World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme: http://www.ifla.org/IV/ifla71/Programme.htm

July 11, 2005

Code Number: 074-F
Meeting: 133 SI - Bibliography

La collecte automatique du web : l'expérience de la Bibliothèque nationale de France

Christian Lupovici
Directeur de l'Agence Bibliographique nationale
Bibliothèque nationale de France
Christian.lupovici@bnf.fr

Résumé :
La Bibliothèque nationale de France, dans le cadre de sa mission de dépôt légal, expérimente les procédures de collecte du web et organise la préservation à long terme des documents numériques. Le travail effectué pour accomplir cet objectif inclus des réflexions fondamentales sur l'essence même du dépôt légal et sur le traitement bibliographique des ressources de l'Internet. Travailler en grandeur réelle, sur l'énorme masse des ressources numériques, est un facteur important d'aide à la décision pour la Bibliothèque nationale.

1. Le dépôt légal à la BnF

1.1 Rappel historique

La législation française sur le dépôt légal remonte au 16e siècle quand François 1er prit l'Ordonnance de Montpellier en 1537. Depuis cette date, l'objectif de toutes les réglementations postérieures fut de prendre en compte chacune des nouvelles technologies qui apparaissaient et de les intégrer dans les dispositions légales, élargissant ainsi chaque fois le champ du dépôt légal.

1537 : les livres ;
1648 : les estampes dont les cartes et plans ;
1793 : les partitions musicales
1925 : les photographies et toute production d'art graphique et les phonogrammes ;
1941 : les affiches ;
1963 : le son enregistré de toute nature ;
1975 : les enregistrements d'images fixes ou animées quel qu'en soit le procédé ;
1977 : les œuvres cinématographiques ;
1992 : tout type de documents sur support, y compris les bases de données et les systèmes experts pour peu qu'ils soient distribués au public.

L'introduction des ressources du web dans la législation française devrait arriver en 2005 puisque cette question doit être discutée au parlement.

1.2 Mission et champ d'application du dépôt légal

Il est clair que le but du dépôt légal est de préserver et de donner accès sur le long terme, à l'ensemble du patrimoine culturel des documents produits et distribués en nombre. Le dépôt légal comprend toute sorte de ressources et n'est pas restreint aux documents en français ni aux seuls documents publiés sur le territoire français.
C'est plus une question de type de contenu et/ou de canaux de distribution (livres, estampes, cartes, son, audiovisuel) qu'une question de support : papier, disque ou en-ligne. Sur l'Internet, la plupart de la littérature grise institutionnelle sera dans le champ de la réglementation du dépôt légal pour autant que les documents soient accessibles sur le web par le public, donc publiés.
Le terme dépôt légal signifie aussi que la collecte n'est pas une acquisition au sens d'entrant dans une politique de développement des collections, mais la collecte de documents de toutes sortes qui donne une image des contenus distribués au public français d'aujourd'hui, pour les conserver et s'en servir dans l'avenir pour des études historiques. Ceci veut dire que toutes les versions différentes d'un même document aussi bien que des oeuvres interdites, sont concernées comme elles l'on toujours été depuis le 16^{e} siècle.
Comment transposer tout ceci aux ressources du web ?

2. La collecte automatique du web

2.1 Le moissonnage de l'Internet dans l'objectif du dépôt légal

La transposition de cette philosophie du dépôt légal aux ressources du web implique :
a) D'exclure tous les noms de domaine qui de toute évidence, ne contiennent pas de ressources appartenant à la culture française (les domaines géographiques hors la France) ;
b) De se concentrer sur les noms de domaines nationaux (*.fr, .pf, .wf, .pm, .re, .tf, .ad, .yt*) ;
c) De collecter tout le reste (les noms de domaines génériques).

