PANORAMA MONDIAL DE L'ARCHIVAGE DU WEB

SC0335_09

Au profit de la Cellule Wall-On-Line – Ministère de la Région Wallonne

Auteurs :

Pour SERDA
M. Mehdi GHARSALLAH, Consultant expert
M. Jérôme MONTFORT, Consultant – Responsable d'affaire
Mlle Audrey CHAUSSARD, Chargée d'études

Destinataires :

Pour la Cellule WALL-ON-LINE
Mme Béatrice VAN BASTELAER, Chef de Projet Adjoint
M Vincent MYNSBERGHE, Expert en Informatique

Date : 23/12/2003
Version : V1

SOMMAIRE

1 – PREAMBULE

Ce document a pour but :

- d'examiner les différentes actions possibles et nécessaires pour pouvoir mettre en place l'organisation d'un système d'archivage des sites web publics de la Région Wallonne ;
- de dresser une cartographie non exhaustive des solutions ayant été mise en œuvre dans différents pays, pour en tirer les enseignements nécessaires qui permettront de faire le choix d'une politique d'archivage adaptée aux besoins de la Région Wallonne et des citoyens ;
- d'avoir une meilleure lisibilité en terme d'intégration technologique, de coût et de délais de mise en œuvre.

1.1 - Identification de la phase

Le présent document répond aux exigences de la phase suivante :

- Phase 1 – Analyse des besoins et des pratiques existantes

1.2 - Définitions

Le panorama des solutions d'archivage de sites web qui est présenté ici, a pour objectif de montrer à travers cinq projets différents, la multiplicité des approches possibles, à savoir :

➔ les approches intégrale, exhaustive, sélective, par échantillonnage, qu'elles soient automatisées, semi-automatisées ou manuelles.

- **L'approche intégrale**

L'approche intégrale consiste à collecter l'intégralité du Web mondial grâce à des robots sans aucune notion de sélection, de valeur patrimoniale ou de dépôt légal. Dans le cadre de cette étude, cette approche ne sera pas abordée en détail car, bien qu'entrant dans le cadre de l'archivage du Web, elle ne peut en aucune manière servir de piste de réflexion pour le projet de la cellule WallOnLine.

En revanche il est important de noter que le projet le plus représentatif de cet archivage « sauvage » est sans aucun doute **www.archive.org** et sa **WaybackMachine**.

- **L'approche exhaustive**

Cette approche souvent automatisée, permet de recueillir un corpus important à moindre frais. Elle consiste donc à collecter l'ensemble des informations d'un corpus sans critères de sélection prédéfinis.

Elle ne permet qu'une sélection sommaire des informations ou documents, elle ne propose pas ou peu d'indexation, et bloque souvent sur des questions techniques.

La qualité même de l'archive et du mode d'archivage est mise en question.

- **L'approche sélective**

Cette approche souvent manuelle offre :
- une archive de grande qualité,
- une sélection et une indexation manuelle des contenus, des autorisations…

En revanche les faiblesses de cette approche se situent au niveau :
- du coût en terme de ressources humaines et en terme de temps,
- de la lenteur relative liée au circuit de validation sur ce qui doit être ou non archivé.

- **L'échantillonnage**

Partant du principe qu'il n'est pas possible de tout conserver (tous les sites, toutes le mises à jour, toutes les versions), l'échantillonnage permet de faire une sélection représentative en terme :
- de nombre de sites,
- de nombre de fichiers,
- de fréquence de collecte.

