de la faire entrer dans l'ère industrielle ; l'informatique et les technologies numériques enfin, qui ont produit leurs propres avatars, les cassettes, disquettes, logiciels, CD et DVD venant compléter la panoplie vertigineuse d'un patrimoine national désormais multimédia.

Entre temps, notre vénérable collègue s'est également souvenu que les chercheurs (qui vivent moins vieux que les bibliothécaires) changeaient de goût et d'intérêt d'une génération à l'autre. Certains ouvrages autrefois prisés n'intéressent plus personne. A l'inverse, des documents imprimés sur papier bon marché considérés comme mineurs, voire indignes, tels les annuaires, les catalogues de vente par correspondance, les magazines de mode, les tracts, les revues érotiques sortent de leurs rayons et de leur Enfer parce que c'est aussi grâce à ces documents de petite vie que les chercheurs contemporains construisent l'histoire. Un blogueur médiéviste m'a écrit en 2005 : « *je trouve formidable que vous archiviez les skyblogs*[1]*, si les adolescents avaient pu bloguer au Moyen Age et que la Bibliothèque nationale avait archivé ces blogs, c'est une autre histoire de France qu'on aurait pu écrire.* »

Pour illusoire qu'elle paraisse au premier abord, l'idée d'archiver le Web s'inscrit donc dans un continuum historique et patrimonial. Plus qu'une option, compte tenu de l'importance qu'a pris Internet dans tous les secteurs du savoir et de la société, c'est devenu une nécessité et même une urgence : bien des supports qui nous sont familiers – livres, périodiques, phonogrammes, vidéogrammes- ont entrepris leur grande transhumance vers l'Internet. De plus, l'Internet génère ses propres modes d'expression et de publication, son matériau scientifique et ses sources, qui n'ont pas d'équivalent sur support. A terme, délaisser l'espace du Web et ne pas reconnaître sa dimension patrimoniale, reviendrait à programmer la transformation définitive de la bibliothèque en un musée incapable d'assurer le renouvellement du matériau constitutif de sa propre mémoire. Il convient simplement d'être raisonnable et réaliste : faire son deuil de l'ambition d'exhaustivité et se contenter d'échantillons, de sélections – de traces.

Dans le cadre de cette session qui traite de l'innovation et de l'accompagnement au changement, mon propos vise à stimuler l'imagination des professionnels à la recherche d'une stratégie pour introduire ce sujet nouveau dans les missions et le fonctionnement de leur institution, qu'il s'agisse de convaincre les politiques chargés de son pilotage et de son financement ou les équipes qui auront à accueillir et à mettre en œuvre un tel projet.

- Où implanter cette activité pour qu'elle s'appuie sur les compétences techniques nécessaires tout en l'inscrivant dans des périmètres garants de la continuité des collections ?
- Comment qualifier, mesurer, valoriser une telle activité, à la fois nouvelle (par la nature de son objet et les techniques utilisées pour le traiter) et ancienne (par ses finalités, qui s'inscrivent dans le développement des collections et du dépôt légal) ?

Ce sont ces deux aspects que j'aborderai ici : l'organisation du travail et l'évaluation des collections – en insistant davantage sur le second point, le premier rejoignant des stratégies mieux connues dans le domaine de la documentation numérique.
Je m'appuierai sur des enquêtes et des discussions récentes conduites parmi les membres du Consortium international pour la préservation de l'Internet (IIPC) et sur l'expérience du dépôt légal de l'Internet à la BnF.

[1] Skyblogs : nom donné aux blogs hébergés par la plate-forme créée par la radio française SkyRock, très populaire chez les adolescents français.

I- Préalables

Pour ceux qui découvrent l'archivage du Web aujourd'hui, j'apporterai en préalable quelques précisions utiles à la compréhension de ce qui suit, tout en rappelant que l'objet de mon exposé n'est pas une introduction technique à l'archivage du Web.

- **Qu'est-ce que le consortium IIPC[2] ?**

Fondé en 2003 à l'initiative d'une dizaine de bibliothèques nationales et de la fondation américaine Internet Archive, le consortium a pour principaux objectifs la promotion de l'archivage du Web dans le monde et la constitution collaborative d'un ensemble de logiciels et de normes permettant la collecte, la consultation et la préservation de long terme des données. Tous les outils développés par IIPC sont des logiciels libres, afin d'encourager l'interopérabilité future des archives constituées à travers le monde, et pour faciliter leur utilisation par les pays les moins nantis. En 2007, le consortium a entrepris son élargissement à de nouvelles institutions. Il compte aujourd'hui une quarantaine de membres en Amérique du Nord (dont Bibliothèque et Archives Canada et Bibliothèque et Archives nationales du Québec), en Europe, et en Australasie.

- **En quoi consiste le projet de la BnF[3] ?**

La BnF s'est engagée dans l'archivage du Web au début des années 2000. Plusieurs années d'expérimentation et de lobbying ont été nécessaires avant le vote, en 2006, de la loi qui a officiellement étendu le dépôt légal français à l'Internet[4]. Sa pratique du dépôt légal explique l'approche française de l'archivage du Web, qui est donc inscrit dans la continuité de ses missions fondamentales. La BnF s'est dotée d'une équipe de 7 personnes pour conduire cette activité, qui repose par ailleurs sur un réseau d'une centaine de bibliothécaires. Les collections qu'elle a réunies en partenariat avec Internet archive représentent aujourd'hui 110 Téraoctets de données, soit 11 milliards de fichiers. Depuis le mois d'avril 2008, elles peuvent être consultées par le public dans ses salles de lecture.

