Date : 24/08/2008 (2nd Version)

La contenu Web de la documentation gouvernementale au Canada : Perspective d'archivage du Web par la Bibliothèque Nationale

GILLIAN CANTELLO and **JOHN STEGENGA**
Library and Archives Canada
Canada

Translated by/traduction faite par :
Julie Vovan
julie.vovan@videotron.ca

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

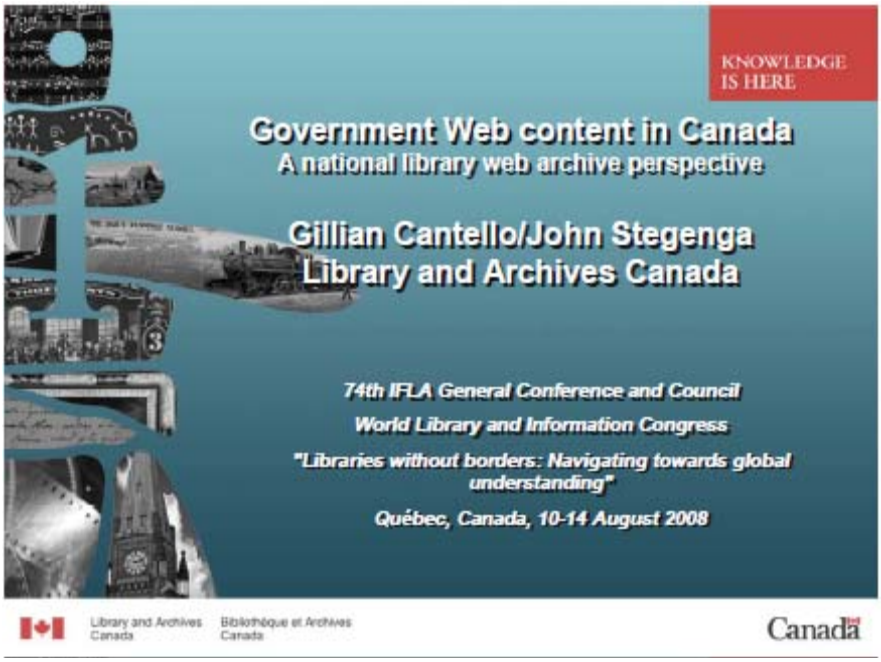
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

Résumé

Rendre accessible l'information de tous les paliers de gouvernement et de toutes les juridictions est un défi mondial au 21e siècle. Les gouvernements sont en général des éditeurs prolifiques de sites internet; et, comme les webmestres des secteurs commercial et privé, ils mettent à jour ou soustraient d'internet le contenu de leurs pages généralement aussi rapidement. Une partie de la solution globale visant à maintenir l'accès à l'information consiste en une sorte d'archivage permanent par lequel on cherche à conserver indéfiniment le contenu Web. Au Canada, conserver l'information gouvernementale est un travail qui relève de Bibliothèques et Archives Canada (BAC). Plusieurs projets sont en cours grâce auxquels on contribue à l'objectif général de conserver l'information des juridictions fédérale, provinciale et territoriale du Canada.

Les auteurs de l'article présentent les efforts du BAC investis dans ses projets de moissonnage du Web pour conserver l'information produite par les gouvernements; ils présentent aussi le lien entre cet investissement et le Dépôt légal dont le champ d'application a été élargi au début de 2007 pour inclure les publications en ligne. L'initiative principale, les *Archives du Web du Gouvernement du Canada* (AWGC), comprend une collection de sites internet provenant du domaine Web entier du gouvernement du Canada, sites recueillis deux fois l'an. Le AWGC est accessible au public depuis novembre 2007. Les auteurs décrivent les progrès des activités de moissonnage depuis la création du AWGC ainsi que l'infrastructure technique qui supporte le projet. On note aussi l'importante contribution, pierre angulaire au projet, du *International Internet Preservation Consortium* (IIPC). On y aborde les défis, qu'ils soient de nature technique, matérielle, légale ou autre et on ajoute quelques mots sur les progrès à venir. On mentionne aussi d'autres projets qui ont été initiés; ceci vise à donner au lecteur un goût de ces progrès futurs.

On étale aussi dans cet article le lien existant entre le moissonnage du web pour les sites gouvernementaux et les activités de Dépôt légal au Canada. Ayant été élargi pour inclure les publications en ligne et étant déjà applicable au gouvernement fédéral, le dépôt légal de publications Web spécifiques et leur conservation subséquente dans un environnement numérique de confiance, tout cela combiné au moissonnage de sites Web entiers permet au BAC de posséder une base importante d'information gouvernementale pour assurer une accessibilité et une conservation à long terme.



INTRODUCTION

Chers collègues, je suis heureuse d'être ici aujourd'hui pour vous présenter ce que nous, au Canada, considérons comme étant un projet innovateur : les *Archives du Web du Gouvernement du Canada.* Bien entendu, vous avez probablement tous décelé un brin de réserve dans cette introduction. Je suis certaine qu'il y a parmi vous des représentants d'autres pays qui se sont engagés, et peut-être le sont-ils depuis un certain temps, dans le même chemin que je décris aujourd'hui. J'attends avec impatience l'après-séance pour comparer mes notes avec vous et, naturellement, vous présenter à d'autres qui pourraient être curieux et curieuses envers ce projet, mais qui n'auraient pas encore initié quoi que ce soit dans leurs pays respectifs.

Les sources de connaissances gouvernementales, non gouvernementales, et intergouvernementales jouent un rôle prépondérant dans notre société de mondialisation. De plus en plus, ces sources naissent déjà en format électronique, ou elles sont numérisées pour en améliorer l'accès par les gens des quatre coins du globe. Les gouvernements à bien des niveaux, les institutions, les organisations non-gouvernementales et internationales, ainsi que les individus collaborent entre eux aujourd'hui sur une étendue locale, nationale, régionale, et internationale pour rendre ces ressources accessibles au public via l'Internet et pour assurer

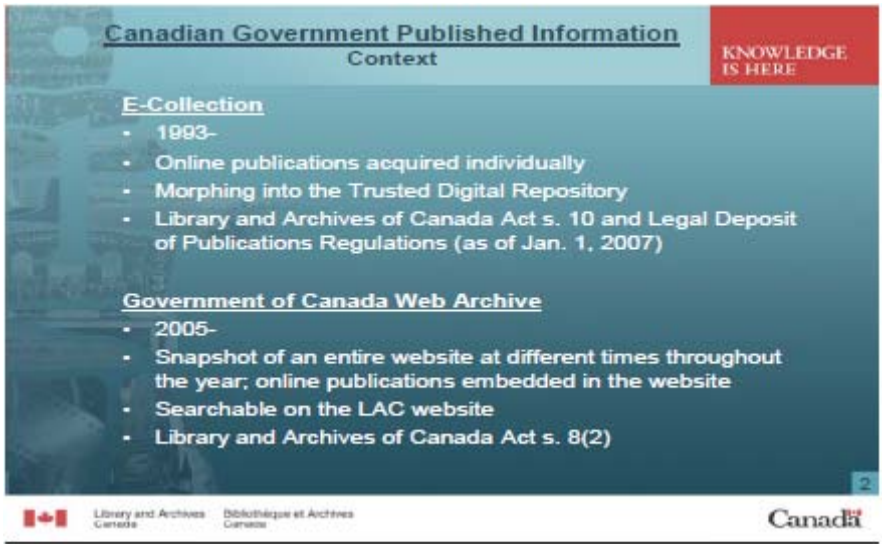
une conservation et un archivage adéquats dans le but d'en maintenir son utilisation par les générations futures.