2 – PANORAMA DES SOLUTIONS

Le tableau ci-dessous présente les différentes solutions qui seront détaillées dans le présent document. Tour à tour seront présentées les différentes solutions qui ont été mises en œuvre dans les pays suivants :

- En Suède, à la Bibliothèque Royale
- Au Québec, à la Bibliothèque Nationale
- En Australie, à la Bibliothèque Nationale d'Australie
- En France, à la Bibliothèque Nationale de France et à l'Institut National de l'Audiovisuel

	Manuelle	Semi-automatisée	Automatisée
Exhaustive	Impossible	Pas de projet en cours	**KulturarW3 - Bibliothèque Royale de Suède (cf. § 2.1)**
Sélective	**Bibliothèque Nationale du Québec (cf. § 2.2)**	**Pandora - Bibliothèque Nationale d'Australie (cf.§ 2.3)**	Impossible
Echantillonnage	**premier-ministre.gouv.fr (cf. § 2.4)**	**Bibliothèque Nationale de France et Institut National d'Audiovisuel (cf. § 2.5)**	sans intérêt par rapport à une approche exhaustive automatisée

Tableau[1] synoptique des projets d'archivage des sites web ayant été menés dans d'autres pays que la Belgique

- Chacune des approches adoptées présente un certain nombre d'avantages et d'inconvénients, que nous détaillerons dans les paragraphes suivants.

[1] Tableau non exhaustif

2.1 – Approche exhaustive automatisée - projet KulturarW3, Suède

Le projet KulturarW3[2] est sans doute le projet exhaustif automatisé le plus abouti. En effet, dans le panorama mondial des projets d'archivage de données en ligne, le projet suédois est apparemment le seul qui ait choisi une approche entièrement automatisée.

Initié en septembre 1996 par la Bibliothèque Royale, le projet KulturarW3 a bénéficié d'une aide financière initiale de 3 millions de couronnes suédoises (~366 000 EUR) pour établir la première étude de définition concernant les différentes méthodes de collecte, de préservation et de mise àdisposition des documents en ligne suédois.

Les Suédois ont effectué 10 collectes du web depuis janvier 1997 dont trois sont maintenant partiellement accessibles en ligne. Le robot utilisé est un robot d'indexation modifié en robot d'archivage.

Les domaines explorés sont *.se*, *.com*, *.net*, *.org*, et *.nu*. La taille du web suédois représentait, en 2000, 5 millions de pages réparties sur 31 000 sites (25000 en *.se* et 6000 sur les autres domaines). Avec les images, les sons etc., la taille de la base s'élève à environ 9,7 millions de fichiers soit 200 Giga bites. Les ingénieurs du KulturarW3 précisent à ce sujet que les problèmes ne viennent pas de la taille de l'archive mais du nombre important de fichiers hétérogènes et de liens àgérer.

Voici quelques données statistiques, intéressantes par leur rareté et leur précision. Elles permettent de dresser une typologie très précise du web suédois.

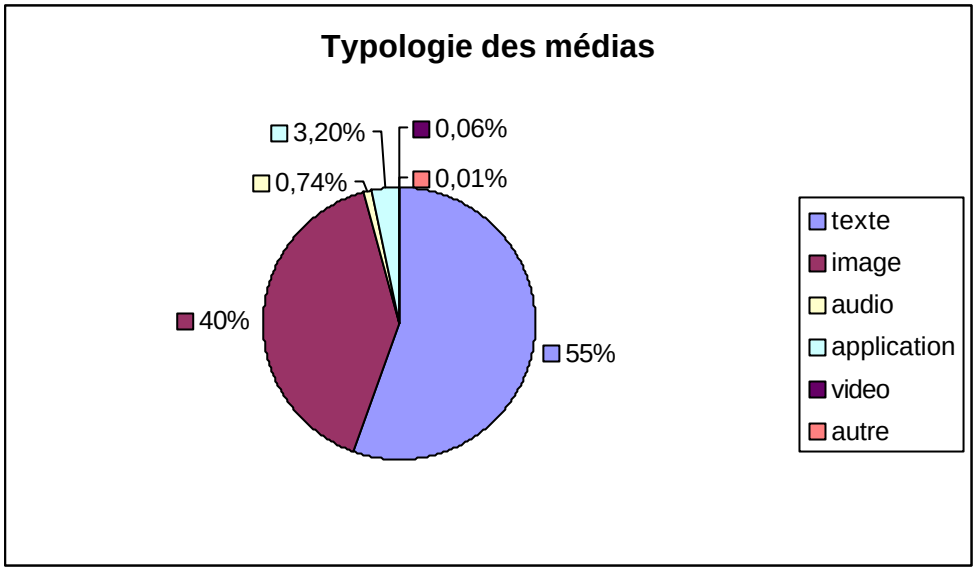
Données générales :