- **L'archivage du Web : mode d'emploi**

Les sites Web sont moissonnés automatiquement au moyen de robots « aspirateurs » appelés *crawlers* ou *spiders,* qui se comportent comme des internautes automatiques : ils cliquent de lien en lien sur tous les sites qu'ils rencontrent. A partir d'une liste d'adresses (URL) de sites appelées graines ou *seeds*, ils parcourent le Web de page en page et copient tous les fichiers qu'ils découvrent, soit en profondeur (liens entrants) à l'intérieur d'un même site, soit en largeur (liens sortants) vers d'autres sites. Le robot le plus fréquemment utilisé par les bibliothèques s'appelle Heritrix, il a été conçu et est maintenu par le consortium IIPC.

Les robots rencontrent régulièrement des obstacles : pièges, exclusions, fichiers et animations complexes (en *Flash*, notamment), contenus payants (le robot ne possède pas de carte bancaire…) ou seulement accessibles par identification, reconnaissance d'adresse IP, mot de passe ou formulaire. Ces fichiers du Web profond ne peuvent pas être capturés sans intervention humaine. De plus, une campagne de moissonnage, même de taille moyenne, est potentiellement illimitée : pour des raisons économiques, il est nécessaire d'en assurer la supervision mais aussi l'arrêt sans qu'on ait l'assurance que tous les contenus aient été effectivement collectés. Pour ces raisons, les archives du Web sont des documents lacunaires puisqu'il peut manquer des fichiers, des pages, mais aussi parce qu'il n'est

[2] www.netpreserve.org
[3] http://www.bnf.fr/pages/infopro/depotleg/dl-internet_intro.htm
[4] LOI n° 2006-961 du 1er août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information (DADVSI), http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000266350&dateTexte=

évidemment pas possible de moissonner tous les sites en permanence et que l'archive capturée ne constitue donc qu'un échantillon, une photographie – une trace- du site à un instant donné.

Les institutions de mémoire ont développé des approches différentes pour exploiter les robots. Ces stratégies varient en fonction de leur situation juridique, de leur politique documentaire et de leurs ressources. Certaines institutions ont l'obligation légale d'obtenir des éditeurs de sites l'autorisation de capturer leurs publications, ou bien ont fait le choix d'une politique délibérément sélective : en ce cas, le robot navigue sur un nombre limité de sites et restreint sa collecte à des publications qui ont été préalablement identifiées par des bibliothécaires. D'autres établissements ont, au contraire, le droit de moissonner les sites sans autorisation préalable et ont développé une approche beaucoup plus massive, préférant collecter le tout plutôt que ses parties. En ce cas, le robot est lancé sur de très gros volumes de sites et travaille pendant plusieurs semaines sur un mode qu'on qualifie d'exploratoire : à partir d'une liste de départ déjà importante, il va découvrir beaucoup d'autres ressources qui n'avaient pas été identifiées au préalable. Beaucoup d'institutions, comme la BnF, conjuguent ces deux approches en réalisant un ou deux instantanés (*snapshots*) annuels de leur domaine national dont ils assurent ainsi une collecte en surface, qu'elles complètent par des campagnes ciblées, plus fréquentes et plus profondes, issues des sélections opérées par leurs bibliothécaires, autour d'un événement ou d'un thème particulier (par exemple : les Web-campagnes électorales).

II- Stratégies d'organisation pour la mise en place d'un projet d'archivage

Pour démarrer une activité d'archivage Web, il faut bien sûr disposer d'un cadre juridique permettant cette activité, au moins à titre expérimental, remporter l'adhésion et le soutien des décideurs, constituer une équipe motivée et compétente. J'insisterai ici sur la stratégie de mise en place d'un tel projet du point de vue de l'organisation, en privilégiant le cas des bibliothèques et plus spécifiquement des bibliothèques nationales, qui sont aujourd'hui les principaux opérateurs de l'archivage du Web, qu'elles intègrent le plus souvent à leur mission nationale de dépôt légal.

1. Le problème : comment créer des mutants ?

Comme de plus en plus d'activités bibliothéconomiques, la gestion des archives du Web nécessite des compétences à la fois documentaires et informatiques. Les spécificités de traitement liées à la collecte des sites Web impliquent un savant mélange de ces deux types de qualification. De plus, l'adjonction de l'archivage du Web aux missions plus classiques de la bibliothèque implique d'être à la fois parfaitement au fait de ses fonctionnements traditionnels et capable d'innovation afin de les remettre en cause ou de les adapter. Les principaux défis sont les suivants :

- appréhender la notion de document patrimonial depuis l'environnement qui est propre au Web, oblige à reconsidérer bien des catégories structurantes de nos classifications documentaires : les divisions liées au territoire (documentation nationale / étrangère), au support physique (imprimés/ audiovisuel/ multimédia…), aux thématiques issues de l'encyclopédisme (arts/ sciences/ sciences humaines…), ou au statut des producteurs (éditeurs patentés/ autoproduction) ne sont plus vraiment pertinentes sur le Web ou nécessitent d'être réinterprétées ;

- conserver une responsabilité documentaire sur ces collections, mais en intégrant le changement d'échelle radical induit par la volumétrie du Web : la collection doit être qualifiée à un niveau de granularité supérieur, qui conduira souvent à renoncer à

définir des choix au niveau des unités (sites, fichiers) au profit d'un pilotage au niveau d'ensembles plus conséquents (domaines, périodes, types d'éditeurs, etc.) ;

- organiser la veille prospective et la sélection de sites, ou d'ensembles de sites dans une perspective d'archivage implique de bien connaître l'architecture technique, la réalité sociale et les usages de l'Internet, ainsi que le fonctionnement des robots de collecte: les bibliothécaires impliqués doivent avoir de solides compétences informatiques et être parfaitement à l'aise avec le média Internet.

- De la même façon, le pilotage des robots de collecte n'est pas une opération purement technique. La qualité d'une collection d'archives est déterminée pendant la capture des sites dont la collecte se programme, se surveille et s'évalue : les informaticiens impliqués doivent être pleinement conscients des contenus qu'ils produisent et des enjeux documentaires et patrimoniaux qui y sont associés.

