L'archivage des données du Web (2ème partie : interviews)

Retour au sommaire de la lettre

Suite à l'article que nous avons publié le mois dernier sur l'archivage des données du Web, nous avons interrogé plusieurs responsables de grandes bibliothèques européennes pour connaître leur point de vue sur ce sujet. L'archivage et la mise à disposition des documents constituent en effet le cœur de métier des grandes bibliothèques nationales et ces institutions sont de plus en plus à l'origine de grands projets d'archivage du Web.
Jean-Noël Jeanneney (Président de la Bibliothèque nationale de France), Mark Middleton (l'un des responsables du UK Web Archiving Consortium) et trois responsables de la Bibliothèque nationale suédoise ont accepté de répondre à nos questions...

Jean-Noël Jeanneney – Président de la Bibliothèque nationale de France (www.bnf.fr)

Pouvez-vous nous décrire les projets actuels de la BnF en matière de numérisation et d'archivage du Web francophone ?
Dans le prolongement de ses missions de conservation patrimoniale et de dépôt légal, la BnF se doit d'assurer la préservation des contenus électroniques francophones et français publiés sur la Toile.
L'Assemblée nationale examinera, début juin 2005, un projet de loi de transposition d'une directive communautaire sur le droit d'auteur et les droits voisins dans la société de l'information, qui prévoit de confier à la Bibliothèque une responsabilité essentielle dans ce domaine.

Quels types de contenus seront archivés en priorité ? Selon quels critères ?
Il faut souligner que les travaux préparatoires à l'archivage de la Toile ont commencé à la BnF dès 1999. Ils nous ont conduit à adopter une stratégie en deux axes :
- d'une part la collecte automatisée dite "de surface" d'un périmètre représentatif de la production nationale ;
- d'autre part un archivage thématique "profond" plus sélectif, faisant appel à des collectes ciblées complétées par le dépôt des éditeurs de sites lorsque cela s'avérera nécessaire pour des raisons techniques, commerciales ou légales (accès protégés par mot de passe ou soumis à des contraintes de consultation particulières, etc.).
En effet dès lors que le dépôt est légal, les éditeurs doivent répondre à la sollicitation de la BnF (à la différence de la situation qui prévaut aux Royaume Uni). Cependant nous nous efforçons de limiter leur charge à ce qui ne peut être réglé par une collecte automatique.

Quel mode d'indexation pensez-vous adopter ?
Le mode d'indexation envisagé offrira des possibilités d'accès par URL, par date et par période d'archivage, ainsi qu'une recherche plein texte. A cet effet, la BnF participe à la mise au point d'un outil dans le cadre des travaux du consortium IIPC évoqué plus bas. Il s'agit d'un développement qui reprend, en l'internationalisant et en l'enrichissant, une interface d'accès développée par les pays scandinaves. Le consortium a par ailleurs opté pour l'intégration de Nutch et Lucene comme moteurs initiaux d'indexation. L'ensemble du système est conçu pour traiter des milliards de pages, avec une architecture ouverte sur les nécessaires évolutions qu'imposeront la croissance de la Toile et l'enrichissement de ses contenus. Les lecteurs qui désirent se familiariser avec ces travaux peuvent dès aujourd'hui consulter un site de démonstration (http://nwa.nb.no/demo/search.php).

Quel public aura accès aux sites archivés et comment ?
C'est la prochaine loi sur le droit d'auteur et les droits voisins et son décret d'application qui fixeront les règles d'accès aux archives électroniques. Notre mission n'est pas de nous substituer aux éditeurs à qui il revient d'assurer, s'ils en ont la volonté et les moyens, une diffusion pérenne des contenus auprès de tous les publics ; en revanche, c'est à la

Bibliothèque d'en garder la trace durable à des fins d'étude et de recherche. Les lecteurs de la BnF devraient ainsi être les premiers bénéficiaires de ces archives. Nous nous attacherons également, autant que cela sera possible, à encourager les productions de contenus dérivés tirant le meilleur parti de ces ressources numériques, dont l'intérêt donnera naissance à des innovations éditoriales et de services qu'on ne peut envisager aujourd'hui.