Rendre accessible l'information de tous les paliers de gouvernement et de toutes les juridictions est un défi mondial au 21e siècle. Les gouvernements sont en général des éditeurs Internet prolifiques; et, comme les webmestres des secteurs commercial et privé, ils mettent à jour ou soustraient d'internet généralement aussi rapidement le contenu web de leurs pages. Au Canada, nous n'avons pas de statistiques sur l'information gouvernementale qui est fournie pas les sites web, que ce soit l'information procurée par le site lui-même ou par une source équivalente publiée de façon conventionnelle. Cependant, il n'est pas rare d'entendre que les webmestres du gouvernement du Canada soient d'accord avec les données d'estimation du *Internet Archive* (« Archives Internet ») voulant que la durée de vie moyenne d'une page Web soit de 44 à 75 jours[1]. Il n'est pas certain que cette estimation puisse viser les pages Web du gouvernement du Canada, mais selon les anecdotes et l'expérience, tout cela semble vraiment concorder.

Pour atteindre ce but, le présent article décrit deux projets intimement liés qui ont été initiés par Bibliothèques et Archives Canada : la Collection électronique[2] et les Archives du Web du Gouvernement du Canada[3]. Cependant, c'est sur cette dernière que l'on s'attardera, puisqu'elle implique quelques compétences inhérentes qui peuvent être admirablement utilisées dans le but de saisir et d'archiver l'information gouvernementale de façon optimale.

Ce sujet va à merveille avec le thème de cette séance : « L'information gouvernementale dans un contexte de mondialisation : la création d'archives numériques pour un accès amélioré. » Cependant, il faudrait comprendre que d'autres programmes du BAC, qui contribuent aussi au contrôle de l'information gouvernementale, créent un contexte dans lequel cette initiative particulière se retrouve : de tels programmes incluent l'élargissement du champ d'application des lois et règlements du Dépôt légal pour inclure les publications en ligne (l'application de la loi aux éditeurs du gouvernement fédéral ayant déjà été mise en place depuis bien des années), l'élaboration d'une collection numérique et le développement de systèmes pour la supporter. Tout cela mis ensemble représente les efforts investis de Bibliothèques et Archives Canada pour acquérir, conserver et maintenir l'accès à l'information gouvernementale à ce jour et bien loin dans le futur.

[1] Internet Archive (http://www.archive.org/web/web.php)
[2] E-Collection (http://www.collectionscanada.gc.ca/electroniccollection/003008-200-e.html)
[3] Government of Canada Web Archive (http://www.collectionscanada.gc.ca/webarchives/index-e.html)

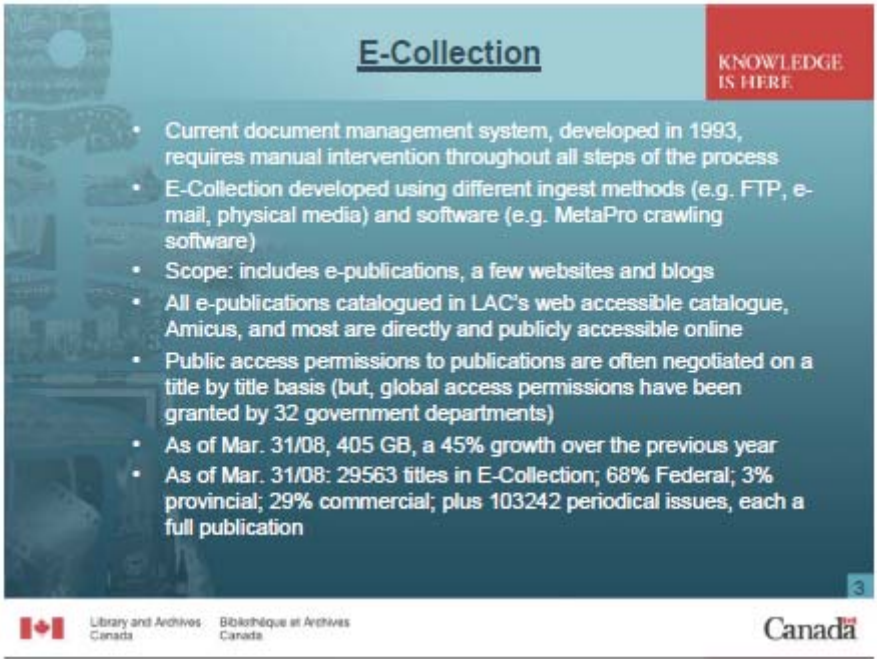


L'information publiée du Gouvernement du Canada : le contexte

Pour comprendre ce que les Archives du Web du Gouvernement du Canada sont essentiellement, il est primordial de faire main mise sur le contexte dans lequel cette archive s'est développée au départ et se retrouve aujourd'hui.

Pour donner une description générale brève, Bibliothèques et Archives Canada (BAC) possède maintenant deux systèmes qui permettent ensemble de saisir et de rendre accessible l'information gouvernementale disponible sur le Web. Le premier, la Collection électronique, est une archive numérique datant de 1993. Malgré qu'elle héberge des publications en ligne de bien de sources domestiques, une grande partie de ces archives est vouée aux publications produites par le gouvernement fédéral canadien. Puisqu'elle n'offre pas un environnement de conservation idéale ou véritable, des progrès seront faits pour la remplacer avec un authentique *Trusted Digital Repository*, sorte de dépôt numérique. Les lois et règlements du Dépôt légal récemment adoptés tombent sous la juridiction du BAC et s'appliquent maintenant ainsi qu'aux publications en ligne; cette législation supporte l'acquisition des publications en ligne domestiques, qui sont téléchargées et conservées dans ces archives.