Une équipe idéale est donc constituée de *mutants* dont les compétences documentaires et informatiques tendent à fusionner de plus en plus. La stratégie d'implantation de ce nouveau service dans l'organisation de la bibliothèque va donc consister à trouver la formule qui favorisera la rencontre de ces compétences tout en respectant le partage des tâches au sein de l'établissement et ses équilibres politiques et fonctionnels.
Je présenterai ici la stratégie appliquée à la BnF, stratégie que l'on retrouve dans la plupart des bibliothèques nationales qui ont atteint un certain degré de maturité dans la conduite de leurs activités numériques.

2. Scénario d'infiltration

• Phase I : implanter l'équipe dans une entité d'innovation numérique

L'archivage du Web se présente toujours comme un projet, une expérimentation. La formation de l'équipe mutante passe par un auto-apprentissage des outils et des procédures de collecte à partir de tests en situation réelle. Toutefois, le dispositif de production ne peut être stabilisé sans qu'on ait parallèlement ou préalablement défini les objectifs documentaires et le cadre juridique de l'activité.

Pendant une première période, il faut donc tout à la fois définir et formaliser les intentions documentaires de l'établissement en impliquant des décideurs et des agents des services responsables des collections et du dépôt légal, et développer progressivement le dispositif de production et les compétences techniques adaptés à ces objectifs. La construction de ce double modèle (à la fois documentaire et informatique) implique une itération constante entre techniciens et bibliothécaires, qu'on aura tout intérêt à placer dans une entité d'innovation déchargée de fonctions de production courante mais bénéficiant d'une visibilité stratégique forte, et donc en situation de dialogue, de communication et de négociation avec les autres services de l'établissement, y compris les services juridiques et financiers. Il faut également prévoir au sein de l'équipe des compétences relationnelles fortes et intégrer au projet une mission de communication et de formation interne auprès des équipes opérationnelles, tant au service informatique qu'au sein des collections. C'est ce travail de sensibilisation (à conduire sous formes d'ateliers, de formations, de séances d'information pour le personnel) qui permettra de faire connaître le projet et d'impliquer une plus large communauté professionnelle dans la définition des choix documentaires les plus fondamentaux.

A la BnF, cette entité d'accueil était le département de la bibliothèque numérique ; à la Library of Congress, il s'agit du Bureau des initiatives stratégiques (*Office of Strategic Initiatives*), à la Bibliothèque nationale de Singapour, c'est la Division des ressources et services numériques (*Digital Resources and Services Division*). Cette activité est souvent

placée à proximité voire sous l'autorité d'autres chantiers numériques stratégiques, la numérisation et la préservation numérique en particulier.
Tactiquement, il est préférable de constituer cette équipe projet dédiée à l'archivage du Web tout en maintenant un lien administratif et fonctionnel entre chacun de ses membres et un département traditionnel : les ingénieurs au service informatique, les bibliothécaires dans les entités en charge du dépôt légal ou des collections.
Parmi les membres du consortium IIPC, plus de 90% déclarent ainsi que leur personnel dédié à l'archivage du Web est implanté dans plusieurs points de l'établissement (principalement : les collections, le dépôt légal, le catalogage, la conservation et l'informatique), ce qui n'est pas contradictoire avec un fonctionnement en mode projet. Au contraire, cette configuration favorisera dès le départ l'émergence de pratiques collaboratives entre des métiers et des cultures professionnelles différentes tout en faisant de l'archivage du Web le projet de l'établissement et pas seulement d'un service.

Le choix d'implantation du pilotage du projet se révèle déterminant. L'expérience des membres du consortium IIPC a montré que les institutions qui avaient confié le pilotage du projet d'archivage du Web à un département de collections ou de dépôt légal avaient plutôt fait le choix d'un modèle d'archivage sélectif, plus proche des références et des pratiques traditionnelles des bibliothécaires. A l'inverse, celles qui ont confié le pilotage au département informatique ont plutôt choisi un modèle d'archivage de masse et à grande échelle. Placer le pilotage dans une entité d'innovation bénéficiant d'une certaines neutralité par rapport aux divisions culturelles entre professions techniques et professions scientifiques est souvent la meilleure solution pour aborder le projet dans toutes ses problématiques. Les institutions qui, comme la BnF, ont fait ce choix, ont généralement abouti à un modèle mixte qui combine collectes larges et collectes sélectives.

- **Phase 2 : la dissémination dans des entités traditionnelles**

Lorsque le projet a atteint un niveau de maturité suffisant, c'est-à-dire qu'il commence à produire régulièrement des données et que celles-ci peuvent être gérées dans un circuit de traitement à peu près complet (collecte, indexation, accès, sauvegarde), il faut se poser la question du redéploiement des fonctions et des personnels à l'intérieur de l'organisation. En effet, maintenir une activité innovante en dehors des grands départements de production lorsqu'elle est arrivée à maturité présente des risques : qu'elle soit marginalisée et, à terme, politiquement ou budgétairement fragilisée parce que placée en dehors des activités les plus pérennes ; qu'elle mobilise les équipes chargées de pousser l'innovation alors qu'elle n'en est plus une, et empêche de fait l'émergence de projets nouveaux ; qu'elle reste cantonnée à quelques individus alors que la mutation culturelle qu'elle implique peut profiter plus largement à d'autres équipes et à d'autres secteurs.

Le choix de redéploiement dépendra bien sûr de la structure et de l'histoire de chaque institution, et de sa maturité vis-à-vis des chantiers numériques. La BnF a fait le choix radical de supprimer son département de la bibliothèque numérique, en considérant que c'était tout l'établissement qui était désormais impliqué dans le numérique. Les activités de numérisation avaient déjà été confiées aux départements de la conservation et de la coopération ; la gestion du site web attribuée au service de communication. L'archivage du Web, encadré juridiquement par le dépôt légal, a été confié au département du dépôt légal qui s'est vu doté en avril 2008 d'un nouveau service : le *service du dépôt légal numérique*.