Dans une première approche il n'est pas possible de savoir ce qui est pertinent pour la culture française dans ce reste et un travail supplémentaire doit être conduit pour trouver un procédé de sélection rapide fondé sur une analyse sémantique sur le contenu des sites. Néanmoins puisque la question du stockage n'est plus un problème crucial, la Bibliothèque national de France stocke et préserve un ensemble de sites bien plus large qu'il serait strictement nécessaire pour les besoins du dépôt légal en attendant une meilleure solution. De toute façon, la part du web qui sera stockée et conservée sera toujours large que la part exclusivement pertinente.

2.2 De l'expérimentation au travail en grandeur réelle

2.2.1 Les premières expérimentations

La Bibliothèque nationale de France collecte des sites web français depuis 2001[1] en utilisant le robot HTTrack. Les premières expérimentations ont été centrées sur les différentes campagnes électorales en France. Ces expérimentations avaient pour objectif de tester différents points :
a) Comment construire une typologie des sites web et en fonction de celle-ci, quelle devait être la fréquence de collecte ;
b) Tester les problèmes techniques (formats, profondeur de la collecte) et le réglage du robot ;
c) Evaluer les procédures manuelles et automatiques de sélection des sites web.

Cette phase fut importante pour évaluer la sélection automatique d'ensembles de sites à traiter manuellement et comment utiliser des outils d'aide à la décision.

2.2.2 La collecte automatique du Web en grandeur réelle

Que représente le dépôt légal français du Web en terme de ressources et de volume à collecter ?

[1] http://bibnum.bnf.fr/conservation/aristote2004/roche.pdf

Une « vue » extraite de la totalité du Web durant le printemps 2004 par le robot Alexa, fut effectuée selon les spécifications, sur les noms de domaines à parcourir (*.fr, .com, .net, .edu, .com, .biz, .info, .int, .pf, .ad, .coop, .name, .aero, .tf, .re, .museum, .pro, .pm, .wf, .yt*) et en rejetant ceux qui devaient être exclus. Le résultat de cette collecte ramena plus de 2 milliards d'URLs et 65 millions de serveurs hôtes, le tout pesant 27 téraoctets.

En complément, une collecte ciblée du domaine *.fr* fut effectuée à la fin de 2004 et en janvier 2005, en utilisant le robot Heritrix. Le résultat de cette collecte ramena 500 000 sites publics et 4 000 sites personnels. Cette collecte équivaut à plus de 118 millions d'URLs et un demi-million de serveurs hôtes, pesant 3 téraoctets.

Pour ces volumes de données, il est inutile d'imaginer un quelconque traitement manuel de catalogage ou d'indexation sinon à la marge. Un traitement manuel pour ce volume de données et seulement pour les sites publics, prendrait au moins 500 hommes/jour à trier et cataloguer. C'est deux fois plus que l'équipe actuelle qui catalogue la bibliographie nationale à la BnF.
D'où la nécessité de combiner différentes techniques.

2.3 Le traitement du dépôt légal du Web

La méthode fondamentale est la collecte automatique. La totalité de la surface du Web et certaines parties du web profond peuvent être collectées automatiquement, en fonction des problèmes techniques auxquels le robot est confronté et en fonction de son paramétrage. Mais en général, une collecte automatique ramènera principalement le Web de surface et créera ainsi une carte du Web.

Une seconde méthode, complémentaire, est de cibler une collecte automatique sur un ensemble de sites, par exemple, le domaine *.fr* et de demander au robot d'effectuer une collecte en profondeur.

Ces deux méthodes vont du haut vers le bas.
Pour obtenir une image complète, il faut combiner ce type de collecte automatique avec une méthode qui part du bas vers le haut et qui consiste à compléter les collections par des ressources du web profond déposées par les producteurs puis de recréer les liens avec le Web de surface qui a été collecté. Ces dépôts devront être traités en partie manuellement, en partie automatiquement. Le dépôt peut présenter des avantages en terme d'assurance de préservation, mais c'est un processus coûteux en temps, en commençant par l'analyse du site qui prend 5 hommes/jour en moyenne.

L'ensemble des données est stocké à la BnF, dans un « magasin » central, ce qui permet, en interne, une utilisation et une appropriation par les professionnels. L'archivage du web fait partie du programme SPAR (*Système de Préservation des ARchives*) de stockage numérique et de préservation, lancé en 2004 et qui est global à la BnF.

