Date : 09/08/2008

Archivage du web pour tous dans les bibliothèques de l'Université d'Indiana : les sites gouvernementaux statistiques étrangers

Andrea Singer
Indiana University Bloomington Libraries
Bloomington, IN., USA

Traduit en français par:
Nadia Pazolis-Gabriel
(Alliance Française de Washington, États-Unis)

Meeting: **130. Government Information and Official Publications**
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

Résumé

Contexte : Dans son étude des diverses initiatives d'archivage du web pour Global Resources Network (http://www.crl.edu/grn/index.asp), James Simon, du Center for Research Libraries, dresse une liste de dix initiatives de capture et de conservation de sites web actuellement en cours, dont le service d'archivage sur inscription Archive-It. Le réseau des bibliothèques de l'Université d'Indiana à Bloomington (Indiana University Bloomington, IUB) figure parmi les institutions inscrites à Archive-It (http://www.archive-it.org) depuis 2006, afin de capturer, conserver, et permettre la recherche de sites web sans assistance technique ni infrastructure préalable.

Cette communication porte sur un projet d'archivage de sites gouvernementaux d'information, constituant l'un des trois fonds d'archives de l'Université d'Indiana. Deux de ces fonds répondent bien sûr à des besoins de conservation locale: les sites de l'Université d'Indiana, et une sélection de sites de l'État d'Indiana et de son administration locale. Le troisième fonds, qui nous intéresse aujourd'hui, est la collecte de sites d'agences statistiques nationales de pays hors de l'Union Européenne, de l'Australie, du Canada et des États-Unis.

Depuis plus d'un demi-siècle à l'Université d'Indiana, la sélection et l'acquisition de documents gouvernementaux provenant de pays autres que les États-Unis a permis de mettre en évidence des genres-clé pour les sciences sociales et les études historiques: annuaires statistiques, comptes-rendus législatifs, plans d'aménagement et autres. Comme cette documentation ne se trouve de plus en plus que sur le web, nous avons

choisi de consacrer une partie de notre budget à la conservation de sites web plutôt qu'à une documentation spécifique par l'intermédiaire d'Archive-It.

Coopération : Les sites archivés sont accessibles gratuitement et une recherche par mot-clé ou par URL est possible sur le site Archive-It. Chaque institution peut y ajouter sa propre interface de recherche, d'autres sites et une aide en ligne extérieure à l'interface d'Archive-It, ainsi que décider de la fréquence de capture, de la suppression ou l'ajout de nouveaux sites.

Comme souvent à l'Université d'Indiana, l'avancée de ce projet s'est faite grâce à la mise en place d'un comité ou groupe de travail. Dans ce groupe étaient présents du personnel des services de catalogage, des archives de l'université et du développement des collections. Les équipes techniques n'étaient pas impliquées. Un bibliothécaire spécialisé tient à jour des ressources web annexes et les «seeds» (amorces) pour chaque collection.