Quels sont les principaux obstacles aux projets actuels en France ?
Je n'en vois guère d'insurmontable. Le travail préparatoire et les expérimentations conduites au cours de ces trois ou quatre dernières années nous ont permis de dessiner une cartographie de la Toile définissant un ensemble national, de repérer les fréquences de mise à jour des sites collectés et leur durée de vie moyenne, d'identifier les principales données inaccessibles de façon automatisée et dont l'intérêt justifierait un dépôt volontaire.
Au-delà de l'expertise acquise et de l'affinement des méthodologies, ce sont déjà près de trente tera-octets d'archives qui ont été rassemblés par la BnF.
Lorsque le législateur aura définitivement arrêté les règles, il nous restera à adapter les moyens techniques aux exigences de notre mission. En matière d'ingénierie il conviendra par exemple d'élargir la bande passante de l'accès au réseau Internet de nos robots d'indexation et, bien sûr, d'anticiper une politique à long terme pour le développement de nos capacités de stockage de l'information numérisée.

Quel rôle doivent selon vous jouer les institutions européennes de conservation du patrimoine dans l'archivage du Web ?
Pour chacune des institutions se pose nécessairement la question de la continuité de sa mission dans l'espace de la Toile. Les modalités d'exercice de cette extension de compétence dépendent du contexte local particulier en matière de législation, d'expertise, de moyens et d'usage. Les contenus en réseau ne connaissant pas de frontière, ils appellent à une nécessaire collaboration pour coordonner et optimiser les efforts afin d'assurer le meilleur archivage du web possible.

Peut-on espérer à terme une interopérabilité des systèmes d'archivage des grandes bibliothèques nationales européennes ?
L'interopérabilité est un des objectifs d'une collaboration indispensable, non seulement dans l'espace européen, mais de manière universelle entre les acteurs majeurs de l'archivage de la Toile. C'est la raison pour laquelle la BnF pilote et coordonne le Consortium International pour la Préservation d'Internet (IIPC - http://www.netpreserve.org/) créé en 2003. Ce Consortium rassemble aujourd'hui la Bibliothèque du Congrès, la British Library et les bibliothèques nationales d'Australie, du Canada, du Danemark, de la Finlande, d'Islande, d'Italie, de Norvège, de Suède, ainsi que la fondation américaine Internet Archive. Il a bien sûr vocation à s'ouvrir à de nouveaux membres qui le rejoindront dès qu'ils en ressentiront le besoin.

Pour quand est prévu le lancement officiel de votre système d'archivage du Web francophone ?
Sitôt que la loi sur le dépôt légal électronique sera promulguée nous pourrons mettre en place dans des échéances très brèves l'officialisation des dispositions que nous avons anticipées, et rendre ainsi accessibles aux chercheurs nos archives du web francophone.

Que pensez-vous des initiatives actuelles des grands moteurs de recherche en matière d'archivage du Web et de numérisation des ouvrages de bibliothèques ?
A ce jour les moteurs de recherche n'archivent les contenus de la Toile que de manière temporaire, à des fins techniques d'indexation. L'enjeu d'un établissement patrimonial est tout autre, puisqu'il lui incombe d'assurer la conservation pérenne des données et leur mise à disposition.
S'agissant de la numérisation des livres, les différents acteurs privés -- par exemple Google avec Google Print ou Amazon avec son moteur A9 -- ont fait des annonces ou des expérimentations à petite échelle de projets dans des registres commerciaux qui, aussi éloignés qu'ils puissent être de la logique des bibliothèques, n'en sont pas moins stimulants pour la réflexion sur l'avenir de l'écrit à l'écran.