Afin de s'assurer du maintien d'une implication forte du département des systèmes d'information, l'équipe projet a été divisée en deux groupes : les bibliothécaires, rebaptisés « chargés de collections numériques » après assimilation de toutes les compétences informatiques utiles, ont rejoint le département du dépôt légal avec pour mission principale d'assurer la *maîtrise d'ouvrage* du dépôt légal numérique en animant notamment un réseau d'une centaine de bibliothécaires dont ils sont les prestataires ; les informaticiens, après

acculturation au contact des bibliothécaires, ont rejoint une nouvelle branche du département des systèmes d'information officiellement chargé de la *maîtrise d'œuvre* du dépôt légal numérique. Cette nouvelle organisation, qui fixe clairement les rôles de chacun, a été complétée par la mise en place d'instances de concertation et de décision qui se réunissent régulièrement afin d'assurer la coordination entre les deux équipes. Dans d'autres bibliothèques du consortium, il est confirmé que c'est la clarté et la qualité du lien qui relie les équipes d'ingénieurs et de bibliothécaires qui constitue le principal facteur de succès du point de vue organisationnel.

III- Quels indicateurs pour les archives du Web dans une bibliothèque ?

1. Le problème : les archives du Web ne seraient comparables à rien de connu

Le mode de production des archives du Web décrit précédemment pose un vrai défi du point de vue de l'évaluation des collections telle qu'elle est habituellement pratiquée en bibliothèque. En effet, les statistiques et les documents générés par les robots de collecte ne ressemblent à rien de connu dans nos établissements et ne sont pas normalisés. En transposant les modèles éditoriaux connus, ne pourrait-on pas comparer un site Web et son archive :
- à un périodique, avec un titre (l'adresse du site Web) et des numéros (ses captures successives) ?
- ou à une monographie (l'adresse du site Web), avec ses éditions (captures) successives ?

La solution semble simple, mais elle est techniquement impraticable.
En effet, la notion de site Web est une construction purement intellectuelle, qui permet de désigner un ensemble éditorial cohérent généralement produit par un même auteur, producteur ou diffuseur clairement identifié : le blog de ma petite sœur, le site de la BnF, le site de l'IFLA, le site du journal *Le Monde*, etc. S'il est capable de restreindre sa collecte à des noms de domaine (*domain names*), à des noms de serveurs (*hosts*)[5] ou à des pages précises au moment de son paramétrage, le robot ne collecte que des fichiers (URI), c'est-à-dire la plus petite unité constitutive d'un site : à la manière d'un recueil (qui compile des coupures de presse, des tracts, des photographies…) il accumule, pêle-mêle, tous les éléments qu'il trouve sur la Toile (des fichiers texte, du son, des vidéos, des rapports voire des livres entiers sur PDF, des logos ou des vignettes comparables à des images) mais sans prendre en compte l'unité éditoriale qui les unit du point de vue de l'internaute, et qui n'est qu'une vue de l'esprit, indépendante de la structure logique de l'Internet. Le Web est, fondamentalement, un vaste système d'adressage et de requêtes permettant de stocker, d'afficher et de lier entre eux des fichiers stockés sur des serveurs, rien de moins, rien de plus.

Cette situation nous limite donc à l'unité atomique du fichier, l'URI, le plus petit dénominateur commun du Web et de ses archives, et le seul qu'on puisse qualifier de document, encore que beaucoup de bibliothécaires rechignent à procéder à la dissection d'ensembles éditoriaux cohérents en tout petits morceaux, à la manière d'un puzzle. Le résultat est à première vue inutilisable, car ce chiffre paraîtra toujours très élevé au regard des collections traditionnelles : la Bibliothèque nationale d'Islande est en possession de 382 millions de fichiers, Bibliothèque et Archive Canada en déclare une centaine de millions, Internet Archive en revendique 115 milliards.

[5] Certains comptent également le nombre de *hosts* (serveurs) en considérant que c'est l'unité qui se rapproche le plus du site Web, ce qui est vrai dans certains cas (lorsque tous les contenus d'un site sont stockés sur un même serveur). Toutefois, on sait que les contenus d'un site peuvent être hébergés sur plusieurs serveurs et qu'un même serveur peut héberger des contenus relatifs à plusieurs sites différents, une tendance qui tend à se répandre avec le Web 2.0 et le Web sémantique, ce qui fausse lourdement les calculs.

Le problème n'est pas anodin, car dans le contexte de la réforme des politiques publiques et de la généralisation des indicateurs de performance, il n'est pas simple de mesurer et de promouvoir la valeur des collections issues de l'archivage du Web en utilisant les statistiques barbares et gigantesques issues des rapports de collecte d'un robot.

2. Des solutions simples pour tous

- **Le fichier : atomique mais pratique**

Le problème précédemment décrit ne nous est pas étranger. Lorsqu'une bibliothèque doit faire le compte de ses acquisitions, n'est-elle pas confrontée à la même difficulté ? La BnF conserve une grande diversité d'objets : des millions de livres, bien sûr (qu'on dénombre en titres et en exemplaires), et de périodiques (dont les titres sont tantôt morts, tantôt vivants, et qu'on dénombre en fascicules ou en volumes reliés), mais aussi des supports audiovisuels et un très grand nombre de collections spécialisées dont l'unité de mesure varie à chaque fois : quoi de commun entre une estampe ou une photographie, une médaille, une carte, un globe, un recueil (de tracts, d'éphémères…) ou la collection des costumes de scène de Sara Bernhardt ? Toute bibliothèque est confrontée à l'impossible agrégation d'unités documentaires très différentes. C'est exactement la même chose pour les archives du Web. L'agrégation des fichiers du web n'a donc rien de honteux ni d'inédit, elle est *pratique*, et on pourra l'utiliser en prenant soin de faire l'analogie avec les collections sur support de la bibliothèque et en ayant l'honnêteté de préciser que ces collections recouvrent des types de documents différents et contiennent nombre d'exemplaires multiples. Le nombre de fichiers collectés sert à mesurer l'évolution de l'activité dans le temps et à faire des comparaisons internationales. L'essentiel demeure, comme pour toute statistique, que la mesure reste uniformément et durablement biaisée.