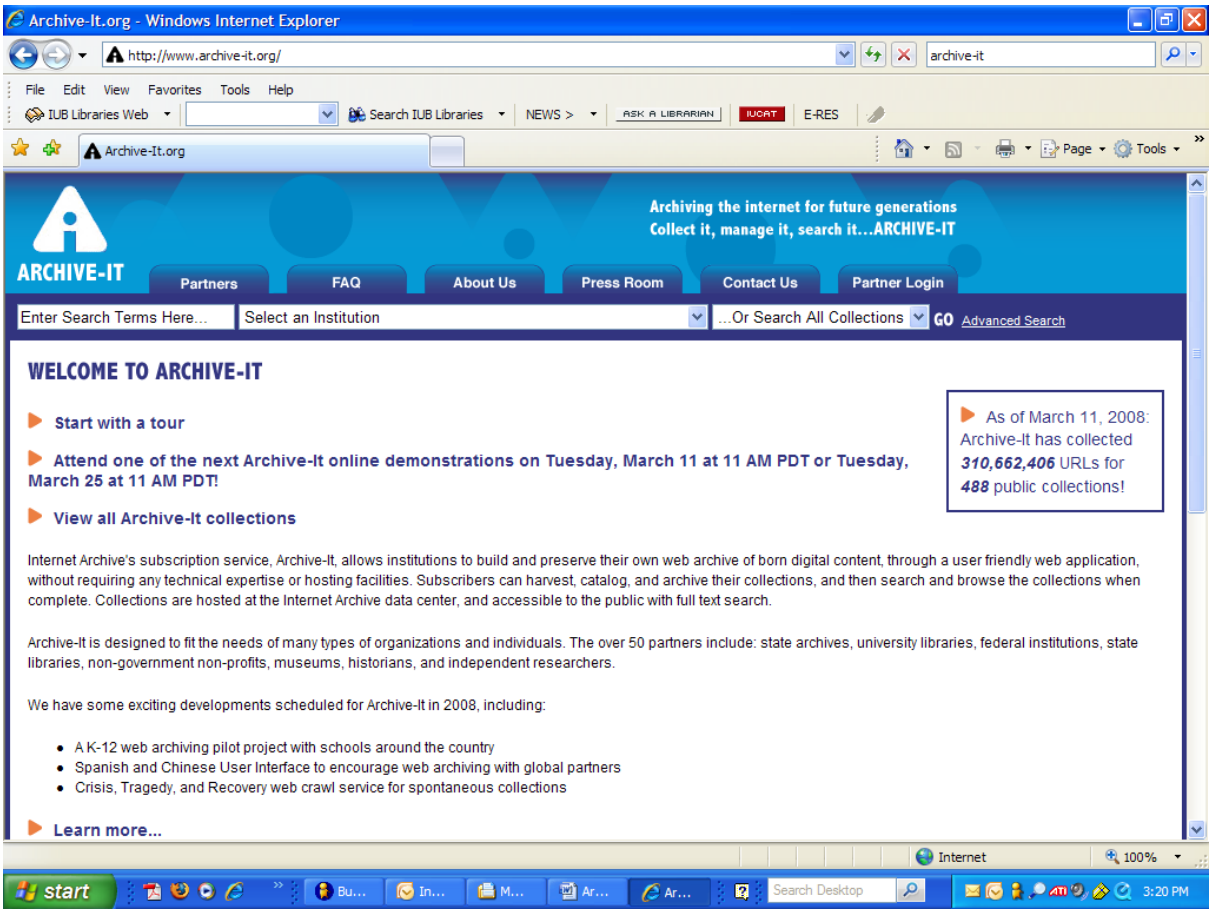
L'aide à la recherche prend également la forme de notices complètes décrivant nos collections dans le catalogue et sur OCLC Worldcat, des brochures imprimées, et des liens vers des sites tels que :
http://www.libraries.iub.edu/index.php?pageId=4302
http://www.archive-it.org/collections/317
http://www.libraries.iub.edu/index.php?pageId=4981

Dans un cadre plus large de coopération, la question est de s'assurer que nous n'effectuons pas un travail déjà fait ailleurs. Nous aborderons donc également les questions de sélection d'amorces qui ne sont pas archivées ailleurs, un registre des adresses URL ainsi que d'autres aspects de la coopération.

Biographie :

Andrea Singer est la bibliothécaire responsable de la documentation étrangère à l'Université d'Indiana à Bloomington, elle est également bibliographe pour le Département d'études indienne et tibétaine. Elle est conservatrice du fonds numérique de sites gouvernementaux étrangers de statistiques.

Merci de me donner l'occasion de partager avec vous cette expérience d'archivage de sites web. C'est en mars 2008 que j'ai commencé à rédiger cette communication, et j'ai vite constaté que certaines des sources consultées sur le web en préparant le résumé plus tôt dans l'année avaient disparues. C'est pourquoi j'ai décidé de ne pas inclure trop de détails sur les techniques de recherche, afin que pour ma présentation en août, nous soyons le plus synchronisé possible avec le web dans son état actuel. En mars 2008, on pouvait lire en en-tête de la page d'accueil de Archive-It : "Archiver internet pour les générations futures. Y collecter, le gérer, le rechercher... l'archiver grâce à Archive-It". Il offrait une recherche par mot-clé pour une institution précise ou pour l'ensemble des collections.



