Date : 07/07/2008

RELU PAR LE CFI

Maintenir l'accès aux sites web gouvernementaux : développement et pratiques du Cybercimetière

Starr Hoffman
Bibliothèque de l'Université du Nord Texas
Denton, Texas, Etats-Unis

Traduit en français par
Claudine Even,
Bibliothèque de Sciences Po, Paris (France)
Juillet 2008

Meeting: **130. Information gouvernementale et publications officielles**

nment Information and Official Publications

Interprétation simultanée : Allemand, Anglais, Arabe, Chinois, Espagnol, Français et Russe

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

RESUME :

A la fin des années 1990, l'information gouvernementale en ligne apparaissait et disparaissait rapidement. En 1999, les bibliothèques de l'Université du Nord Texas (UNT) ont conclu un partenariat avec l'US Government Printing Office (GPO) pour étudier la question de l'archivage électronique des sites web gouvernementaux.

Ces archives - connues sous le nom de Cybercimetière - offrent un accès public permanent aux sites des agences et commissions gouvernementales américaines disparues. Ce partenariat entre l'UNT et le GPO s'est étendu et inclut la National Archives and Records Administration (NARA). Cette contribution porte sur le développement du Cybercimetière, les procédures de sélection, de capture et d'édition du contenu dans les archives.

TEXTE INTEGRAL

Introduction

A la fin des années 1990, l'information gouvernementale sur Internet apparaissait et disparaissait rapidement. Cette information était souvent publiée uniquement en ligne, sans version imprimée. Aussi si elle était mise à jour ou supprimée, la version originale était perdue. En 1999, les bibliothèques de l'Université du Nord Texas (UNT) ont conclu un partenariat avec l'US Government Printing Office (GPO) pour étudier la question de l'archivage électronique des sites gouvernementaux.

Ces archives - connues sous le nom de Cybercimetière - fournissent un accès public permanent aux sites web et aux publications des agences et commissions gouvernementales disparues. Les archives web sont hébergées sur des serveurs dans la bibliothèque de l'UNT et sont accessibles à l'adresse : http://govinfo.library.unt.edu. Elles contiennent des informations sur des thèmes variés et des sources importantes comme le site web de la National Commission on Terrorist Attacks upon the United States, aussi appelée Commission du 11 septembre. Comptant un seul site en 1999, le Cybercimetière en contient 45 en 2008 et 9 autres sont en cours d'archivage. Chaque mois il reçoit 230 000 visites et 1 million de pages sont consultées.

Les usagers peuvent accéder au Cybercimetière de partout gratuitement. Cela répond à la mission du Federal Depository Library Program (FDLP) qui est de fournir un accès public aux bibliothèques du FDLP, et permet ainsi d'accéder à des informations de grande valeur. 79% des consultations proviennent des Etats-Unis, mais une minorité significative sont faites depuis d'autres régions du monde : Europe (5,67 %), Asie (3,77 %). Les groupes d'usagers les plus importants, hors des Etats-Unis, sont en Chine (1,6 %), au Canada (1,5 %) et au Royaume-Uni (1,2 %).

Il est important que le contenu de ces sites reste accessible en raison de sa grande valeur. Les sites du Cybercimetière portent sur un grand nombre de sujets qui montrent l'étendue de l'information gouvernementale. Les archives contiennent des statistiques, des rapports de commissions, des recommandations, des photographies, des vidéos, des transcriptions d'auditions et d'autres documents. Cette information est intéressante pour les employés gouvernementaux, les Américains, mais aussi la communauté internationale. De plus, il est important de conserver des documents pour l'avenir, en tant que trace de l'histoire du pays.

Le partenariat

En 1995, l'US Government Printing Office a publié dans un rapport son plan stratégique qui mettait l'accent sur la nécessité de conserver les publications électroniques de l'organisme. Le GPO et les bibliothèques dépositaires s'inquiétaient du fait que les institutions retiraient les informations de leurs sites sans en assurer la conservation. Dans son plan stratégique de 1996 « Study to identify Measures necessary for a Successful Transition to a more Electronic Federal Depository Library Program », le GPO faisait un appel à partenariat avec les bibliothèques dépositaires pour repérer et archiver les matériaux (1).

Cathy Hartman, une bibliothécaire spécialisée dans les publications officielles à l'UNT a engagé des discussions avec le GPO, elles ont abouti à un partenariat. La première collection sélectionnée pour l'archivage a été le site web de l'Advisory Commission on Intergovernmental Relations (ACIR), une agence gouvernementale disparue. Comme l'agence n'existait plus, aucun contenu ne serait ajouté et le site devait être désactivé. Le choix s'est porté sur ce site en raison, d'une part, de son statut « à risque » et, d'autre part, du fait de son contenu statique, le coût de conservation et de stockage restant donc faible. Le Memorandum of Understanding (MOU) signé par le GPO et l'UNT précise que l'objectif premier du projet est d'offrir un accès gratuit, sans restriction et permanent aux sites web de l'organisme. Beaucoup de spécifications du MOU répondent aux directives traditionnelles du Federal Depository Library Program (FDLP) qui offre un accès gratuit et permanent aux ressources imprimées.