De son côté, les Archives du Web du Gouvernement du Canada fait partie d'une crue plus récente. Alors que le but de la Collection électronique est d'archiver les publications Web individuelles, les Archives du Web du Gouvernement du Canada, lui, saisit des sites Web entiers de tous les ministères du gouvernement. Naturellement, les publications en ligne qui ont été archivées individuellement dans un système se retrouveront aussi sous forme de publications annexées aux sites Web archivés dans le second système. Tout ce qui est hébergé dans ce site Web – les publications, l'information générale, les formulaires, etc. —est accessible au public sur internet. De plus, puisque le Dépôt légal est régi par la Loi sur Bibliothèque et Archives Canada, la Loi régit aussi bien la collecte de sites disponibles en ligne.



En premier lieu, faisons un tour en arrière pour étudier en plus grands détails le système de base, le « *Electronic Collection : a virtual collection of monographs and periodicals* » (intitulé dans cet article « La Collection électronique »). Pendant presque quinze ans, ce système de gestion documentaire numérique fut la seule source du BAC dans lequel les publications domestiques en ligne canadiennes ont été acquises et archivées.

Bibliothèques et Archives Canada et son prédécesseur, la Bibliothèque nationale du Canada, ont toujours cru qu'il était vital de bâtir une collection de publications du gouvernement du Canada la plus exhaustive qu'il soit. Les responsabilités entourant ce rôle se sont élargies de plus belle en 2004 avec la fusion de la Bibliothèque nationale du Canada et des Archives nationales du Canada, son institution sœur dont le mandat inclut la gestion des documents administratifs du gouvernement fédéral. Depuis, Bibliothèques et Archives Canada voit son rôle général lié à la gestion de l'information gouvernementale au niveau fédéral, rôle probablement mieux déclaré dans le préambule de sa Loi d'institution statuant « que cette institution soit la mémoire permanente de l'administration fédérale et de ses institutions. »

Sans aucune intention de laisser pour compte la gestion de documents gouvernementaux, qui est en soi un exploit, cet article se concentre sur la partie publiée de l'information gouvernementale, c.-à-d., tout ce qu'un ministère du gouvernement produit dans le but d'être disséminé (et donc, une « publication ») au public. Bien qu'il existe un nombre considérable de publications conventionnelles ou analogues produites aujourd'hui, la présence sur le Web ou la disponibilité en format électronique de ces mêmes publications et autre type d'information sur Internet (c'est-à-dire publiées) du gouvernement du Canada s'amplifie année après année.

Comme je l'ai mentionné précédemment, BAC est en affaire depuis longtemps; il est facile d'affirmer qu'elle l'a été depuis plus de cinq décennies pour collectionner l'information gouvernementale publiée.

Au début, BAC (et son prédécesseur) collectionnait chaque année des milliers de publications gouvernementales produites de façon conventionnelle. Cependant, en 1993, elle commença à élargir son procédé d'acquisition sur une base expérimentale pour inclure les publications reproduites sur les sites Web du gouvernement du Canada.

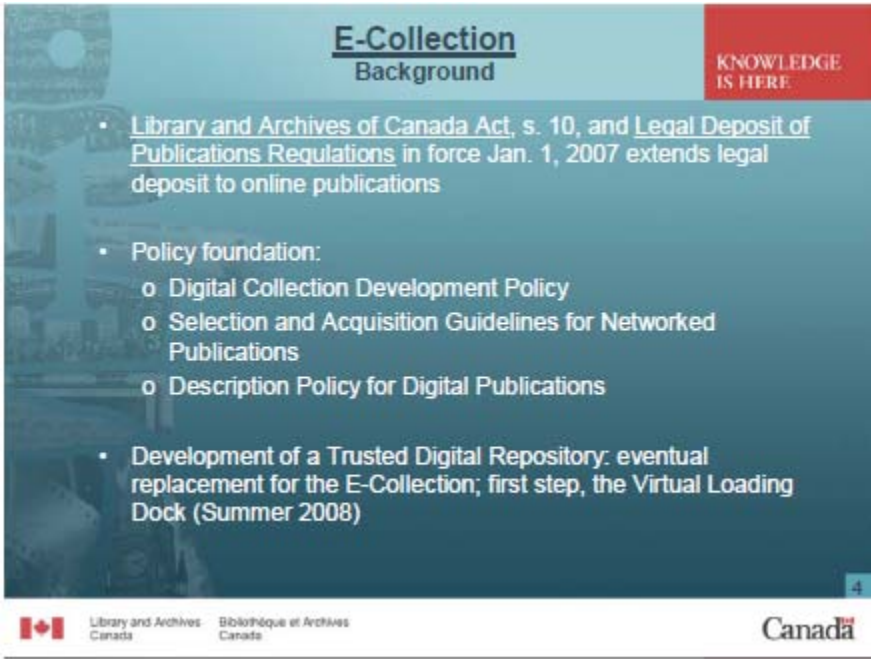
Ces publications gouvernementales en ligne ont été acquises et archivées dans la Collection électronique en utilisant une panoplie des méthodes d'acquisition. L'équipe a négocié titre par titre l'acquisition et l'archivage des publications en ligne du gouvernement du Canada (p. ex. Rapports, Journaux, Gazettes, Rapports annuels, et, sur une base très expérimentale, quelques sites Web et blogues) et leur accès. Au départ, les documents se voyaient versés dans la Collection électronique sur une base volontaire, mais ce procédé a changé depuis une modification de la loi, modification en vigueur depuis le 1er janvier 2007, dans laquelle le champ d'application de la loi et les règlements du dépôt légal alors existants a été élargi pour inclure les publications électroniques dont ceux du gouvernement du Canada. Cependant, leur acquisition était essentiellement un processus de négociation qui se faisait titre par titre, et c'est toujours le cas.

Nous avons assigné un descriptif aux publications hébergées dans la Collection électronique de la même manière que nous procédons pour les publications conventionnelles. Pour accéder aux publications numérisées contenues dans ces archives, la clientèle utilise, pour les retracer, le catalogue en ligne Amicus, un répertoire du BAC accessible sur Internet. Les clients sont ensuite dirigés vers un lien contenant la publication conservée dans la Collection électronique. Les organismes, qui publient, dictent le mode d'autorisation d'accès des utilisateurs au contenu complet de la publication. Les droits d'accès de chaque publication sont négociés avec les éditeurs, et ce, même avec les éditeurs gouvernementaux. BAC essaie de négocier les meilleures conditions pour sa clientèle sur une base générale (p. ex. une autorisation pour toutes les publications sans restriction de durée) telle qu'elle l'a déjà fait avec un bon nombre de ministères, mais ce n'est pas toujours le cas malgré les tentatives. Lorsque l'éditeur dépose les publications mais désire restreindre l'accès au public pour les copies de document dans la Collection électronique, nous établissons un accès limité. Cela signifie que les chercheurs désirant lire la publication peuvent le faire au siège social du BAC, aux terminaux spécialement conçus. À ces endroits, la clientèle peut lire la publication mais ne peut pas la télécharger ni l'imprimer.