Strates	Début	Fin	Nbre de fichiers (en million)	Poids (Gb)
1	24/03/97	26/08/98	5 771	143
2	01/09/98	01/12/98	9 383	198
3	02/04/98	15/06/98	11 600	180
4	25/08/98	31/10/98	11 246	179
5	22/12/98	25/03/99	14 576	235
6	09/07/99	13/12/99	12 536	361
7	23/03/00	23/11/00	30 764	1 221
8	02/05/01	18/10/01	29 978	1 334
9	28/11/01	22/01/02	12 874	646
Total :			138 728	4 4497

Après avoir mis 17 mois pour collecter 143 Giga bits, l'équipe suédoise a réussi àne plus mettre que 2 mois pour 180 Gb.
Vers la fin de l'année 1999, les chercheurs suédois avaient tellement affiné leur outil de collecte qu'ils pensaient être en mesure d'archiver le web national en seulement trois semaines. Mais cette période a vu l'expansion colossale d'Internet dans le monde en général et en suède en particulier, le nombre de fichiers en ligne augmentant de manière exponentielle.

Bien que leur robot soit de plus en plus performant, le temps de collecte d'une strate s'est remis àaugmenter. Cela étant, la collecte de 1 334 Gb en 5 mois est un véritable exploit.

[2] http://www.kb.se/kw3/


Typologie des médias
3,20%
0,06%
0,74%
0,01%
40%
55%
texte
image
audio
application
video
autre

2.2 – Approche sélective manuelle - projet de la Bibliothèque Nationale du Québec

Certainement l'une des institutions ayant le plus communiqué sur son projet, la Bibliothèque Nationale du Québec (BNQ) a ouvert la voie aux institutions dans le domaine de l'archivage des publications électroniques émanant du web.

Son approche extrêmement sélective a conduit la BNQ à prévoir une préservation à long terme non pas pour les sites web eux-mêmes mais pour les documents textuels qu'ils pourraient contenir, considérés comme importants et n'existant pas sur d'autres supports.

Le projet se découpe en trois phases :

- Phase I : février - septembre 2001

  Dépôt d'environ 1 000 titres signalés par une vingtaine de ministères et organismes gouvernementaux dans la Banque des publications gouvernementales accessibles par Internet. Cette Banque est diffusée par le MRCI (Ministère des Relations avec les Citoyens et de l'Immigration) sur le portail du gouvernement du Québec[3].

- Phase II : mars - décembre 2002

  Dépôt rétrospectif de l'ensemble des publications assujetties au dépôt légal diffusées sur les sites Web de quatre ministères invités (Culture et Communication, Education, Finances et Ressources Naturelles). On estime le corpus à3 000 titres.

- Phase III : janvier 2003

  Dépôt de l'ensemble des titres diffusés par les ministères et organismes gouvernementaux (250 ministères et organismes, près de 50 000 titres).

La politique générale est donc de préserver en priorité les documents des sites gouvernementaux, essentiellement au format *Adobe Acrobat (.pdf)* ou traitement de texte.

➔ Le choix de solutions propriétaires non normalisées est étrange venant de la part de documentalistes. En effet, l'utilisation de format de fichiers non normalisés (.pdf, .doc, .xls etc..) par rapport à des fichiers tels que les .rtf ou .htm entraîne une **dépendance à un éditeur**.

Si Microsoft ou Adobe venaient à changer de format, ou à péricliter, c'est toute une politique d'archivage qui serait remise en question. Il est évident que le fait que ces formats soient des standards de fait entraîne d'une part une économie importante en les conservant au format original et d'autre part garantie leur intégrité.

Lors d'un exposé concernant l'avancement des travaux de la BNQ *au 28ème Congrès de l'ASTED*[4] en novembre 2001, Danièle Léger, coordonnatrice à la section de dépôt légal, proposait une analyse des approches des autres pays.