Le contexte :

Le site de Internet Archive propose un système d'archivage de sites web disponibles sur inscription, Archive-It, par l'intermédiaire d'adresses URL qui servent de clés, ou, dans la terminologie de Archive-It, d'amorces. Les partenaires, participants et abonnés sélectionnent des sites web regroupés en collection et gèrent la fréquence de moissonnage des sites concernés. Ils peuvent aussi limiter la recherche d'une certaine adresse de plusieurs manières[1]. Les documents sont collectés, hébergés et interrogeables dans Archive-It sans qu'aucune expertise technique ne soit exigée des partenaires qui peuvent ainsi se concentrer sur le contenu.

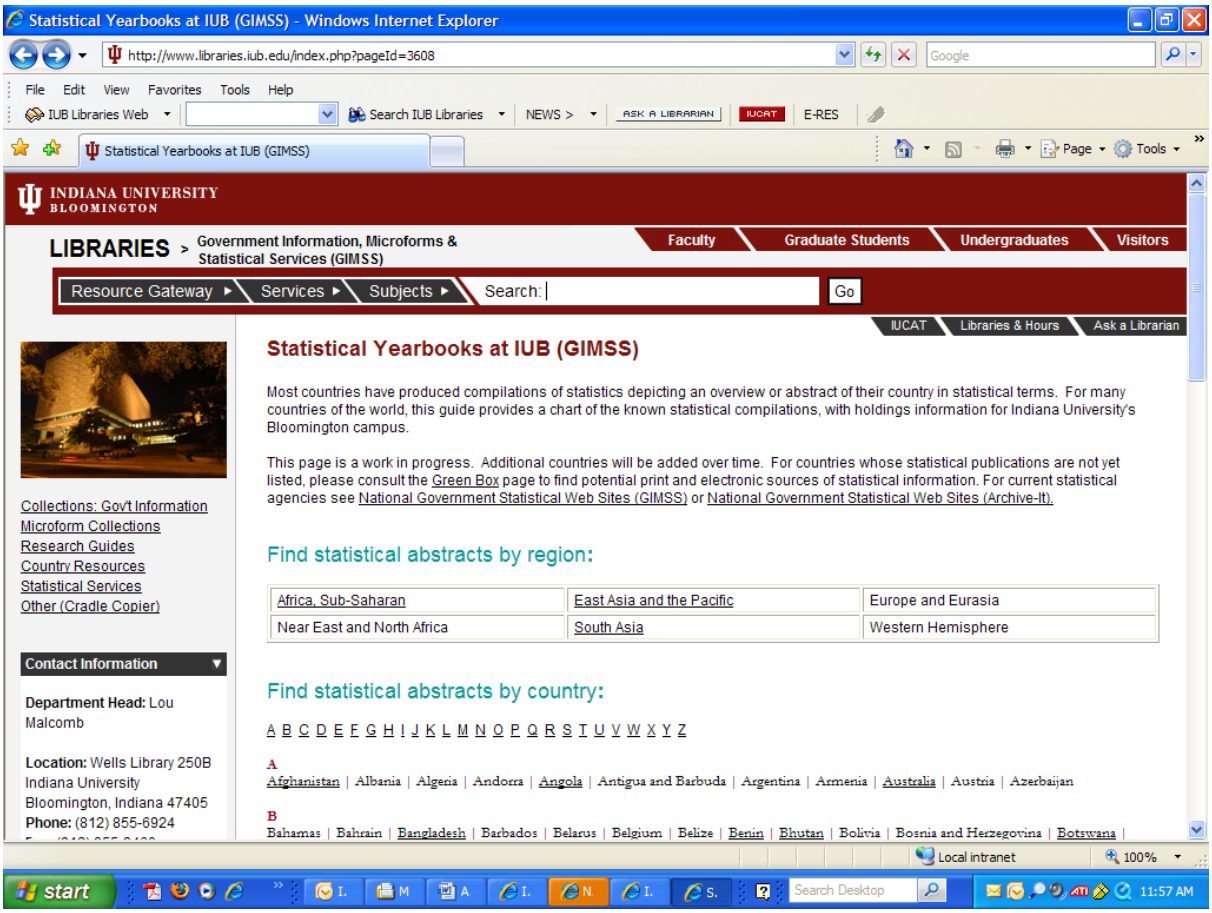
Le réseau des bibliothèques de IUB, bibliothèque de recherche servant une grande université publique dans le centre des Etats-Unis, s'est inscrite à Archive-It dès 2006, après une période initiale de planification entamée en 2005. Nous avons sélectionné des adresses URL pour construire trois fonds. Pour rester fidèle à la tradition du "Penser mondialement, agir localement", deux de ces fonds représentent un grand intérêt local. La collection de sites web de l'université, gérée par le personnel des archives de l'université, est la collection pour laquelle nous nous sentons le plus responsables. Le fonds "Indiana : État et administration locale", géré par la bibliothécaire responsable de la documentation sur l'État et l'administration locale, compte également beaucoup pour l'histoire de notre État et de notre région (nous pensons que ces deux fonds ne se recoupent pas avec