Cet accord a été étendu en 1999 à d'autres sites, tous provenant d'organisations gouvernementales ayant cessé leurs activités. Cette révision a été réalisée car il devenait urgent d'assurer la conservation de ces documents. Ces informations sont souvent produites uniquement sous forme électronique, aussi quand l'organisation qui les publie disparaît et que le site est désactivé, il n'y a plus aucun accès public à ces sources. Quand les archives se sont développées, elles ont été connues sous le nom de Cybercimetière, ce qui indique bien leur rôle d'archivage électronique des sites des organismes disparus.

Ce partenariat entre le GPO et l'UNT a été révisé à nouveau en 2006 pour inclure la National Archives and Records Administration (NARA). L'UNT est maintenant affiliée aux Archives nationales et constitue l'une des trois institutions du secteur de l'éducation ayant cette distinction. Quand les sites sont archivés, la NARA y donne accès comme partie des archives nationales.

Les sites archivés

Comme dit précédemment, l'archivage concerne les sites de toutes les organisations gouvernementales disparues. Une fois qu'on a repéré une agence ou une commission qui va prochainement cesser ses activités, nous étudions attentivement son site. Nous faisons une seule capture et archivons la dernière version la plus complète. La capture est effectuée quand

l'organisme a été supprimé, quand la date d'expiration est dépassée ou après la publication du rapport final. La totalité du site est archivée, ce qui inclut tous les fichiers.

Les procédures d'archivage

La procédure d'archivage dans le Cybercimetière a évolué au cours de la dernière décennie. L'UNT vérifie tout d'abord quelles sont les agences et commissions gouvernementales qui risquent de disparaître. Cela constitue une partie de mon travail en tant que bibliothécaire en charge des collections numériques, au sein du département des publications gouvernementales. D'autres bibliothécaires du réseau FDLP m'envoient aussi des références de sites. Comme le Cybercimetière est maintenant bien connu, il arrive que des organisations gouvernementales contactent directement le GPO ou l'UNT pour s'enquérir de la possibilité d'archiver leurs sites.

Ensuite le site est évalué. Il doit répondre aux critères suivants :

1) Ce doit être le site officiel d'une organisation gouvernementale
2) L'agence ou la commission doit être sur le point de fermer, doit avoir publié son dernier rapport ou encore signaler qu'elle risque de disparaître.

Si l'institution a demandé elle-même l'archivage de son site, nous envoyons le questionnaire suivant à son représentant ou à l'administrateur du site :

- Quel système d'exploitation a été utilisé pour héberger le site ?
- Quel est le logiciel utilisé pour le serveur web ?
- Est-ce que des serveurs annexes (SSI) ont été utilisés ?
- Etait-ce un site statique en html ou était-ce un site dynamique ?
    - Dans ce cas, quels langages de programmation ont été utilisés (php, perl, python) ?
- Est-ce qu'une base de données a été mise en place pour ce site ?
    - Si oui quelle base de données
    - Et quels étaient les moyens de connexion à la base de données
- Est-ce que du streaming était utilisé sur le site ?
- Est-ce qu'il y a des fichiers de formats propriétaires sur le site ?
- Y a-t-il autre chose que vous souhaiteriez ajouter ?

Ces questions permettent de savoir si la capture par moissonnage du site peut être faite facilement et en totalité. Nous avons jusqu'ici utiliser HTTrack, un logiciel libre de moissonnage du web (http://www.httrack.com/page/1/en/index.html). Ce logiciel permet de télécharger la totalité d'un site, y compris les fichiers html, les images et tout autre type de fichiers.

Nous changeons actuellement de logiciel de moissonnage et passons de HTTrack à Heritrix (http://crawler.archive.org). Heritrix est un robot de recherche libre développé pour l'Archive Internet. Heritrix archive les sites web en fichiers ARC qui stockent des ressources multiples dans un fichier unique dont la taille va de 100 à 600 MO. Quand nous aurons développé une interface pour ces fichiers nous pourrons présenter et signaler le statut de l'archive sans altérer le code original. Cela constituera une amélioration par rapport à la méthode précédente qui demandait des changements mineurs pour désactiver les pages de contacts et pour indiquer le statut de l'archive.

Dans certains cas, c'est l'institution qui nous a fourni le contenu du site. Pour cela les données doivent être copiées sur un support amovible pour être envoyées à l'UNT. C'est ce qu'ont fait, par exemple, l'Office of Technology Assessment, la Commisison du 11 septembre et d'autres. A l'avenir, cette méthode sera utilisée uniquement quand le site contiendra du contenu impossible à moissonner. Le moissonnage offre de meilleures garanties en ce qui concerne la conformité au site original lors de sa clôture.