En comparant les approches exhaustives automatisées avec l'approche sélective manuelle, elle écrit ceci :

« Chacune des approches a ses mérites…et ses faiblesses. L'approche intégrale présente l'avantage de l'exhaustivité, de l'inclusion immédiate sans trop s'embarrasser des problèmes

[3] http://www.gouv.qc.ca

[4] Association pour l'avancement des Sciences et techniques de la Documentation (http://www.asted.org)

de découpage d'une ressource particulière, de son extraction par rapport à son site d'origine. »

2.3 – Approche sélective semi-automatisée – projet PANDORA, Australie

En juin 1996, la Bibliothèque Nationale d'Australie a mis en place le projet **PANDORA**[5] (Preserving and Accessing Networked Documentary Ressources of Australia).

De juin 1996 à fin 1997, ses chercheurs ont développé une politique et des procédures de sélection, de capture et d'archivage pour l'accès à long terme des publications électroniques australiennes.

Dès la fin 1997, le concept est déposé et une première archive d'environ 229 documents est créée. A l'heure actuelle, PANDORA a conçu grâce au **SCOAP** (Selection Committee on Online Australian Publications), un processus de travail ainsi qu'une liste des problèmes rencontrés et des spécifications techniques.

Le processus d'archivage se déroule de la façon suivante :

a. Evaluation de la publication de manière à déterminer sa structure, ses particularités, etc.,

b. Obtention de la permission de l'éditeur d'archiver sa publication,

c. Catalogage de la publication sur la base de données de la Bibliothèque Nationale Australienne de manière à s'assurer que cette publication est accessible, notamment par la création d'un hyperlien vers elle,

d. Envoi d'une requête pour archiver la publication dans « l'*Archive Management System ».* Cette action lance le robot *Harvest* (logiciel de collecte de fichiers en ligne) pour qu'il récupère les fichiers sur le web. Parfois c'est *WebZip* (autre logiciel de sauvegarde de site qui a la particularité de compresser les archives qu'il produit) qui est utilisé.

e. Comparaison entre l'archive récupérée du web et la source en ligne de manière à vérifier que toutes les pages et tous les liens ont été sauvegardés correctement,

f. Renvoi d'un rapport de vérification signifiant si la copie est conforme ou s'il faut corriger des erreurs,

g. Construction d'une page d'entrée pour la publication archivée et l'allocation d'une PURL[6] (Persistent Uniform Ressource Locator) qui est plus stable qu'une URL classique car elle permet de suivre les changements d'adresses,

h. Vérification périodique de la collecte et comparaison entre les pages en ligne et l'archive, surtout pour les publications périodiques.

Tel que décrit, le travail de PANDORA n'est que partiellement automatisé. Les réalisateurs de ce projet tentent ainsi de conserver un équilibre entre rapidité de collecte, pertinence et qualité de l'information archivée.

[5] http://pandora.nla.gov.au

[6] Chaque adresse PURL est associée à une - et seulement une - adresse URL, et chaque adresse PURL est unique. Les adresses PURL ne peuvent être permanentes que si quelqu'un maintient la relation entre l'adresse PURL et son adresse URL correspondante.

Leurs critères de sélection concernant la qualité des publications sont d'ailleurs assez sévères. Leur argument est le suivant : contrairement à l'imprimerie, les éditeurs en ligne ne font pas de sélection, elle doit donc être faite au niveau de la collecte.

Il est intéressant de souligner les principales limites du projet PANDORA :

- Difficultés d'obtenir la permission de l'éditeur d'archiver ses publications (pas de réponses reçues, incompréhension, refus etc.),
- Problèmes liés à la collecte de certains fichiers tels les applets Java, les bases de données ou les sites dynamiques, qui ne peuvent être collectés par *Harvest*[7],
- Problèmes liés au catalogage et à l'indexation car les publications électroniques sont plus difficiles à décrire que les publications « classiques ».