d'autres programmes d'archives, mis à part celui de la Wayback Machine [2]). Le troisième fonds est une sélection de sites gouvernementaux nationaux créés par des agences statistiques dans le monde entier. Le nombre de sites choisis pour archivage a évolué entre 70 et environ 200 tout au long de notre expérience. Nous avons privilégié en permanence les sites de pays en dehors de l'Europe occidentale, de l'Australie, de Nouvelle-Zélande, du Japon et d'Amérique du Nord, zones où les tentatives d'archivage sont naissantes ou déjà bien avancées [3].

Le choix des sites est basé sur la philosophie et la politique de développement des collections en place à l'Université d'Indiana depuis des années, grâce à la collecte de publications officielles de pays étrangers. L'idée que les sociologues, historiens et autres chercheurs comptent en permanence sur certains types de publications officielles (statistiques, recensements, plans d'aménagement, rapports sur l'éducation, l'environnement ou autres sujets d'un intérêt constant pour la recherche) a été le point de départ de notre sélection de ressources imprimées, puisque couvrir l'ensemble des publications officielles de nombreux pays s'avère impossible[4]. (La meilleure description de ces types de publications générales imprimées et leur localisation est probablement celle de Gloria Wesfall dans son "Guide to Official Publications of Foreign Countries"[5], qu'elle qualifie de "résultat d'une coopération internationale et des efforts de nombreux bibliothécaires dans le monde".)

Au fil du temps, les sites gouvernementaux se mirent à fournir une information de plus en plus à jour et de manière de plus en plus exhaustive et nous nous sommes associés aux bibliothécaires qui utilisent ces sites quotidiennement pour compléter les ressources imprimées. Nous nous sommes empressés d'ajouter à notre boîte à outils un puissant outil de référence accessible gratuitement : Governments on the WWW [6] de Gunner Anzinger. L'auteur a tenu ce site à jour entre 1995 et 2002. D'autre part, nous avons demandé à des catalogueurs d'ajouter les notices de sites web gouvernementaux choisis au catalogue, IUCAT et à OCLC Worldcat. Les notices comprenaient des liens persistents (PURLs) mais ne pouvaient bien sûr pas diriger vers une information entre-temps retirée du web. Cette sélection et le catalogage sont évidemment coûteux, à cause de l'attention requise par un personnel hautement qualifié. Nous avions donc hâte d'expérimenter d'autres méthodes de capture et de conservation de contenu numérique.