2.4 Question et méthode de sélection

En principe, la Bibliothèque nationale de France aimerait archiver les sites les plus pertinents. Il est dont intéressant de trouver une méthode qui puisse sélectionner ou proposer une sélection parmi les la totalité des sites à collecter. Un projet de recherche appelé « Watson » a été mené sur les possibilités des traitements linguistiques applicables à l'archive du web à la BnF et destinés aux utilisateurs comme aux professionnels. Le projet a duré de 2003 à 2004[2]. L'une des voies qu'il a explorée fut la caractérisation des sites en utilisant l'analyse linguistique des métadonnées et des phrases du texte intégral. Cette méthode a fourni des résumés sur le contenu des sites. Cette méthode n'est pas aujourd'hui assez fiable pour être utilisée sans vérification humaine, mais elle est assez bonne pour constituer un outil d'aide à la sélection de sites.

[2] José Coch, Julien Masanès**. -** Language engineering techniques for web archiving. In : 4th International Web Archiving Workshop (IWAW04), 16 septembre 2004, Bath, UK - http://www.iwaw.net/04/index.html (visited 30/05/2005).

La BnF organise actuellement une équipe de veilleurs composée de professionnels qui travaillent dans les différents départements de la bibliothèque essentiellement ceux qui sont en charge du dépôt légal. L'objectif de cette équipe est d'apprendre comment traiter ce nouvel environnement de ressources numériques et comment faire le lien nécessaire avec les collections traditionnelles antérieures sur support. Des outils d'accès et les méthodes de veille ont été mises au point pour évaluer la qualité du résultat de la collecte automatique. L'équipe dispose d'un outil qui lui permet de faire des vérifications à la fois sur l'archive de la BnF et sur le web actif, de façon à proposer des sites spécifiques à collecter en profondeur par une collecte automatique ciblée. Cet automne, la prochaine collecte automatique prendra en compte le travail fait par l'équipe de veille en incorporant les « graines » (URLs) proposées dans les paramètres du robot.

De toute façon il ne sera jamais possible de sélectionner un ensemble de sites web qui corresponde exactement au dépôt légal français et il est évident que l'archive du web fait par la BnF sera toujours d'une plus large taille. Si tous les pays font de même, il y aura des recouvrements, mais c'est plutôt une bonne chose puisque le stockage n'est plus désormais un problème, et que les doublons dans différents pays assureront une meilleure sécurité pour la préservation à long terme. D'un autre côté, la méthode des « vues instantanées » n'assure pas l'exhaustivité des collections, mais sont un échantillonnage représentatif de la culture actuelle. Ce qui est l'essence du dépôt légal.

Néanmoins, la nécessité d'un « robot intelligent » capable de cibler des ensembles de sites selon des paramètres donnés et capable de descendre en profondeur quand le rang de pertinence est élevé, a été souligné par le consortium international IIPC (International Internet Preservation Consortium)[3]. La British Library et la Bibliothèque nationale de France ont donc lancé ce printemps, un appel d'offres pour un robot moissonneur « intelligent ».

2.5 Nouvel environnement, nouveau document

En traitant le web nous devons abandonner l'approche traditionnelle de traitement des documents et aller vers de nouvelles façons de gérer les ressources documentaires qui ressemblent plus aux techniques archivistiques.
L'archive du web a de nouvelles dimensions:

La navigabilité doit être conservée, elle est donc recréée en interne. Aucun document n'est une ressource isolée. Le document lui-même se transforme. Dans le modèle de la BnF, l'unité documentaire est le site web lui-même (sur le modèle d'une collection). Ce qui était considéré comme un document est devenu un composant du document principal.

2.6 Les techniques linguistiques pour l'analyse des ressources

Face à l'immense masse de données chargées dans le système de stockage, les professionnels comme les lecteurs, doivent disposer d'outils pour chercher l'information dont ils ont besoin. Ils peuvent vouloir rassembler les ressources en sous-ensembles pour chercher une information précise, mais ils peuvent également vouloir traiter une grande masse d'information pour y appliquer des traitements statistiques ou linguistiques.