Enfin, quels sont vos projets pour la numérisation des ouvrages écrits et leur mise à disposition sur le Web ?
Vous me permettrez de vous renvoyer à mon dernier ouvrage [1] que la presse a abondamment commenté, et de revenir sur un point particulier. A l'heure du web on ne peut plus penser la numérisation des bibliothèques de manière isolée et nationale. Ce qui, il y a

vingt ans, pouvait apparaître comme un formidable défi -- créer une grande bibliothèque numérique francophone -- paraît aujourd'hui une approche surannée. La francophonie et le patrimoine français ne peuvent s'enrichir et assurer leur bonne diffusion que dans une mise en valeur et une confrontation en bonne intelligence avec les cultures européennes. Je suis donc très heureux que mon appel ait été aussi largement entendu et repris, et que toutes les bibliothèques de l'Union européenne s'engagent avec détermination dans la mise en place d'une bibliothèque numérique commune.

[1] - Quand Google défie l'Europe, Plaidoyer pour un sursaut - Mille et une nuits.

Pour commander ce livre sur Amazon :
http://www.amazon.fr/exec/obidos/ASIN/2842059123/171-7281317-9061039

Mark Middleton – Web Archiving Programme Manager (www.webarchive.org.uk)

Qu'est-ce que le Web Archiving Consortium et quand a-t-il été créé ?
Le UK Web Archiving Consortium (UKWAC) est un partenariat entre différents organismes dont le but est de mettre au point un banc d'essai pour l'archivage sélectif des sites internet britanniques. UKWAC a vu le jour en juin 2004 et le projet va durer initialement 2 ans.

Qui est à l'origine du projet et quelles sont les bibliothèques partenaires ?
La British Library, le Wellcome Trust et le JISC (Joint Information Systems Committee), un organisme lié à l'enseignement supérieur, en sont à l'origine. Les Archives Nationales, la Bibliothèque Nationale du Pays de Galles et la Bibliothèque Nationale d'Ecosse les ont ensuite rejoints.
C'est suite à une étude sur la faisabilité de l'archivage du Web menée en février 2003 par le Wellcome Trust et JISC qu'un projet d'archivage en commun a été fortement recommandé.
D'autres renseignements sont disponibles sur :
http://library.wellcome.ac.uk/node228.html ou
http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

Quels sont les objectifs de ce partenariat ?
UKWAC, réunion de six organismes majeurs du Royaume-Uni, a pour ambition d'archiver une liste sélective de sites britanniques dans le respect des droits de propriété. Son but est de garantir la conservation pour les générations futures d'inestimables ressources culturelles ou travaux de recherche.
Chaque membre du consortium va sélectionner et collecter les contenus relatifs à son sujet / domaine.
1) La British Library a la charge des sites nationaux dans les domaines de la culture, de l'histoire et de la politique. Cela concerne par exemple les pages Web traitant d'évènements clés dans la vie nationale, par exemple les élections. De même, les pages Web des musées, les sites récompensés par des prix, certains blogs qui traitent de la recherche ou qui présentent les projets littéraires et artistiques des citoyens britanniques pourraient être conservés.
2) La Bibliothèque Nationale d'Ecosse conservera les sites ayant trait à la culture et à l'histoire de l'Ecosse.
3) La Bibliothèque Nationale du Pays de Galles conservera les sites reflétant la vie contemporaine du pays de Galles.
4) Le Wellcome Trust s'intéressera principalement à la médecine en archivant les sites britanniques liés à la santé et à la médecine. Cela inclut les organismes de recherche, les associations caritatives, les associations professionnelles, organismes de certification et les groupes de pression.
5) Les Archives Nationales sélectionneront et archiveront certains sites en rapport avec six domaines précis de la politique gouvernementale :
- La politique étrangère et de défense.
- La justice et la sécurité intérieure
- La gestion des ressources nationales
- L'offre de services non fournis par le marché (santé, éducation, culture)
- La régulation et la coordination des services offerts par le marché
- Le fonctionnement des administrations gouvernementales et délocalisées.
6) JISC conservera les sites les plus innovants des projets TIC menés dans l'Enseignement Supérieur.