Le système a admirablement résisté au passage du temps et continue de jouir d'une forte croissance annuelle. Au 31 mars 2008, le nombre calculé de publications s'élevait à environ 30 000. La composition de cette base de données constitue 68 % de publications du gouvernement fédéral, 3 % du provincial, et le reste, 29 % de sources commerciales et non-commerciales.

Cependant, le rendement de la Collection électronique est limité. Par exemple, en consolidant le système avec une équipe de onze personnes, l'acquisition des titres plafonne à 6000 nouveaux titres et 15 000 volumes de périodiques que l'on ajoute annuellement dans les archives. Il est ironique de voir que même si nous travaillons avec des publications numériques, la grande partie du travail d'organisation des documents de la collection dans les différents répertoires et structures de dossier est encore majoritairement faite manuellement. Mais surtout, pour ce qui touche la conservation, l'environnement

de la Collection électronique n'est pas idéal et ne constitue pas un répertoire électronique de confiance : nous en avons décrit les qualités ailleurs[4]. Conséquemment, malgré qu'il ait surmonté l'épreuve du temps, ce système sera mis au rancard et sera remanié pour devenir le *Trusted Digital Repository* lors de l'implantation complète de ce dernier.



Déjà mentionnées précédemment, toutes les publications archivées dans la Collection électronique sont acquises sur une base volontaire. Jusqu'au 1er janvier 2007, les lois et règlements concernant le Dépôt légal[5] sous la responsabilité du BAC n'atteignaient pas les publications en ligne. Cependant, grâce au fait que les lois canadiennes du dépôt légal incluent maintenant les publications en ligne, les conditions ont été modifiées. La partie légale du casse-tête est en place.

Les politiques, comme la *Politique de développement des collections numériques*[6] et les *Lignes directrices relatives à la sélection et à l'acquisition de publications en réseau*[7], constituent une deuxième pièce très importante de notre trousse d'outils générale. Un troisième énoncé de politique, le *Description Policy for Digital Publications* (« Politique de la description des publications numériques ») qui sera en place plus tard cette année, prépare une place pour un point de vue différent de l'accès bibliographique aux ressources en ligne. Cette politique tient compte des méthodes descriptives visant la gestion d'un grand nombre de données. Ensemble, ces règles complètent les lois et démontrent l'orientation vers

4 Trusted Digital Repositories: Attributes and Responsibilities (2002) (http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf)

5 Library and Archives of Canada Act (http://laws.justice.gc.ca/en/showtdm/cs/L-7.7) and the Legal Deposit of Publications Regulations (http://laws.justice.gc.ca/en/showtdm/cr/SOR-2006-337)

6 Digital Collection Development Policy (http://www.collectionscanada.ca/collection/003-200-e.html)

7 Selection and Acquisition Guidelines for Networked Publications (http://www.collectionscanada.ca/collection/003-206-e.html)

laquelle BAC entend se diriger.

La BAC attend de pouvoir concevoir une capacité d'acquisition plus forte pour initier une application systématique et sérieuse des lois du Dépôt légal. Ce rêve prend forme avec le *Trusted Digital Repository* (TDR) qui en est au stade de développement. On prévoit que la première partie de ce système, le *Virtual Loading Dock* ou la partie acquisition du TDR sera mise en fonction à l'été 2008.

Les composantes du *Virtual Loading Dock* incluent :

Quatre formes possibles d'acquisition : un format Web, courriel, FTP, et CD-ROM ou autres moyens de conservation envoyés par courrier régulier.

Cueillette des métadonnées.

Une fois que le programme d'acquisition sera établi et aura été testé dans des conditions où le nombre de demandes est optimal, les autres parties du TDR verront leur mise en fonction se poursuivre l'année prochaine :

Développement des parties « gestion documentaire » et « accès au public » du TDR.

Migration de toutes les publications de la Collection électronique au TDR.

Une collecte des publications en ligne possible, dont la provenance est le programme de moissonnage Archives du Web du Gouvernement du Canada.

L'acquisition des documents électroniques ministériels du gouvernement du Canada.

Tournons-nous maintenant vers la deuxième et plus récente initiative, les Archives du Web du Gouvernement du Canada[8].

Que sont au juste les Archives du Web du Gouvernement du Canada? Comme son nom l'indique, les AWGC sont une archive composée de sites Web de ministères, agences, commission et autres organismes similaires appartenant au gouvernement fédéral du Canada. Environ deux fois l'an, BAC lance une recherche par robot indexeur de tous ces sites Web. Jusqu'à présent, BAC a conduit trois recherches de ce genre. Le robot prend une « image » du contenu entier accessible au public des sites Web et les remet dans les archives. Puisque chaque site est en fonction au moment de la recherche, la copie est une représentation fidèle du contenu, de l'apparence du site ainsi que des liens tant internes qu'externes disponibles le jour, l'heure ou la minute à laquelle le robot en a fait une copie. Comme vous pouvez l'imaginer, plus le temps passe, la clientèle peut accéder aux générations successives des sites Web d'un ministère.

BAC débuta ses expériences sur les technologies de moissonnage du web en 2005. Les premières tentatives de collecte de site Web semblaient indiquer un avenir quelque peu positif : la méthode semblait extrêmement efficace pour la collecte d'une grande quantité de contenu Web dans un court laps de temps. Après avoir résolu la plupart des difficultés rencontrées au stade expérimental, BAC initialisa une première tentative sur le domaine Web du Gouvernement du Canada vers la fin de 2005. Le corpus de sites Web était quelque peu condensé et relativement facile (c'est ce que nous pensions à ce moment-là!) à repérer par la dernière partie typique de son adresse : « .gc.ca ». BAC terminait sa première collecte quelques semaines plus tard.

[8] Government of Canada Web Archive (http://www.collectionscanada.gc.ca/webarchives/index-e.html)

Même si nous pouvons programmer dans le robot d'indexation le nombre de niveaux des sites Web à capturer dans la recherche, nous avons dicté à notre robot d'aller le plus loin possible. Résultat : une image complète et exhaustive de la présence sur le Web des sites de chaque ministère accessible au public. On devrait cependant noter que le robot ne peut faire la collecte de tout le contenu Web. Par exemple, il ne pénétrera pas un pare-feu pour accéder au contenu de l'intranet d'un ministère. Il est aussi empêché d'accéder aux sites d'inscription exigeant la participation de la clientèle (p. ex. une page de recherche dirigeant les internautes vers une base de données de rapports; toute information Internet pour laquelle la clientèle doit payer et qui requiert une preuve inscriptible de paiement).