Pour le moment la taille du site à archiver ne pose pas de problème. Nous avons actuellement 13 GO disponibles sur le serveur pour de nouveaux contenus. Quand nous aurons besoin de davantage, les services des Government Documents and Information Technology Services (ITS) analyseront les besoins et étudieront les différentes solutions. Nous ne pensons pas que l'expansion future pose problème, car les coûts de stockage sont relativement faibles.

Quand tout le site a été moissonné - ou reçu - nous vérifions s'il y a des erreurs et s'il est complet. C'est une procédure à la fois manuelle et automatique. D'une part, on navigue manuellement dans le site et on le compare à l'original et, d'autre part, on utilise le logiciel de vérification des liens URL Xenu Link Checker pour repérer les liens cassés, tant dans les archives que dans le site d'origine (http://home.snafu.de/tilman/xenulink.html). Nous comparons ensuite les résultats des deux rapports afin de repérer les fichiers qui auraient été oubliés et qu'il faut capturer avant que l'archive soit mise en ligne. Dans la mesure du possible, on reprend les fichiers manquants sur le site d'origine. Il existe d'autres solutions : récupérer le fichier dans les Archives d'Internet ou contacter l'agence ou la commission par l'intermédiaire du GPO.

Le Memorandum of Understanding (MOU) signé entre le GPO et l'UNT autorise une légère altération du code du site pour indiquer qu'il n'est plus actif. Pour ce faire, la méthode standard utilisée a été de marquer chaque page du site. Afin de modifier au minimum le code, l'accord permet d'ajouter le mot « Archive » en Times New Roman 8 en haut de chaque page. Pour respecter les objectifs du Cybercimetière tant en matière d'accessibilité que d'autorité, notre nouvelle procédure de moissonnage avec Heritrix, nous permettra d'indiquer la nature

du site sans modifier le codage. Et aussi d'économiser le long travail de modification de toutes les pages de chacun des sites archivés.

Nous avions l'habitude de désactiver les liens des contacts et les adresses électroniques, dans la mesure où ils ne sont plus actifs, ni utiles car l'organisme n'existe plus. Cependant quand Heritrix sera utilisé régulièrement pour le moissonnage aucune modification ne sera plus apportée au code. La page d'accueil du Cybercimetière indiquera que les sites sont des archives et que les contacts et les liens sont caducs.

Après, le site est chargé sur les serveurs de l'UNT et un lien ajouté sur les pages du Cybercimetière listant les sites classés par noms, par branche gouvernementale ou par date d'expiration. Si le GPO, l'agence ou la commission sont à l'origine de l'archivage, nous leur signalons à ce moment-là que l'archive du site est active.

Quand une organisation gouvernementale a elle-même demandé l'archivage dans le Cybercimetière, il arrive qu'elle souhaite conserver le nom de domaine pendant quelques temps après l'archivage du site. Dans ce cas, nous lui fournissons - ou au GPO quand il a servi d'intermédiaire - l'adresse IP du site dans les archives. C'est eux qui doivent renouveler l'abonnement au nom de domaine et assurer la redirection à partir de cette page vers la version archivée. Cela permet au public de trouver le site durant une période transitoire pendant laquelle l'ancienne adresse URL du site apparaît en tête des réponses dans les moteurs de recherche. Quand l'archive aura été active pendant quelques temps, c'est son adresse qui arrivera en tête des résultats de recherche, alors la redirection ne sera plus nécessaire. Jusqu'à présent, l'expérience nous a montré que les institutions conservaient leur abonnement au nom de domaine pendant un ou deux ans.

Quand le site archivé est opérationnel, nous l'ajoutons à la Bibliothèque numérique des bibliothèques de l'UNT (UNT's Libraries Digital Library System – DLS) http://digital.library.unt.edu. C'est le système en ligne qui regroupe toutes les collections numériques de l'UNT ainsi que les documents gouvernementaux produits sous forme numérique et les sites web capturés. Chaque site archivé dans le Cybercimetière est signalé dans la Bibliothèque numérique (DLS) afin de donner un accès supplémentaire au contenu. Les sites sont classés par sujets, en utilisant le Vocabulaire législatif d'indexation (Legislative Indexing Vocabulary) créé par le Congressionnal Research Service – CRS pour décrire la littérature législative et des politiques publiques. Ces indexations sont de bons moyens de découverte car elles permettent aux utilisateurs de l'UNT, qui sont familiarisés avec les collections numériques de l'UNT mais ignorent les noms des sites archivés, de trouver les références pertinentes dans le Cybercimetière. La Bibliothèque numérique permet aux usagers de rechercher grâce à une interface unique des sites, des images, des documents multimédias, à partir d'une variété de collections et de sujets. Enfin, les enregistrements du DLS nous permettent de partager les métadonnées relatives au Cybercimetière avec d'autres institutions via l'Initiative Archives ouvertes (Open Archive Initiative - OAI) et la recherche fédérée, accroissant ainsi l'accès du public et la connaissance du contenu.