Quelles sont selon vous les principales difficultés qui peuvent être rencontrées lorsqu'on archive le web ?
Les données que l'on trouve sur le Web sont fragiles, éphémères et peuvent disparaître soudainement. L'archivage de données Internet est une activité récente pour les bibliothèques du Royaume-Uni. L'un des objectifs d'UKWAC est de tester la viabilité de cet archivage des données Web au Royaume-Uni.
Il y a quatre difficultés principales à surmonter :
1) La valeur des données du Web tient à leur haut niveau d'actualité. Cependant, cela fragilise certaines informations potentiellement intéressantes car elles sont très souvent remplacées

par des informations plus récentes. Il est donc nécessaire de les capturer dès leur publication sous peine de les perdre à jamais.
2) Il n'existe pas de standard unique pour les formats de fichiers ou pour la construction de sites Internet. C'est pourquoi la technique d'archivage doit être adaptée à chaque site en fonction de ses caractéristiques et de sa complexité propre. Cela nécessite du temps et une haute technicité.
3) Internet est un moyen de communication flexible, dynamique et accessible à tous. On y trouve beaucoup de formats de fichiers variés, qu'il s'agisse de texte, d'image, de son ou de films. Certains formats sont propriétaires et nécessitent des applications spécialisées supplémentaires pour permettre de visualiser ou d'utiliser le fichier. Celles-ci sont quelquefois payantes. De même certains formats de fichiers sont obsolètes après quelques années et nécessitent un traitement spécialisé pour permettre leur migration dans un format plus courant. Dans certains cas rares, certains procédés de construction de sites, notamment l'utilisation de certains types de Java, rendent l'archivage d'une copie du site impossible.
4) A long terme, conserver un site dans l'état où il était lors de son archivage nécessite une gestion rigoureuse et planifiée. Chaque site archivé est unique et ne peut pas être remplacé s'il est endommagé ou perdu. La fonctionnalité à long terme de certains formats de fichiers et même la viabilité future du protocole HTTP soulèvent encore des questions.

Que pensez-vous des initiatives d'archivage du web lancées par les moteurs de recherche ?
Le but des moteurs de recherche est de fournir à l'utilisateur un moyen facile de trouver sur Internet une information fraîche et pertinente. Globalement, ils fournissent des index généraux des contenus hébergés sur des sites actifs. Au fur et à mesure que de nouveau sites et de nouvelles données apparaissent, les vieux index sont réécrits et remplacés. Même avec les meilleurs moteurs de recherche, certains sites ou certaines pages restent inaccessibles. Certains moteurs permettent l'accès à un cache limité, c'est-à-dire à une archive de pages ou de sites qui ne sont plus disponibles. Cependant ces sites ne sont pas organisés ou archivés systématiquement dans l'optique d'une conservation à long terme.
UKWAC a un objectif différent. Nous ne sommes ni un moteur de recherche, d'indexation, ni même un portail d'informations. UKWAC structure un système d'archivage recensant des sites Web sélectionnés avec une gestion sur le long terme pour que la recherche future puisse en profiter. Les sites sont choisis en fonction de politiques permettant d'évaluer leur pertinence future. Le fait de sélectionner avec précision les sites nous permet de les archiver en les altérant le moins possible et en leur laissant un haut degré de fonctionnalité. De plus, cela nous permet de ne dupliquer que l'information nécessaire et d'en assurer d'autant plus aisément la pérennité.

Selon vous, quelle est l'importance de l'archivage des données du Web pour les bibliothèques ?
De plus en plus, Internet est perçu, en matière d'édition (surtout professionnelle), comme un lieu de choix. C'est pourquoi les bibliothèques et les centres d'archives sont bien inspirées lorsqu'ils acquièrent, gèrent et mettent à disposition du public du contenu numérique. L'augmentation du contenu numérique et la nécessité de le gérer efficacement représente, d'une certaine manière, un prolongement de l'activité des bibliothèques. Beaucoup d'entre elles fournissent d'ailleurs déjà un accès à des journaux en ligne. Alors que les méthodes d'édition traditionnelles évoluent, les bibliothèques ont, depuis plusieurs années, compris l'importance de la gestion du contenu numérique.
En évaluant la viabilité de l'archivage du web en Grande-Bretagne, UKWAC espère montrer la voie à d'autres acteurs.