- **L'octet : le mètre linéaire de demain**

L'autre solution est à rechercher, cette fois encore, dans les pratiques existantes. Que fait une bibliothèque qui doit défendre un projet d'extension, de déménagement ou de reconditionnement de ses collections ? Elle englobe la diversité de ses fonds afin de les rapporter à une unité de gestion simple et pratique pour évaluer les coûts et organiser les opérations : le mètre ou le kilomètre linéaire. La même démarche peut s'appliquer aux archives du Web mais en transposant le besoin de stockage physique (dans des magasins) aux modalités du stockage numérique où c'est le poids des données qui sert d'étalon. A l'heure où beaucoup d'institutions s'engagent dans des projets ambitieux d'archivage numérique (*digital repositories*) et que le coût du stockage numérique est devenu un enjeu budgétaire crucial, cette démarche apparaît d'autant plus pertinente. C'est donc en *gigaoctets* (pour les institutions aux pratiques sélectives) ou en *téraoctets* (pour les plus gourmandes) voire en *pétaoctets* (demain) que les archives du Web peuvent également être mesurées. Là encore, on aura tout intérêt à faire l'analogie avec le kilomètre linéaire et à filer la métaphore du magasin pour introduire cette mesure auprès des décideurs, d'autant plus qu'elle est pratiquée par les informaticiens, ce qui s'avère très commode pour le calcul des coûts de traitement, de stockage et de préservation à long terme.

Au niveau international, une enquête récente conduite auprès des membres du consortium IIPC a montré que si les pratiques de mesure des collections variaient sensiblement d'une institution à l'autre, plus de 90% utilisaient l'octet et le fichier (URI) comme unités de mesure de leurs collections. C'est d'ailleurs au moyen de ces indicateurs que les Web-archivistes du monde entier ont pris l'habitude de qualifier leur activité. Ces unités sont fausses[6], bien sûr,

[6] La présence de nombreux doublons dans les archives biaise fortement la mesure en nombre de fichiers. Ce problème pourrait s'estomper dans le futur avec l'introduction de modules de déduplication dans le robot Heritrix, qui aura pour effet paradoxal la

mais de manière stable : ce sont les mesures les plus simples et les plus efficaces que l'on puisse utiliser aujourd'hui pourvu qu'on ait pris la peine de les expliquer aux décideurs et aux agents qui auront à les utiliser.

3. Des solutions sur mesure

Sans épuiser les possibilités, je souhaiterais compléter mon propos par quelques initiatives mentionnés dans cette enquête et qui m'ont paru particulièrement intéressantes dans une perspective d'accompagnement au changement. On peut calculer un très grand nombre d'indicateurs à partir des archives du Web. Le bon gestionnaire se distinguera en n'en choisissant que quelques-uns, adaptés à chaque type d'objectif et d'interlocuteur.

- **Des indicateurs pour les informaticiens et les administrateurs numériques**

Les responsables du système de production (qui supervisent les opérations de collecte, d'indexation, et de préservation des archives) s'embarrassent rarement des questions philosophiques qui préoccupent les bibliothécaires : le problème bibliothéconomique de l'unité et de la cohérence documentaire et chronologique des archives du Web n'en est pas vraiment un pour eux. Pour l'informaticien, un fichier est un fichier, mais ses préoccupations portent sur le nombre, la diversité et l'importance des flux et des stocks de fichiers qui sont associés aux opérations de traitement. Outre le poids des données, qui reste l'unité de gestion la plus utile pour identifier les points critiques dans le circuit de traitement (*workflow*), les informaticiens utilisent des indicateurs de production, de charge et de performance qui ne sont pas spécifiques aux archives du Web mais qui permettront d'évaluer l'adéquation d'une infrastructure informatique aux besoins de la production, de repérer des goulets d'étranglement ou de donner une visibilité sur le service rendu. Par exemple : la durée moyenne entre la capture d'un site et son indexation permet de donner aux utilisateurs professionnels comme au public une indication utile sur le moment où l'archive d'un site identifié est susceptible de devenir visible depuis les interfaces d'accès.

Le *crawl*

L'unité du *crawl* ou du *job* est propre aux collectes du Web. Elle correspond au processus de capture d'un ensemble de sites à partir d'une même instance du robot et répondant à une même « commande » (une collecte pouvant être découpée en plusieurs commandes). C'est une unité importante pour calculer les charges supportées par la production et pour planifier et répartir le travail des ingénieurs et des machines. Par extension, l'évolution du nombre de crawls sera également utilisée par les institutions pratiquant la collecte de grande échelle (telles Netarchive.dk au Danemark, la BnF ou la Bibliothèque nationale d'Islande) afin de mesurer l'évolution du rythme de l'activité de collecte. Elle constitue de fait un indicateur objectif de l'accroissement des collections et du degré de maturité du projet.

La répartition par types MIME

Les rapports statistiques fournis par le robot Heritrix produisent également des statistiques relatives aux types MIME des fichiers collectés, par exemple : 70% de fichiers .html, 10% de fichiers .txt, 10% de fichiers .jpeg, etc. On sait que les types MIME sont généralement truffés d'erreurs, cependant, à l'échelle d'une production de masse, ils restent des indicateurs pertinents pour apprécier à grandes mailles la répartition globale d'une collection entre familles de fichiers. L'information présente un intérêt documentaire, car c'est un moyen utile de qualifier la répartition des contenus collectés en masse (texte, vidéo, images, son…).

baisse tendancielle du nombre de fichiers collectés pendant quelques années. Pour ce qui concerne la mesure en poids (octets), elle est biaisée par la prise en compte, variable selon les institutions, des processus de compression des fichiers d'une part, et l'intégration ou non des index, d'autre part.