Comme la présentation ci-haut le démontre, les Archives du Web du Gouvernement du Canada comporte trois index : recherche dans le texte entier, par nom de ministère et par URL. Une preuve anecdotique indique que les utilisateurs et utilisatrices sont à ce stade-ci très satisfaits des index de recherche, si simplifiés soient-ils, pour trouver et localiser une myriade d'information à laquelle ils n'ont jamais eu accès auparavant.

L'acétate suivant montre une autre composante importante des archives : la possibilité de différencier le contenu Web archivé d'un ministère du contenu actuel. Puisque cette différenciation est primordiale pour la clientèle, non seulement dans cette archive-ci, mais aussi dans les autres environnements archivistiques et les engins de recherche en général, BAC a développé, il y a de cela quelques années, un menu grâce auquel on peut distinguer une page archivée de son contenu actuel. Sans la présence de ce menu vert criard annonçant l'heure précise à laquelle la page a été acquise et autre avertissement (p. ex. que les liens externes peuvent ne pas fonctionner) et offrant d'autres options (e.g. visionner les versions antérieures ou postérieures de la même page), la clientèle pourrait facilement confondre le contenu Web actuellement en vigueur avec celui qui est périmé.

Quelques observations intéressantes peuvent être faites lorsqu'on compare le moissonnage web initié par BAC au dépôt de titres individuels, approche qui se reflète dans la Collection électronique et où la responsabilité du dépôt repose sur les épaules de l'organisme éditeur. De par notre expérience, nous considérons la collecte de sitesWeb comme étant un instrument d'acquisition important du système de Dépôt légal; c'est une façon de rendre optimale la quantité d'information gouvernementale collectée et de réduire l'effectif des ressources humaines en gros, tant celui des ministères que celui des archives.

En comparant les deux projets, on peut aisément être poussé à remarquer leur relation de symbiose. Alors que les collectes par robot des Archives du Web du Gouvernement du Canada peuvent acquérir une grande quantité de contenu Web, quelques environnements restent inaccessibles (p. ex. base de données d'information; information pour laquelle la clientèle doit payer pour y avoir accès). Dans un tel cas, le Dépôt légal peut être utilisé pour s'assurer que l'information « sous clé » est déposée et rendue accessible en temps opportun. Dans un même esprit, BAC a observé que la collecte de publications en ligne elle seule surpasse de loin le dépôt de titres individuels puisque les publications électroniques, celles-là mêmes que les éditeurs gouvernementaux devraient déposer titre par titre, sont insérées dans le contenu Web que les robots indexeurs collecte à chaque moissonnage. De plus, bon nombre d'éditeurs gouvernementaux ont demandé au BAC de faire la collecte exclusivement, puisque ceci leur enlève la pression, en ce qui a trait aux ressources, provenant de la responsabilité de déposer chaque nouvelle publication.

Cependant, on devrait aussi noter qu'il existe des inconvénients au AWGC, des problèmes qui peuvent être résolus en développant le système pour le mieux. Par exemple, le dépôt des documents dans la Collection électronique assure aussi que les publications acquises sont classées selon les vraies méthodes de bibliothéconomie traditionnelles qui ont été éprouvées (p. ex., chaque rapport est catalogué

individuellement; les numéros de périodiques sont tous organisés en ordre chronologique à une seule place), tandis que dans les AWGC, la clientèle doit faire une recherche couvrant tous les sites Web archivés pour trouver les occurrences du rapport en question. Les rapports ne sont pas facilement isolés de leur contexte, ils peuvent se trouver en copie; dans le cas des numéros de périodiques, ces rapports ne sont pas emmagasinés en ordre séquentiel qui est facile à utiliser et à localiser. Cependant, cet inconvénient est largement compensé par le fait que bien du contenu Web a été stocké ensemble à une place et est accessible immédiatement par la clientèle qui est déjà familière avec les principes de navigation Internet. À dire franchement, les utilisateurs et utilisatrices sont reconnaissants de pouvoir trouver un tel contenu Web gouvernemental sur le bout des doigts dans les AWGC.

Bien que ces informations illustrent un nombre de manières par lesquelles les deux approches diffèrent, il est évident que lorsque leurs potentiels sont mis à contribution ensemble, ces systèmes constituent un développement encore meilleur pour assurer l'ultime accessibilité de l'information gouvernementale.

Les Archives du Web du Gouvernement du Canada : les fondements

Le fondement juridique régissant la collecte des sites Web canadiens, dans l'intérêt de la postérité, réside dans la *Loi sur la Bibliothèque et les Archives Canada*[9], une loi relativement récente qui est entrée en vigueur en avril 2004. Selon la section 8(2) de cette loi, BAC est autorisée à faire une telle collecte. En effet : « Pour l'application de l'alinéa (1)*a*), l'administrateur général peut, à des fins de préservation, constituer des échantillons représentatifs, selon les modalités de temps ou autres qu'il détermine, des éléments d'information présentant un intérêt pour le Canada et accessibles au public sans restriction dans Internet ou par tout autre média similaire. »

[9] Ibid. s. 8(2)

Comme discuté précédemment, le Dépôt légal et la collecte sont deux dispositions différentes dans la Loi.

Vu le jeune stade des connaissances sur le moissonnage au Canada à cette époque, la Loi serait remaniée de telle sorte que la collecte deviendrait aussi une partie intégrale des dispositions sur le Dépôt légal, si un tel remaniement existait aujourd'hui. D'une certaine manière, la loi sur le Dépôt légal est quelque peu anachronique en ce sens qu'on y définit le concept « publication en ligne » comme étant similaire aux publications publiées en format conventionnel. Lorsque cela est considéré conjointement avec les bénéfices que représentent une méthodologie et une technologie d'acquisition de masse du contenu Web, il est tentant d'émettre une hypothèse sur la façon dont la loi sur le Dépôt légal pourrait être structurée dans le futur. Serions-nous en train de prédire un moment auquel ce dépôt serait un mélange du contenu Web acquis par BAC et d'une approche traditionnelle au procédé d'acquisition dans laquelle la responsabilité de faire un dépôt de document repose sur les épaules des éditeurs?

Les politiques seront toujours requises pour compléter un projet dans sa totalité. Dans un tel cas, BAC a mis en place deux politiques documentaires clés, la *Politique sur le développement des collections numériques*[10] (mentionnée précédemment) et les *Directives de sélection et d'acquisition des sites web canadiens*[11], qui valent la peine d'être étudiées. À cela, nous ajouterons d'ici peu un énoncé concernant la stratégie de collecte qui complètera la direction vers laquelle BAC se dirige.