Le résumé de cette communication fournit les adresses de plusieurs pages web en rapport avec le fonds d'Archive-It que nous sommes en train de développer, comprenant une synthèse, une aide à la recherche, et le formulaire de recherche du fonds. Sur le site des bibliothèques de IUB, vous trouverez une page qui donne plus d'explications quant au contexte du fonds au sein de notre institution, ainsi que des liens vers le fonds d'archive et une page encore incomplète d'aide à la recherche du fonds d'annuaires statistiques[7]. Vous trouverez également des aides pour le site des bibliothèques et les fonds présents dans Archive-It.



Coopération :

Le projet interne :

Cette description de notre coopération démarre avec une décision des administrateurs des bibliothèques de IUB de rejoindre d'autres bibliothèques de recherche dans le projet RLG avec Archive-It en 2005[8]. À l'interne, le project était dirigé par les directeurs du développement des collections ainsi qu'un groupe de travail formé par les services des archives, du catalogage, des collections et du service public. Ils sont soutenus par l'assistant du directeur du développement des collections (la composition du groupe, qui n'inclut pas le personnel du Programme de la bibliothèque numérique ni des technologies de l'information, est importante parce qu'elle confirme que l'expertise technique est fournie par Archive-It). En 2006, le groupe de planification avait décidé d'organiser des collections pour Archive-It et de les représenter et était prêt à s'inscrire pour les trois fonds précédemment décrits.

Le groupe de travail a décidé de décrire les trois fonds de manière complète (au nom de chaque fonds) dans IUCAT et dans OCLC Worldcat. La création de métadonnées est minimale et on avait informé les usagers internes de l'existence des fonds avec une simple page web descriptive et une page donnant des conseils de recherche pour chaque fonds. Comme pour d'autres fonds de l'Université d'Indiana à Bloomington, nous avons

créé une brochure descriptive imprimée que nous continuons à distribuer. Les fonds sont aussi mentionnés sur la page d'accueil des bibliothèques ainsi que dans le rapport annuel 2006/2007, mis à la disposition du personnel enseignant, des administrateurs et des sympathisants de nos bibliothèques.

En même temps, un début d'évaluation des collections a été mené par chacun des responsables. Des liens ont été ajoutés ou supprimés selon les informations fournies lors des captures initiales ou de la collecte automatique (crawl). Nous avons étudié diverses fréquences de collecte, tout en nous assurant régulièrement que nous ne dépassions pas le budget prévu pour ce projet[9].

D'autres exemples de coopération à un niveau institutionnel ont consisté en une participation au premier rassemblement des partenaires de Archive-It à San Francisco en octobre 2007. Pour ce qui est de ce fonds en particulier, il existe également une coopération à une plus large expérience liée aux outils de recherche grâce à un partenariat expérimental entre Archive-It et le projet LibraryFind à l'Université d'État d'Orégon[10].

Un cadre plus large :

Tous les participants à des projets d'archivage du web sont confrontés à de nombreuses questions. En plus des questions liées à la sélection et aux technologies, des inquiétudes notables portent sur la conservation à long terme, le souci de ne pas répéter des efforts déjà fournis, le contrôle de la qualité et l'évaluation.

À petite échelle, peut-être même à micro-échelle, des contributions à des efforts d'archivage dans le monde entier, comme le fonds numérique de sites nationaux gouvernementaux de statistiques, représentent l'archivage de minuscules quantités d'adresses URL par rapport à l'étendue de projets lancés par les bibliothèques nationales et autres. Souvent, les conservateurs de petites collections ont nombre d'autres responsabilités dans le non-numérique et ne sont certainement pas des experts dans l'archivage du web (je pense que la plupart d'entre nous vont chercher des informations sur les différents projets, les techniques et les méthodes d'évaluation auprès de nos institutions ou sur le web, ce qui est un effort coopératif toujours en mouvement).

On peut trouver facilement nombre d'outils sur le web, comme le Bureau des systèmes d'information à l'Université de Harvard et sa page sur les ressources pour l'archivage du web[11]. Y sont réunies des informations sur les listes, droits IP, métadonnées et catalogage, archivage du web, ateliers et conférences, ainsi que d'autres informations générales sur les archives. D'autres travaux sur des efforts pilotes de description des données obligatoires dans les registres d'information numérique sont accessibles sur le web.