Mais elle est également très précieuse du point de vue de la préservation, car c'est la connaissance des principaux types de fichiers présents dans la collection qui permettra d'apprécier les risques, les priorités et de définir des stratégies de migration ou d'émulation pour l'archivage pérenne des données.

Les fichiers (W)ARC

Enfin, quelques institutions utilisent le fichier (W)ARC (fichier container propre aux archives du Web, en cours de normalisation à l'ISO) comme unité complémentaire de mesure de leurs collections[7]. Cette approche est elle aussi dictée par un souci de préservation. En effet, pour tous ceux qui gèrent de gros volumes, le traitement des fichiers à l'unité constitue un véritable défi compte tenu du nombre de paquets de données à prendre en compte. Le format (W)ARC, qui permet le stockage en vrac des fichiers à mesure qu'ils sont copiés par le robot constitue une unité de préservation et de gestion qui fera référence tout au long du cycle de vie du document né numérique. Dénombrer ses collections en (W)ARC et croiser cette information avec les statistiques de types MIME permet de documenter ses collections dans la perspective de leur préservation.

Ces indicateurs sont particulièrement utiles à la planification et à la documentation des activités de traitement et à l'évaluation des besoins en ressources informatiques.

- **Des indicateurs pour les bibliothécaires**

Sites et graines

Lorsque les sites sont sélectionnés à l'unité, il est bien sûr possible de compter le nombre de sites « commandés » au robot (nombre de graines) mais cela ne préjuge pas du nombre de documents effectivement collectés, ni ne peut s'appliquer aux collectes de grande échelle qui fonctionnent en mode large et exploratoire : ce qu'on mesure alors, c'est l'effort de sélection des bibliothécaires, pas la collection à proprement parler. La plupart des bibliothèques qui pratiquent l'archivage sélectif utilisent l'unité du « site web » ou de la « graine », éventuellement assimilés à des « titres » (de publications) et assortis d'un nombre d'exemplaires (ou d'instances), comme le font la British Library ou la Bibliothèque nationale d'Australie. Ces unités, assimilables aux pratiques de traitement pour les documents sur support, permettent d'évaluer le travail de repérage, et, le cas échéant, de catalogage, effectué par leurs agents en utilisant un système de références qui leur est familier. Elles servent également, par extension, à qualifier l'étendue et l'évolution des fonds collectés.

Collections

Au niveau de granularité supérieur, certains comme la *Library of Congress* ou la *California Digital Library* utilisent également la mesure de la « collection » comme un sous-ensemble

[7] Lorsqu'un fichier est collecté sur le Web, on le copie au sein d'un fichier container ARC (http://www.archive.org/web/researcher/ArcFileFormat.php), avec les métadonnées recueillies au cours de l'opération de collecte. Les fichiers ARC peuvent contenir de multiples objets numériques : on ne les ferme (c'est-à-dire qu'on arrête d'y copier des fichiers) que lorsqu'ils ont atteint leur taille maximale de 100 mégaoctets. Les fichiers ARC fonctionnent ainsi comme l'équivalent numérique des cartons d'archives, que l'on remplit avec des documents de toutes sortes. On parle désormais de fichiers (W)ARC car ce format est destiné à évoluer vers le format WARC (http://bibnum.bnf.fr/WARC/index.html), qui propose des fonctionnalités de traitement étendu (comme la gestion des migrations ou des doublons). Sa taille cible a été élargie (1 gigaoctet) pour faire face à l'augmentation de la taille des objets sur le Web. Sa normalisation définitive est attendue à l'ISO dans l'année 2008.

présentant une unité thématique ou, plus souvent, événementielle (le Tsunami, l'ouragan Katrina, les attentats du 11 septembre, une élection nationale…). La couverture d'un événement nécessite en effet une mobilisation importante des personnels dans un temps limité, cette mesure permet donc de qualifier et de pondérer l'activité en nombre de projets spécifiques, qui sont à la fois des indicateurs organisationnels (ils mesurent le degré de maturité, de dynamisme ou d'acceptation de l'archivage du Web dans les équipes) et documentaires (chaque collection particulière issue d'un projet constitue un ensemble documentaire et scientifique et un point d'entrée intéressants pour le public).

Prestations

Certaines institutions qui sont organisées entre un service « prestataire » (maître d'œuvre), chargé de la réalisation technique des collectes, et un groupe ou un service d'utilisateurs « commanditaires » (maître d'ouvrage), chargés de la sélection des sites, ont développé des indicateurs spécifiques permettant de mesurer la qualité du service rendu (par exemple : ratio nombre de sites sélectionnés / nombre de sites capturés) mais aussi la pertinence des propositions de collecte (un site peut ne pas être capturé s'il a disparu ou si la syntaxe de l'adresse communiquée aux ingénieurs est incorrecte). Ces indicateurs permettent d'organiser le dialogue entre informaticiens et bibliothécaires, d'instaurer la confiance, ou, au contraire, d'objectiver des dysfonctionnements qui peuvent être liés à des défaillances du système de production ou à une formation insuffisante des bibliothécaires.

L'ensemble des indicateurs mentionnés jusqu'ici tant pour les informaticiens que pour les bibliothécaires constituent, chacun dans leur domaine, des indicateurs de production internes, utiles à l'évaluation et à l'amélioration des procédures et des services. Ils pourront être utilisés en fonction du degré de maturité, et de l'échelle de production propres à chaque établissement. Dans tous les cas, ils favorisent l'accompagnement au changement en objectivant, voire en banalisant une activité nouvelle selon une approche qui est comparable à ce qui se pratique dans d'autres secteurs de la bibliothèque. D'autres sont à mettre en place spécifiquement en fonction des environnements juridiques (indicateurs liés aux demandes d'autorisation de collecte auprès des éditeurs…) ou culturels (indicateurs sur la langue des documents collectés…) propres à chaque établissement.