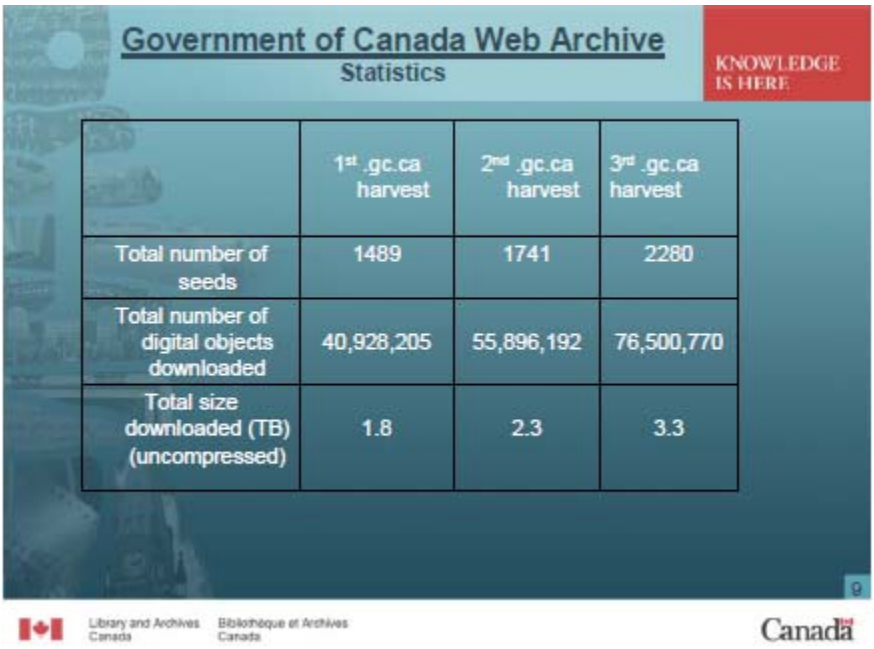
En passant, BAC a aussi entrepris la collecte et l'archivage des sites Web des gouvernements provinciaux et territoriaux. Ces recherches particulières sont lancées seulement sur une base annuelle. Les sites Web archivés possèdent un accès sécurisé, et ils ne sont donc pas accessibles au public ou même à l'interne avant que BAC ne puisse négocier des ententes avec les gouvernements respectifs ou trouver des moyens pour collaborer avec ces partenaires afin d'en ouvrir l'accès.

Enfin, il vaut la peine de noter que BAC a aussi ajouté à son système expérimental de collecte de sites Web, la collecte sélective de domaines plus petits, les événements spéciaux et autres applications Internet assorties (p. ex., blogues, Facebook, Wikipédia, YouTube). Ces dernières applications font tout particulièrement partie des expériences archivistiques dernier cri qui sont conduites même au gouvernement. Je voudrais réitérer que les discussions se poursuivent toujours sur la façon dont ces applications Internet pourraient être acquises efficacement, et introduites par un accès interne ou même, grâce à des négociations, rendues accessibles publiquement sur l'Internet.

---

[10] Op. cit. (http://www.collectionscanada.ca/collection/003-200-e.html)

11 Selection and Acquisition Guidelines for Canadian Web Sites (http://www.collectionscanada.ca/collection/003-203-e.html#e)



Les statistiques d'archivage des sites Web

Malgré que les statistiques semblent quelque peu parler d'elles-mêmes, elles méritent quelques commentaires.

Le total du nombre des sources (adresses Internet)

Comme vous pouvez le constater, chaque recherche successive par le robot suit une courbe de croissance. Cela parle du fait que même dans un environnement relativement « limité » de sources ou liens Internet, nous découvrions constamment de nouveaux sites Web au fil du temps, sites dont nous n'avions pas connaissance auparavant. Cela était dû en partie au fait que le public général ainsi que les ministères qui ont créé des documents, soucieux de s'assurer que leur site Web en entier soit représenté tel qu'ils le connaissent, nous annonçaient des sites Web connexes manquants. Il va sans dire que dans un environnement aussi grand que le gouvernement et dans un environnement où la gestion de sites Web s'apparente à une conquête de terres sauvages, les ministères n'agissent pas nécessairement de concert. Ceux-ci peuvent aussi perdre la trace des personnes qui créent leurs sites ministériels. La gestion des sites Web semble rarement centralisée.

Le nombre total des produits numériques

Ces chiffres ont aussi augmenté au fil du temps. Leur croissance atteste non seulement de l'efficacité grandissante de nos collectes par robot à travers notre expérience, mais aussi de la grande quantité indiscutable de contenu Web que BAC acquiert par ses recherches. Bien que le doublement du contenu d'une collecte à une autre existe, il subsiste tout de même une croissance nettement significative de la quantité de contenu. C'est peut-être une preuve que l'utilisation de l'Internet par les gouvernements, comme moyen de communication pour informer les citoyens et citoyennes, est florissante.

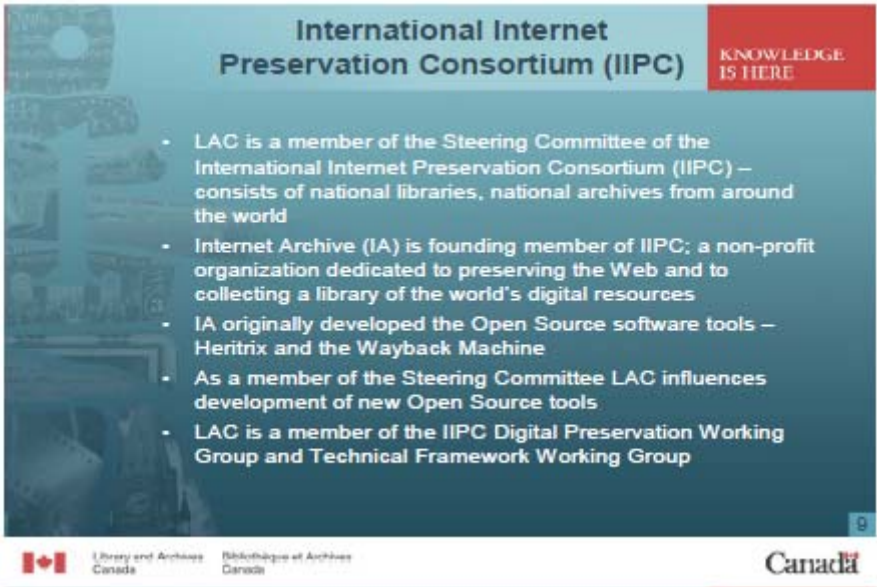
À noter aussi le cas des publications potentielles faisant partie du contenu Web. Comme vous le savez, un site Web gouvernemental se compose d'informations générales ainsi que des publications incluses dans le site. Alors que nous ne détenons pas de chiffres précis sur le nombre de publications dans chaque collecte, nos expériences sur quelques sites nous ont aidés à découvrir que, du point de vue « publications », il en existe littéralement des milliers dans un site Web, un nombre plus grand de ce qui nous a été donné par le dépôt titre par titre requis par la *Loi sur le Dépôt légal*. Plus tard, je dirai quelques mots sur le potentiel que nous voyons dans l'utilisation du moissonnage du Web en tant qu'une technique d'acquisition alternative de Dépôt légal.