Le projet Watson a eu pour but de sélectionner des ensembles personnalisés de ressources correspondant à un domaine déterminé, pour leur appliquer des analyses linguistiques ; ce peut être, pour donner une idée du sujet, pour renseigner l'utilisateur sur la structure, les mots-clé, sur les noms et les lieus etc. contenus dans les documents ou en donnant un résumé navigable. Le projet a produit des outils spécifiques à l'utilisateur pour l'extraction du sujet, la navigation contextuelle et l'analyse sémantique du contenu des sites. Ces expérimentations ont été effectuées sur les archives des élections françaises.

[3] IIPC est un consortium de 11 bibliothèques et comprenant aussi Internet Archive (en 2005). Voir : http://netpreserve.org/about/index.php

3. Nouvelles perspectives du catalogage

3.1 Les métadonnées de catalogage

Les métadonnées sont utilisées traditionnellement pour décrire, identifier des objets physiques et pour l'interrogation. Cette activité de catalogage est un travail lourd et totalement manuel, mais elle est adaptée aux objets physiques et on ne peut pas s'en passer dans ce contexte.
Dans l'environnement numérique, la situation est différente :
Les objets numériques sont (ou devraient tous être) prêts pour un traitement automatique de l'information. Ils contiennent leur auto-documentation sous forme de métadonnées qui peuvent être d'ordre technique, administratif et juridique, permettant leur gestion dans un système d'information.

Quand on a affaire à du mode texte, c'est le contenu lui-même qui peut être indexé, trié, sans aucune intervention humaine (à cette étape). Ceci concerne la structure comme le contenu sémantique de l'information.

Dans le contexte du web, les métadonnées de description peuvent être pauvres car elles sont compensées par « l'effet réseau ». L'objet numérique n'est pas un objet isolé dans l'archive du web. Il est lié par des liens explicites à d'autres ressources, à d'autres contenus et à des descriptions de contenus, comprenant des instruments de recherche qu'ils soient traditionnels ou totalement nouveaux.

3.2 Autres types de métadonnées

Avec le document numérique, de nouvelles questions émergent, comme la gestion des droits, comme l'assurance de l'authentification, non seulement pour des raisons juridiques, mais aussi pour assurer la reproduction indéfinie des ressources numériques et leur circulation immédiate. La reconnaissance des formats, qui sont souvent associés dans un même document, est un des problèmes cruciaux car l'information donnée par les serveurs ou contenue dans les documents n'est pas assez fiable. Les documents collectés doivent faire l'objet d'une vérification de leur format.

Dans le contexte d'une Bibliothèque nationale, il est aussi très important d'étendre les métadonnées à l'information de préservation adaptée à l'archive du web.

3.3 L'avenir du catalogage

La situation du catalogage dans le contexte du web peut se résumer à faire les spécifications pour la création des documents numériques et effectuer l'analyse des ressources au moment de la phase d'entrée des données (modèle OAIS) dans le système, y compris les instructions au robot de collecte automatique et à l'interaction avec les producteurs de sites web.

4. Que peut être une Bibliographie nationale du web ?

La législation française sur le dépôt légal (même la nouvelle) dit qu'il y aura une bibliographie nationale des documents déposés.

Mais dans ce contexte, à quoi une bibliographie nationale peut-elle bien ressembler ?
Si l'objet de la bibliographie nationale est de faire connaître les nouvelles publications disponibles, on peut envisager de publier sur le site web de la Bibliothèque nationale de France, la liste des nouveaux sites qui viennent d'apparaître (pas les modifications), en triant automatiquement et en « cataloguant » automatiquement dans la phase de stockage dans l'archive (modèle OAIS). Le résultat en sera une liste d'URLs avec un index des noms, un index des sujets, pour autant que ce soit possible avec une analyse linguistique automatique, proche des traitements qui ont été effectués dans le projet Watson.