Pour ajouter un enregistrement dans la Bibliothèque numérique, je commence par créer une « imagette » de la page d'accueil du site. Pour cela j'affiche la page et presse la

touche « Impr. Ecran » , puis j'ouvre un éditeur d'image, comme Adobe Photoshop et colle l'image dans un nouveau fichier. Je coupe l'image et la redimensionne, en général au format 118 x 90 pixels et 100 dpi. L' «imagette » finale pèse moins de 50 ko.

Ensuite je crée les métadonnées, elles reposent sur le schéma de métadonnées de l'UNT (http://www.library.unt.edu/digitalprojects/metadata). L'un des champs est la description du contenu, un paragraphe de trois à quatre phrases contenant le descriptif et les mots-clés, qui améliore la pertinence des résultats de recherche dans la Bibliothèque numérique (DLS). Cet enregistrement est directement lié au site archivé sur le serveur du Cybercimetière. Quand il est complet, je préviens l'Unité des projets numériques (Digital Projects Unit) que les métadonnées et l'icône sont prêtes à être chargées dans la Bibliothèque numérique (DLS).

Matériel, environnement et sauvegarde

Pour assurer un accès permanent à ces sites, les serveurs du Cybercimetière sont placés dans un lieu contrôlé, au sous-sol du bâtiment de la bibliothèque. La température de la salle des machines est maintenue aux environs de 3° C (38° Fahrenheit) et le taux d'hygrométrie à 50 %. Le Cybercimetière est hébergé sur un cluster de quatre serveurs redondants utilisant une baie SAN comme espace de stockage. Cybercimetière est hébergé dans un cluster à tolérance de panne composé de quatre nœuds utilisant un volume SAN pour le stockage. Il est actif sur un seul serveur à la fois, les trois autres servant de sauvegarde en cas de panne de matériel ou de logiciel ou lors de la maintenance. Le Cybercimetière comprend actuellement 22,2 GO de contenu et un volume serveur de 40 GO. L'allocation des espaces de stockage est discutée entre les départements des Government Documents and Information Technology Services. Pendant les week-ends des sauvegardes complètes sur bandes magnétiques sont réalisées, celles-ci sont conservées hors site par la société Iron Mountain (http://ironmountain.com).

Migration des données

Certains fichiers du Cybercimetière sont uniques et de formats propriétaires. A l'avenir, ils risquent de devenir difficiles à utiliser ou à lire avec les nouvelles versions des logiciels, et les logiciels peuvent devenir complètement obsolètes. La documentation sur les moyens d'accès à ces fichiers n'est pas toujours bonne. Comme l'objectif premier du Cybercimetière est de conserver l'accès et l'utilisation par le public, les fichiers seront transférés sur de nouveaux supports accessibles quand ils deviendront inutilisables.

Pour cela, nous allons faire l'inventaire des différents formats de fichiers actuellement présents sur le Cybercimetière. Nous identifierons ensuite ceux qui présentent des risques et élaborerons un plan de migration de ces formats de fichiers. Ce plan traitera de deux points importants : l'accès fonctionnel et le respect de l'aspect original du site. Sa mise en oeuvre nécessitera la participation du Superintendant of Documents, comme cela a été spécifié dans le Memorandum of Understanding (MOU) signé par le GPO et l'UNT. Les fichiers présentant

des risques seront migrés vers de nouveaux formats quand ce sera nécessaire, afin de maintenir l'accès au site archivé. Les fichiers originaux seront conservés sur le serveur du Cybercimetière pour assurer leur pérennité. Ceci est en conformité avec le MOU qui prévoit le maintien de l'accès au contenu. Tous les changements de ce genre seront signalés sur le site.

Conclusion

Le Cybercimetière a été créé pour assurer un accès public permanent à l'information gouvernementale qui disparaît d'Internet. Alors qu'au départ il contenait un seul site, il en compte aujourd'hui 45, maintenus grâce à un accord entre l'US Government and Printing Office, la National Archives and Records Administration et les bibliothèques de l'Université du Nord Texas. Comme la technologie évolue sans cesse, nous allons chercher de nouvelles méthodes pour conserver cette information et son accessibilité. En tant que membre du Federal Depository Library Program, l'UNT doit donner un accès permanent à cette information variée et de grande valeur.

(1) Hartman, C. N. (2000). Storage of electronic files of federal agencies that have ceased operation. A partnership for permanent access. Government Information Quarterly. 17, 299-307.