Pourtant, lorsqu'un usager lambda part d'une information actuelle et voudrait remonter le cours du temps pour localiser les versions antérieures d'un site, il lui est difficile de savoir si ce site a été archivé quelque part. La Wayback Machine, mentionnée plus tôt, fournit cette information pour les sites qu'elle archive.

En conclusion, rêvons un peu, avec une courte analyse de la valeur d'une création en coopération d'un registre de sites archivés, qui pourrait profiter aux archivistes autant qu'aux usagers.

Dans le cas du fonds numérique de sites nationaux gouvernementaux de statistiques, les principaux usagers sont ceux qui ont besoin d'aide dans la recherche d'une information statistique difficile à localiser au sein de notre institution. Nous utilisons ce fonds comme n'importe quel autre outil en notre possession. Nous espérons que les usagers extérieurs découvriront une information sur le site Archive-It au cours d'une recherche de documentation archivée et non parce qu'ils avaient besoin de renseignements sur l'accès à l'université. Cependant, nous serions heureux de contribuer à un registre ouvert, qui pourrait ne nécessiter qu'un adresse URL en guise de localisation, l'URL du site archivé et la ou les date(s) de capture, afin d'être utile aux chercheurs et aux développeurs.

J'espère que les membres du public partageront leurs idées sur ce sujet pendant la séance de questions-réponses ou bien après la session.

Merci.

---

[1] Les adresses URL du fonds numérique de sites gouvernementaux de statistiques ne font délibérément pas l'objet d'une limite. Si la capture n'est pas bloquée ou limitée par le site sélectionné pour archivage lui-même, la capture par défaut comprendra les deux premiers écrans du site.

[2] La Wayback Machine archive des sites web depuis 1996. Voir http://www.archive.org/web/web.php (consulté le 1 avril 2008)

[3] La liste initiale des sites possibles provient de la liste d'agences internationales de statistiques du Bureau de recensement américain. Voir http://www.census.gov/main/www/stat_int.html (consulté le 1 avril 2008)

[4] L'Université d'Indiana a de solides programmes d'études régionales. Les ressources de base sont donc fournies de manière poussée pour certaines parties du monde.

[5] Gloria Westfall, éd. Guide to Official Publications of Foreign Countries (Bethesda, Md.: CIS, 1990) American Library Association/Government Documents Round Table, 2e éd.

[6] Toujours disponible en anglais et en allemand, cette source propose des liens vers des sites web gouvernementaux par régions, pays ou catégories d'information telles "institutions dans le domaine des statistiques" ou "Parlements". Voir http://www.gksoft.com/govt/ (consulté le 1 avril 2008)

[7] Les bibliothèques de l'Université d'Indiana à Bloomington. "Statistical Yearbooks at IUB (GIMSS)". Voir http://www.libraries.iub.edu/index.php?pageId=3608 (consulté le 13 mars 2008)

[8] Fondé en 1974 par la New York Public Library et les universités de Columbia, Harvard et Yale, le programme RLG fait partie de OCLC depuis juillet 2006. Voir http://www.oclc.org/programs/about/default.htm (consulté le 1 avril 2008)

[9] La décision de limiter les captures à des collectes annuelles par exemple est due à des facteurs économiques internes. Actuellement financés de manière centrale, notre désir de pouvoir poursuivre, si nécessaire, le projet avec des fonds des collections de documents officiels a eu une influence sur les décisions quant au contenu.

[10] Archive-It a invité des institutions à participer à cette expérience, collection par collection, en février 2008.

[11] Office for Information Systems. "Web Archiving Resources" http://hul.harvard.edu/ois/systems/wax/resources.html (consulté le 13 mars 2008). Le jour de la consultation, le site n'avait pas été modifié depuis le 24 octobre 2006.