[7] Harvesting : technique automatisée de collecte d'information reposant sur l'utilisation d'aspirateurs de site

2.4 – Approche par échantillonnage manuel – site du Premier Ministre, France

Conformément à la demande des Archives Nationales de France, le site du Premier Ministre français, comme tous les autres sites gouvernementaux, en *.gouv,* est livré régulièrement sur support physique (en l'occurrence le disque compact) a ces dernières.

Cependant, conscient de l'aspect nécessaire mais insuffisant de cette démarche, le responsable du site, Monsieur Benoît Thieulin, a entrepris d'archiver les différentes versions de ce site et de les proposer au public.

➔ ***Données existantes :***

La prévoyance de ses prédécesseurs qui n'avaient pas détruit les versions antérieures du site a permis d'accéder aux données sauvegardées.

En revanche, la multiplicité des techniques employées sur ces sites a rendu la tâche plus ardue. Le souci de Benoît Thieulin était d'une part de conserver le patrimoine que représentent ces sites mais aussi de le restituer au public.

➔ ***Fréquence d'archivage :***

Le site n'est pas archivé de manière régulière. La notion de version intéresse Benoît Thieulin, c'est-à-dire qu'il s'efforce de conserver une image exacte de la version d'un site avant évolution vers une autre formule ou un autre gouvernement. Il fonctionne avec la méthode de l'instantané (Snapshot).

➔ ***Uniformisation des technologies :***

Les premiers sites du gouvernement sont des sites statiques, n'utilisant que des technologies normées comme le HTML et les formats d'images Gif ou Jpeg.

Puis, pour des questions pratiques, les sites sont devenus dynamiques : pour créer les pages, il est nécessaire de faire appel à différents éléments dans une ou plusieurs bases de données.

Dans le cas du site du Premier Ministre, la technologie propriétaire Cold Fusion avait été choisie. A ce sujet, Benoît Thieulin regrette d'ailleurs que les logiciels libres et ouverts n'aient pas été plus populaires à cette époque car, dit-il, s'il devait refaire un choix, ce serait vers ces solutions qu'il s'orienterait.

Pour l'accès aux archives, il fallait uniformiser les technologies. Le site a été standardisé. En effet, il a été demandé à un prestataire de convertir les sites dynamiques en sites statiques renforçant encore l'image de photographie instantanée. Conscient de l'aspect perfectible de ce choix, Benoît Thieulin a préféré, à l'époque, et en l'absence de recommandation sur le sujet, aller vers les technologies standardisées et, par conséquent pérennes.
A l'adresse http://www.archives.premier-ministre.gouv.fr, on accède à un menu de navigation permettant de voir les quatre versions du site de 1996 à 2002 :


Archives
Premier
ministre
archives.premier-ministre.gouv.fr
La base de données des sites archivés des précédents Gouvernements
Un accès libre aux archives des Premiers ministres depuis 1996
Le site www.archives.premier-ministre.gouv.fr constitue ainsi une véritable base de données de l'activité gouvernementale depuis la création du site du Premier ministre. Il offre un accès libre :
au site du gouvernement d'Alain Juppé (1996-1997)
à la version 1 du site du gouvernement de Lionel Jospin (1997-1998)
à la version 2 du site du gouvernement de Lionel Jospin (1998-2000)
à la version 3 du site du gouvernement de Lionel Jospin (2000-2002)
ARCHIVE
Préparé depuis plusieurs mois, le changement de Gouvernement a conduit la rédaction du site du Premier ministre à une remise à jour complète de www.premier-ministre.gouv.fr, portail du Gouvernement français. Ainsi, la base de données du site a été totalement refondue et ce sont près de 1500 pages qui ont été modifiées entre la démission de Lionel Jospin et la nomination de Jean-Pierre Raffarin, le 7 mai dernier.

2.5 – Approche par échantillonnage semi-automatisé – INA et BNF, France

La **Bibliothèque Nationale de France** (BNF) et l'**Institut National de l'Audiovisuel** (INA) sont en charge du dépôt légal des publications électroniques en ligne.

Chacune de ces deux institution devant se répartir la tâche en fonction du type de contenu du site collecté et dans l'optique d'assurer la continuité de ses propres collections. Ainsi l'Ina sera en charge de la conservation des sites web axés sur la vidéo ou la radio. La BNF prendra les autres en charge.