La taille totale

Nous croyons que la grandeur totale des 3.3 TB (non compressés) du contenu web du gouvernement fédéral est quelque peu surprenant pour un pays comme le Canada dont la grandeur est modérée. Comme vous pouvez le constater, chaque collecte continue de suivre une croissance importante. Nous ne pourrions ni n'aurions l'audace de prédire à ce stade-ci où cette croissance s'arrêtera.

Pour l'année qui vient, nous prévoyons installer un autre programme - le *Smart Crawler* - qui pourrait diminuer la taille du système à chaque collecte. Comme mentionné auparavant, le doublement du contenu Web se manifeste de collecte en collecte. Le programme informatique que nous sommes en train de considérer comparera à chaque collecte le contenu actuel du site Web avec le site acquis lors de la collecte précédente; à moins qu'une variation ne soit décelée, le robot indexeur ne copiera pas le site Web actuel. Si quelque chose a changé sur le site Web, la clientèle verra tout de même le site Web entier, le contenu étant simplement fusionné en parties avec le contenu ancien et inchangé lors de la dernière collecte.

Nous croyons que cela aura pour effet de réduire le nombre actuel de documents archivés à chaque collecte, d'améliorer, en toute probabilité, la performance du système et de nous procurer des statistiques sur le taux de changement de contenu pour chaque site. Ce dernier élément, en l'analysant, pourrait nous mener à décider de faire une collecte des sites Web qui changent fréquemment et plus souvent que ceux qui ne changent presque jamais.



International Internet Preservation Consortium (IIPC)

BAC a commencé à suivre le chemin de l'archivage du contenu Web gouvernemental en étant inspiré de près par les développements à l'échelle internationale, dont elle prit connaissance, grâce à son statut de membre à *l'International Internet Preservation Consortiumm* (IIPC). Sans ce support, ce partage des technologies et de l'information, BAC ne serait pas où elle l'est aujourd'hui.

Le coût implique le développement de certaines parties comme le standard concernant l'accès, ainsi que la contribution de nos programmes élaborés aux membres du IIPC. Ce coût en a valu son pesant d'or.

Les défis et les sujets de réflexion

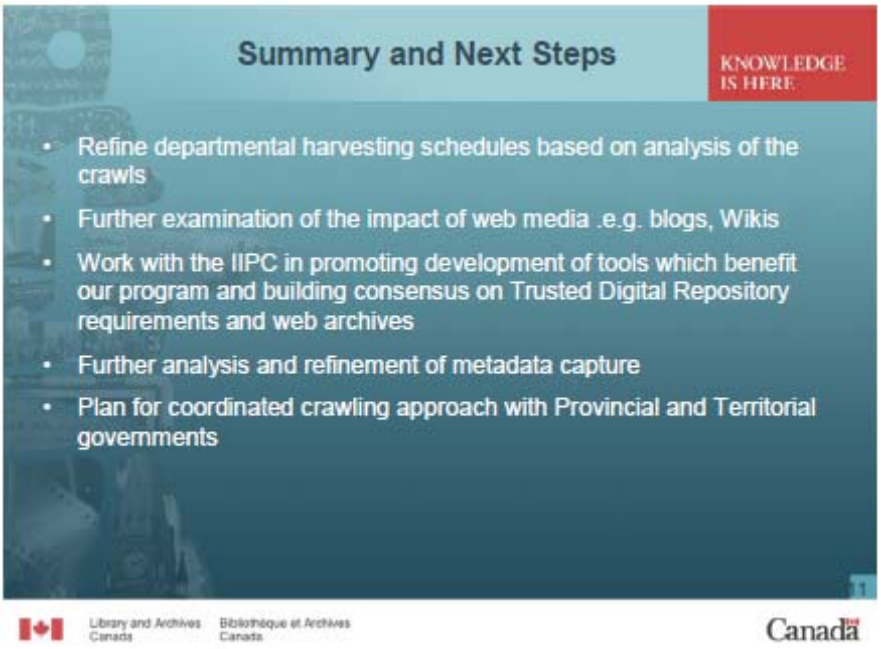
Alors que nous nous sommes engagés dans le processus de collecte de contenu Web des sites gouvernementaux depuis maintenant un peu plus de deux ans, nous avons certainement beaucoup appris. Mais de plus, nous avons toutes sortes de questions auxquelles, j'en suis certaine, nous trouverons les réponses en travaillant sur les défis qui y sont impliqués.

Les défis techniques existeront toujours. Plusieurs, dont l'élimination du doublement du contenu Web (c.-à-d. chaque collecte double la plupart des recherches précédentes), seront relevés par l'installation d'un programme, solution créée par un des membres du IIPC. Trouver des moyens brillants d'économiser un peu d'espace sur l'ordinateur améliore la rapidité à laquelle les utilisateurs et utilisatrices peuvent retracer l'information dans les archives. D'autres défis ayant de mêmes résultats sont de nature plutôt interne (p. ex. mémoire disponible, allocation de temps d'utilisation d'ordinateur). De plus, puisque la clientèle n'est pas la seule à exprimer un intérêt accru pour le contenu de ces archives, les ministères qui ont créé des sites veulent aussi que les AWGC soient modifiées. Un tel de ces ministères a demandé s'il existait un moyen, dans une recherche sur Google, de séparer les résultats acquis du contenu actuel des sites Web touchés et les versions que BAC a archivées. Pour les ministères, il est important que la clientèle d'aujourd'hui ait accès au contenu actuel et non celui d'hier. C'est un point important à considérer. À ce jour, le contenu du AWGC ne peut se retrouver dans Google malgré que la page d'accueil même le soit. Un autre élément à considérer pour la clientèle et, jusqu'à un certain point, les ministères créant les documents, est de savoir si la clientèle peut accéder aux données contenues dans les bases de données à l'intérieur du site. Il arrive souvent de voir que beaucoup de ministères installent des bases de données sur leurs sites Web, bases qui consistent en une collection de publications, de rapports, de données sur les programmes, etc. Puisque le robot indexeur conduisant une collecte s'arrête à la page de recherche qui introduit une base de données, c.-à-d. une sorte de point d'inscription (il en existe beaucoup, mais chaque point d'accès est caractérisée par le fait que la clientèle doit inscrire des termes de recherche ou des données indiquant leur présence), une grande quantité d'information gouvernementale incluse dans ces bases de données ne peut tout simplement pas être archivée et, de ce fait, elle reste inaccessible. On discute beaucoup à l'échelle internationale pour cerner la faisabilité d'un tel acte d'accès à ces bases de données. Cependant, à la fin de l'automne 2007, lorsque Google a annoncé que l'entreprise avait découvert une façon d'obtenir un accès au contenu de ces bases et de les indexer, la possibilité qu'une telle fonction puisse être adaptée à la collecte Web a vu le jour.