- **Des indicateurs pour les décideurs**

La préoccupation principale des décideurs est d'ordre économique. Il convient donc de définir des indicateurs relatifs au coût et à la valeur des archives. Pour toute institution engagée dans un processus de négociation (pour obtenir des moyens dédiés à cette activité) ou de lobbying vis-à-vis de sa tutelle ou du législateur (pour obtenir le vote d'une loi les autorisant à archiver le Web), l'exercice consiste donc à démontrer que l'archivage du Web est une activité qui coûte moins cher qu'elle ne rapporte.

Le coût des archives

Si la bibliothèque externalise tout ou partie de sa production auprès d'un prestataire extérieur, on prend en compte le coût facturé. Si la production est internalisée, la dépense peut être décomposée en quatre catégories : La première concerne le coût d'investissement, de maintenance et d'amortissement des matériels informatiques utilisés pour la collecte, l'indexation et le stockage des données : serveurs de collecte et d'indexation, baies ou entrepôts de stockage, quote-part des archives du Web dans l'entrepôt numérique de la bibliothèque, à calculer en fonction de leur volumétrie. La seconde concerne le coût de fonctionnement de ces machines, c'est-à-dire le courant électrique et l'utilisation des réseaux (bande passante) qui sont consommés pour assurer le traitement des données.
La troisième concerne les coûts logiciels. La plupart des bibliothèques utilisent les logiciels libres et gratuits développés par le consortium IIPC, ce qui réduit foncièrement la facture

logicielle par rapport à d'autres applications. Néanmoins, ce coût n'est pas nul et peut même s'avérer élevé, notamment dans les phases d'installation et de mise à jour des versions. Les logiciels doivent en effet être adaptés à l'environnement informatique de chaque institution, éventuellement traduits dans la langue locale ou complétés par des interfaces utilisateurs correspondant aux besoins locaux. Ceci implique des compétences et une activité régulière de développement et d'intégration logicielles.

Le dernier poste budgétaire à évaluer est le coût humain, qui concerne aussi bien la supervision de la production (ingénieurs, administrateurs et techniciens informatiques) que le travail de prospection, de sélection et de traitement (le cas échéant : catalogage, valorisation…) documentaire. C'est sur ce point précis que le modèle retenu pour la collecte des sites fera toute la différence. Les collectes larges, gourmandes en matériel, en traitement informatique et en stockage, s'avèrent très peu coûteuses en personnel, surtout lorsqu'on rapporte ces coûts au nombre de documents collectés. La Bibliothèque nationale d'Islande a ainsi collecté plus de 380 millions de fichiers depuis 2004 et capture chaque semaine plus de 268 gigaoctets de données avec seulement deux personnes à temps partiel sur ce projet.

A l'inverse, les collectes sélectives, qui impliquent la mobilisation de nombreux bibliothécaires, coûtent cher en personnel alors qu'elles rapportent comparativement moins de documents que les collectes larges. Lors de la couverture de la Web-campagne française de 2007, la BnF a procédé à une évaluation fine des charges de travail induites par cette opération qui a mobilisé près de 25 personnes pendant 8 mois. Ce calcul a permis de définir un coût moyen de 51 € par site sélectionné, 90% de ce chiffre étant constitué par le coût humain. L'information est doublement utile. Elle permet d'abord de valoriser le travail accompli par les agents (souvent sur la base du volontariat). Elle permet ensuite de prendre la mesure du coût réel d'un tel effort et d'en tenir compte dans la programmation de futurs projets de ce type, en démontrant ainsi à la bibliothèque qu'un projet aussi coûteux doit être réservé à des actions dont le niveau de priorité – la valeur documentaire- a été clairement établi.

La British Library, qui, pour des raisons juridiques, se limite depuis 2004 à la collecte ciblée d'une sélection de sites présentait en avril 2008 les chiffres suivants : sur 6500 sites identifiés par une équipe de 4 bibliothécaires en trois ans, seuls 1800 ont pu être archivés (les autres n'ayant pas reçu l'accord des éditeurs), pour constituer une collection d'un volume total de 840 gigaoctets. John Tuck, Directeur des collections britanniques a fait état des résultats d'une étude économique[8] conduite par la British Library pour démontrer l'économie que permettrait l'adoption d'une loi supprimant l'obligation de demander aux éditeurs de sites l'autorisation de collecter leurs ressources. Selon cette étude, sans modification de la loi, le coût par téraoctet est évalué à 6 476 £ et permettrait d'assurer la couverture de 0,6% du domaine britannique en 10 ans. Si la loi était modifiée, ce coût serait ramené à 215 £ par téraoctet et permettrait d'assurer la couverture de 81% du Web britannique en 10 ans. Ces chiffres témoignent à la fois du coût relatif d'une politique exclusivement basée sur la sélection humaine et de l'efficacité des indicateurs utilisés : le coût total par téraoctet et le taux de couverture du domaine national[9].

[8] John Tuck, présentation à l'Assemblée générale du consortium IIPC « Web Archiving at the British Library: How Many ? How Much ?» Canberra, Australie, le 7 avril 2008.

[9] Pour efficace qu'il soit, on peut être plus réservé sur ce second indicateur, dont le calcul suppose de pouvoir mesurer le périmètre exact du domaine national relevant de la compétence d'un Etat, ainsi que son évolution dans le temps. Pour beaucoup de pays, dont le domaine national ne se réduit pas aux noms de domaines enregistrés sous leur Top Level Domain (TLD, comme le .fr pour la France ou le .ca pour le Canada, par exemple), l'exercice est quasiment impossible.