Johan Mannerheim, Allan Arvidsson et Krister Persson – Kungliga biblioteket – Suède – www.kb.se)

Qu'est-ce que KulturarW3 ?
Le projet ambitionne de conserver la contribution suédoise au Web mondial. Il ne concerne pas le dépôt légal des supports numériques.
Vous remarquerez que le nom du projet est Kulturarw3, avec un "w" puissance 3 comme "www". "Kulturav" signifie "patrimoine culturel" en suédois.

Qui est à l'origine du projet et quand a-t-il été lancé ?
La Royal Library, la Bibliothèque nationale de Suède, en est à l'origine. La bibliothèque conserve depuis 1661 tout ce qui est publié en Suède et soumis au dépôt légal. En 1995 et 1996 nous avons eu une réflexion interne sur le volume grandissant du contenu publié sur internet. Comme celui-ci est volatile et évolue rapidement, il nous a semblé qu'il fallait que quelqu'un s'occupe de le conserver. La Royal Library a démarré le projet en 1996 et depuis 1997, une collecte régulière est en place. Au départ, lorsque la Royal Library s'est attelée à réduire la perte du patrimoine culturel contemporain en collectant de manière globale les pages suédoises, il n'y avait pas de cadre légal. Cette activité concernant les pages web s'inscrivait dans le cadre de la loi suédoise sur le dépôt légal.

Quels contenus sont archivés et quels sont vos critères de sélection ?
Nous essayons de sélectionner tout ce qui est « suédois ou présente un intérêt pour la Suède ». Nous ne regardons que la partie serveur de l'URL. Tout ce qui appartient au domaine suédois « .se » est archivé. Les domaines « .nu » sont aussi populaires en Suède car « nu » signifie « maintenant ». Dans ce cas, une liste des noms de domaine « .nu » appartenant à des personnes ou à des organisations suédoises nous est adressée par les gens qui gèrent ces domaines. Pour les autres noms de domaine, nous vérifions que le serveur est basé en Suède et si c'est le cas nous les incluons. Auparavant, nous utilisions la base de données « Whois » mais cette base ne nous est plus accessible. Nous nous basons dorénavant aussi sur les adresses IP. Nous venons juste de commencer à utiliser ce dernier critère et n'avons pas encore évalué la valeur de ce dernier critère.
Nous avons deux archives : la première où nous collectons tout ce qui est suédois selon les critères ci-dessus. La seconde qui concerne tous les journaux conservés chaque jour.
Nous n'effectuons pas d'autres sélections, en principe. Cependant il arrive que nous excluions certains contenus pour des raisons techniques, par exemple lorsque certains serveurs causent des problèmes.
Nous n'utilisons pas de critères thématiques ou évènementiels, par exemple nous n'avons pas de catégorie concernant le 11 septembre.

Comment peut-on consulter l'outil KulturarW3 ?
L'accès est restreint par la loi sur la protection de la vie privée. Le gouvernement autorise la consultation des archives à des fins de recherche et d'étude mais simplement dans les murs de la Royal Library. Une interface a été mise au point qui permet de rechercher des sites web, de surfer dans les archives et de naviguer entre les différentes sauvegardes dans le temps.

Quelles sont les difficultés principales que vous rencontrez lors de l'archivage des sites web ?
Les sites interactifs posent problème. Un outil de collecte [harvester] ne peut pas utiliser les bons mots-clés. Les scripts et d'autres techniques de liens non html posent aussi des difficultés. La plupart de ces difficultés sont surmontables mais il existe tant de versions différentes qu'il faut procéder au cas par cas.
Les serveurs Web mal configurés. En particulier ceux qui ne fournissent pas de code d'erreur lorsqu'une fausse URL a été demandée.

Merci à tous pour vos contributions !