Comme mentionné précédemment, BAC a découvert que le moissonnage du contenu Web gouvernemental satisfait nos attentes plus que le petit nombre de publications en ligne que nous devrions théoriquement recevoir par dépôt par les ministères créateurs, titre par titre. Il est probable que même en ayant les meilleures intentions, la collecte de contenu Web surpasse le dépôt individuel de publications. De plus, il y a le contexte des publications ainsi que toute l'information les entourant, information non classée en tant que document. Serait-il alors possible de se demander s'il serait concevable que les versions postérieures des règlements du Dépôt légal intègreront dans le futur l'obligation traditionnelle de dépôt dans le AWGC où la responsabilité reste sur les épaules de l'éditeur et où l'acquisition est initiée par l'agence de dépôt légal, ce qui enlèverait alors une telle obligation à l'éditeur?

Il existe aussi un lien intriguant entre les collectes de contenu Web public que le gouvernement crée et les documents des transactions d'affaires mis en arrière-plan qui enrichissent l'information publique. Comme les publications, les documents qui étaient traditionnellement sous forme papier, se retrouvent maintenant en version électronique. BAC prévoit cette année la migration de la première série de documents électroniques vers le *Trusted Digital Repository* (malgré qu'ils soient versés dans une section différente), celui-là même dans lequel les publications en ligne apparaissent déjà. La présence de ces deux environnements (c.-à-d., les publications et les documents gouvernementaux) dans un seul système offrira au fil du temps à la clientèle une collection d'information entièrement intégrée.

Enfin, BAC se questionne sur l'utilisation du AWGC. Qui utilise réellement les données? Quel est le portrait type d'un utilisateur ou une utilisatrice? Est-ce que les habitudes d'utilisation changeront à travers le temps? Est-ce que l'utilisation publique aura un effet domino sur la façon dont les gouvernements publieront et emmagasineront l'information accessible au public? Il est trop tôt pour répondre à ces questions à ce stade-ci, mais néanmoins, ce sont des sujets qui restent à être explorés par des réflexions absorbantes. Il se peut qu'au prochain congrès de l'IFLA, nous ayons cueilli assez d'observations pour rapporter les faits sur les habitudes de la clientèle du AWGC.



Résumé et prochaines étapes

Comme je l'ai mentionné dans l'introduction de cet article, « rendre accessible l'information gouvernementale de tous les ordres du gouvernement dans toutes les juridictions est réellement un défi mondial au 21e siècle. » Ce que Bibliothèque et archives Canada a réussit par les deux projets décrits dans cet article est de faire un pas simple et rudimentaire vers l'atteinte des objectifs d'acquisition, d'archivage et d'accessibilité à l'information gouvernementale du gouvernement fédéral du Canada. Je mets l'emphase sur le fait que ceci constitue « un pas » et non une solution, pour la simple raison que rendre accessible l'information gouvernementale restera toujours un défi, malgré les solutions si

brillantes que nous puissions trouver pour cette gestion documentaire. Les gouvernements continuent de produire une quantité d'information qui change rapidement; dans un environnement numérique, cette information peut et sera communiquée tant de différentes façons que nous ne pouvons l'imaginer. Non, ceci constitue seulement le premier pas.

Les points à considérer dans cet acétate témoignent des défis qui se présenteront à nous. En analysant les collectes par robot indexeur, nous pourrons être en mesure de mieux structurer les collectes de façon à capturer un plus grand nombre d'images de certains sites gouvernementaux, alors que d'autres sites qui sont modifiés moins souvent rendent nos collectes moins larges. Serons-nous capables de répondre au défi qui consiste en la collecte de nouveaux produits médiatiques (p. ex., blogues, Wikis) que les gouvernements commencent à utiliser, ou même de ces sources médiatiques non gouvernementales qui contiennent des commentaires sur les actes des gouvernements pour que les futurs chercheurs et chercheuses puissent avoir une image entière dans son contexte de la façon dont les citoyens et citoyennes réagissent à ce qu'ils lisent?

BAC continuera, bien certainement, de collaborer avec la communauté internationale qui a initié des projets de collecte Web par robot indexeur. C'est ce chemin même qui nous rapprochera éventuellement de notre objectif de « la mondialisation de l'accès à l'information gouvernementale .» Serait-ce un rêve d'imaginer un standard universel d'accès à ce type d'information et de conservation? Est-ce que les futures lois du Dépôt légal seront modifiées assez rapidement pour accommoder les nouvelles technologies?

De plus amples analyses et un raffinement des collectes de métadonnées sont encore aussi un point situé au premier plan de notre liste de priorités. En ce sens, toute amélioration à l'extraction de métadonnées réduit l'intensité du nombre de ressources faisant l'objet d'une description et ceci aide la clientèle à trouver ce qu'elle cherche. Les index actuels dans le AWGC fonctionnent bien, mais ils pourraient être considérés comme étant rudimentaires. Tout moyen additionnel d'améliorer ces index serait utile à la clientèle.

Pourquoi s'arrêter là? Au Canada, nous sommes certainement en train de progresser vers une collecte plus universelle, non seulement celle visant l'information gouvernementale à l'ordre fédéral, mais aussi vers une collecte plus systématique de l'information provenant des provinces et des territoires. Les prochains objectifs pourraient consister en l'élaboration d'ententes de collaboration avec chacune de ces juridictions pour rendre le contenu accessible au public. De la même façon, nous étudions présentement la possibilité de faire une collecte de la présence Web entière du pays.

En résumé, les deux projets que j'ai décrits ici ne sont réellement qu'un petit pas vers le but de cette séance : rendre accessible l'information gouvernementale à la clientèle. C'est une aventure en constante évolution; en fin de compte, elle est imprévisible. Les éditeurs gouvernementaux, comme tout éditeur du 21^{e} siècle, continueront d'évoluer, et les suivront de façon permanente dans leur développement, nous-mêmes évoluant pour répondre à ces défis.

Information biographique des auteurs et présentateurs

Gillian Cantello, Directrice générale à la Direction du patrimoine de l'édition à Bibliothèques et archives Canada, est impliquée avec le BAC depuis 2002. Elle a assuré une aide dans des fonctions variées, incluant les programmes d'accès à l'information gouvernementale.

John Stegenga a été associé d'une manière ou une autre aux publications gouvernementales ainsi qu'au Dépôt légal au BAC depuis 1974. Il travaille présentement à la Direction du patrimoine de l'édition.