La valeur des archives

Convaincre un décideur de la valeur patrimoniale des archives du Web (ou tout simplement du Web) n'a rien d'évident. On l'a dit, Internet véhicule nombre de contenus dont la valeur patrimoniale peut paraître discutable. Faut-il collecter les (très nombreux) sites pornographiques, les sites publicitaires et toutes les lettres d'information qui se déversent dans nos boîtes à lettres? Les plates-formes de blogs, de vidéos où des internautes ordinaires racontent leur vie intime, leurs vacances, leurs amours, leurs hobbies, sont-elles bien du ressort d'une bibliothèque ? Et collecter ces contenus ne peut-il lui attirer des ennuis? Quelques cas se sont présentés (au Danemark, aux Etats-Unis) où des hébergeurs de sites ont intenté un procès à la bibliothèque parce que son robot avait consommé une part important de sa bande passante ou compromis son activité commerciale. La justice elle-même ne pourrait-elle se retourner contre la bibliothèque si celle-ci archivait automatiquement ou délibérément des sites tombant sous le coup de la loi, sites pédopornographiques ou incitant à la haine raciale ? Toutes ces questions, chacun d'entre nous a eu à y répondre un jour, et la réponse est toujours la même : dire que les archives valent la peine de prendre de tels risques, que leur valeur patrimoniale est considérable et digne d'investissements. En ce cas, la valeur que l'on accordera aux archives du Web ne sera pas tant une valeur économique qu'une valeur d'usage et une valeur symbolique.

L'usage, ce grand absent

Peu d'institutions sont aujourd'hui en mesure de proposer un accès public à leurs archives et celles qui y sont parvenues le font souvent dans un cadre contraint qui en réduit l'impact. Depuis 2007, Bibliothèque et Archives Canada propose un accès en ligne depuis son site Web à la plupart de ses archives, mais celles-ci restent limitées aux sites gouvernementaux pour lesquels il a été relativement aisé d'obtenir une autorisation de collecte et de publication. A l'inverse, la BnF donne depuis Avril 2008 accès à la totalité de ses archives sans autorisation préalable des éditeurs, mais la consultation en est fortement restreinte puisque réservée à des chercheurs accrédités, dans l'emprise de l'établissement. Seule la fondation Internet Archive propose un accès en ligne à la plus grande collection du monde, s'assurant de fait une forte popularité auprès des internautes et de ses donateurs. Internet Archive publie ainsi très régulièrement les chiffres mesurant les consultations de son site www.archive.org , soit entre 150 et 200 requêtes par seconde…La valeur des archives du Web ne pourra réellement être attestée que lorsque l'usage en aura été facilité et développé auprès du public et qu'il sera possible de le mesurer à une échelle significative. C'est pourquoi la mise en ligne des collections est aujourd'hui une priorité pour nombre d'institutions confrontées à des obstacles techniques et surtout juridiques pour parvenir à ce but. Souvent, elles se trouvent enfermées dans un cercle vicieux : pour obtenir un cadre juridique facilitant cette communication, elles doivent convaincre les décideurs de la valeur d'un objet qui reste souvent invisible du grand public…faute d'une législation favorable.

La rareté et la valeur de l'éphémère

C'est donc sur d'autres registres qu'il faut jouer pour démontrer cette valeur. La rareté est un de ces registres. Plusieurs bibliothèques ont pris l'habitude de mesurer le taux de disparition des sites Web qu'elles archivent. Par exemple, la BnF calcule régulièrement le taux de présence en ligne des sites Web qu'elle a archivés lors des Web-campagnes nationales.
Plus d'un tiers des sites archivés lors de la campagne présidentielle de 2002 n'existaient plus cinq ans plus tard – ce taux ne tient pas compte des sites devenus inactifs après la campagne. Ces chiffres s'avèrent très efficaces et donnent immédiatement à la collection la valeur de l'éphémère, si bien que pour désigner les collections les plus anciennes de ses archives (1996), la BnF évoque souvent ses « incunables » du Web français.

En conclusion : éloge de l'analogie

C’est finalement grâce au symbole et à la métaphore que les bibliothécaires ont peut-être le plus de chance de séduire et convaincre leurs décideurs.

Une première démarche consiste à traduire les unités logiques propres aux archives du Web en des unités physiques intelligibles par votre interlocuteur. Certains ont ainsi suggéré que l’on mesure les collections du Web en heures d’émissions de radio ou de télévision ou en kilomètres linéaires ou carrés de pages imprimées. A la BnF, la collection électorale de 2007 représente une pile de 6 000 CD. Et nous ne manquons pas d’indiquer régulièrement que chaque instantané annuel de notre domaine national rapporte autant de documents que la BnF n’en a réunis en cinq siècles, ce qui est évidemment très contestable.

Une seconde démarche, sur laquelle je me suis appuyé durant la totalité de cet exposé, consiste à user plus généralement de l’analogie avec les activités et les missions traditionnelles de la bibliothèque au lei d’aborder l’innovation comme une révolution. Il suffit de s’appuyer sur des exemples concrets et des histoires, qui démontrent que malgré les apparences, rien ne change vraiment. Les fichiers du Web sont des documents et constituent des collections. La collecte du Web à grande échelle s’apparente à du dépôt légal, les captures sélectives à des acquisitions. L’archivage numérique des données correspond à leur conservation et à leur magasinage. L’accès aux archives s’intégrera aux pratiques de consultation et de recherche. C’est pour les mêmes raisons qu’il n’y a pas lieu d’isoler les équipes en charge de cette nouvelle activité mais qu’il faut viser à leur intégration et à leur banalisation dans les organigrammes existants.

Notre connaissance, notre souvenir et notre utilisation de l’histoire et des histoires de la bibliothèque peuvent nous aider à la mettre à l’abri de l’oubli du futur en l’aidant à reconnaître dans des objets nouveaux des signes familiers. De ce point de vue, la transmission entre générations de professionnels est devenue un point plus critique que jamais, et c’est un facteur important à prendre en compte du point de vue de l’accompagnement du changement : faire en sorte que les jeunes qui portent l’innovation n’aient pas l’illusion d’être les premiers à le faire et, d’une certaine manière, sachent aussi être vieux.