- L'approche choisie pour mener à bien ce projet est un échantillonnage semi-automatisé. Partant du principe qu'il ne serait pas possible de tout conserver, les deux institutions ont choisie d'archiver en priorité les sites àforte notoriété.

Ce critère de notoriété étant déterminé par le nombre de visiteurs, le nombre de liens qui pointe vers le site et la fréquence des mises àjour. Les sites moins "populaires" seront peut-être collectés ultérieurement et àune fréquence moindre.

A chaque fois que cela sera possible, les sites seront collectés par des robots mas dans tous les cas où les difficultés techniques ne permettent pas cette collecte automatisée, une démarche de dépôt volontaire sera demandée aux propriétaire de site.

2.6 – Autre démarche intéressante - ministère de l'Emploi et de la solidarité, France

Une opération pionnière a été menée depuis 2001 par le ministère de l'Emploi et de la solidarité :

Extrait du compte rendu d'expérience :

> ***Expérience d'archivage des sites internet du ministère de l'Emploi et de la solidarité***
> *Un groupe de travail, créé à l'initiative du centre de ressources du ministère puis piloté par la mission des Archives Nationales, s'est mis en place à l'été 2001. Ce groupe est composé pour le ministère, d'informaticiens, de documentalistes, de webmestres et d'archivistes et en ce qui concerne les partenaires extérieurs, d'un représentant de la BNF, d'un représentant de la Direction des Archives de France (chargée de mission auprès de la directrice) et des Archives Nationales (CAC, programme Constance). Les archivistes du ministère ont su profiter de l'occasion. En effet, moyennant quelques corrections, le système de gestion développé en interne par le service informatique correspond aux besoins formulés par la mission des Archives Nationales et le CAC pour la collecte. De façon schématique, les opérations doivent se dérouler de la façon suivante : au préalable, l'équipe a rédigé un tableau de gestion hiérarchisé des types de documents. Par un système de filtre, dès qu'un webmestre met un document en ligne, les documents à destination des Archives Nationales sont transférés sur un serveur consacré à l'archivage (en application du tableau de gestion). L'archiviste peut valider la sélection opérée automatiquement par l'ordinateur et choisir de conserver ou d'éliminer les documents stockés sur le serveur "archives". Les méta données peuvent lui être fournies par la base de données que le webmestre doit obligatoirement renseigner pour pouvoir poster son document (la liste des méta données retenues est celle proposée par l'ATICA[8]). Cette opération d'archivage est actuellement en phase-test. La DGAFP-DIRE[9] des Services du Premier Ministre s'appuie sur l'expérience en cours au Ministère de l'Emploi avec des spécificités locales. Le travail coopératif est en particulier un outil beaucoup plus utilisé par cette direction. La réflexion en cours porte sur : l'élaboration d'une étude fonctionnelle de l'archivage, "l'exportation" éventuelle de la base de données développée en interne par le Ministère de l'Emploi vers d'autres ministères par convention, l'étude des coûts (base de données, personnel, matériel). En conclusion, plusieurs défis restent à relever : l'archivage des pages dynamiques, la conception et l'élaboration d'instruments de recherche, les techniques de conservation.*

[8] Agence pour le développement de l'administration électronique http://www.atica.pm.gouv.fr

[9] Direction Générale de l'Administration et de la Fonction Publique – Délégation Interministérielle à la Réforme de l'Etat.

3 – ACTIONS A MENER

D'un point de vue méthodologique ce document a pour objectif :

- de stimuler et de favoriser la prise de décision des différents donneurs d'ordre intervenant sur ce projet,
- de lancer un **cycle de validation** nécessaire sur les grandes orientations qui se profilent quant à l'archivage des sites web publics (cf. document SC0335_07a : Note d'organisation),
- de mettre en perspective les résultats issus de la typologie des sites.

Le groupe SERDA se tient à votre disposition pour vous accompagner dans cette phase, ainsi que pour éclairer les points nécessitant d'être approfondis.