Date : 27/05/2008

Government Web content in Canada
A national library web archive perspective

Gillian Cantello and **JOHN STEGENGA**
Library and Archives Canada
Canada

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

Session theme: "Globalization of government information: creating digital archives for increased access"

Abstract:

Rendering access to government information at any level of government in any jurisdiction is indeed a 21st Century global challenge. Governments are generally prolific web publishers and, very similar to the commercial or private website creators, generally update and/or remove their web content as quickly. Part of the overall solution to the maintenance of access is some form of permanent archiving that seeks to preserve web content indefinitely. In Canada, the work of preserving government information is being undertaken by Library and Archives Canada (LAC). Several initiatives are underway which jointly contribute to the overall goal of preserving information from Canadian federal, provincial and territorial jurisdictions.

This paper explains the nature of the investment in web harvesting which Library and Archives Canada has made in order to preserve government-produced information, and its relationship to Legal Deposit which in early 2007 was broadened to include online publications. The principal initiative, the Government of Canada Web Archive (GCWA), comprises a collection of bi-annually harvested

websites of the entire Government of Canada web domain, publicly accessible since Nov. 2007. The progress of harvesting activity since the Archive's inception in early 2006 is described, as well as the underlying technical infrastructure. The important foundation the International Internet Preservation Consortium has provided is noted. Challenges, be they technical, resource, legal or otherwise related are also touched on, as well as a word about future developments. Other initiatives are also mentionned to give the reader a flavour of what future developments may look like.

Also developed in this paper is the relationship that web harvesting of government web sites has with Legal Deposit activity in Canada. Having been broadened to include online publications and already applicable to the federal government, Legal Deposit of specific web publications with subsequent storage in a trusted digital environment combined with the harvesting of entire websites will ensure that LAC will have a significant base of government information to ensure long term access and preservation.

Introduction

Dear colleagues, I am delighted to be here today to present what for us here in Canada is considered a bold new initiative, the **Government of Canada Web Archive**. Of course you may all have caught the faintest hint of a qualification in this opening sentence as I am certain that there are among you representatives of other countries who are engaged – and may have been engaged for some time – in the same pursuit that I am describing today. After the session I look forward to trading notes with you and naturally introducing you to others who may be curious about this initiative but may not have yet started anything in their respective countries.

Government, non-governmental and intergovernmental knowledge resources play an important role in our global society. Increasingly these resources are either born digital or are being digitized for enhanced access by people everywhere. Governments at many levels, institutions, non-governmental and international organizations, as well as individuals are collaborating locally, nationally, regionally and internationally to make these resources accessible to the public via the Internet and to ensure that they are properly preserved and archived for sustained use by future generations.

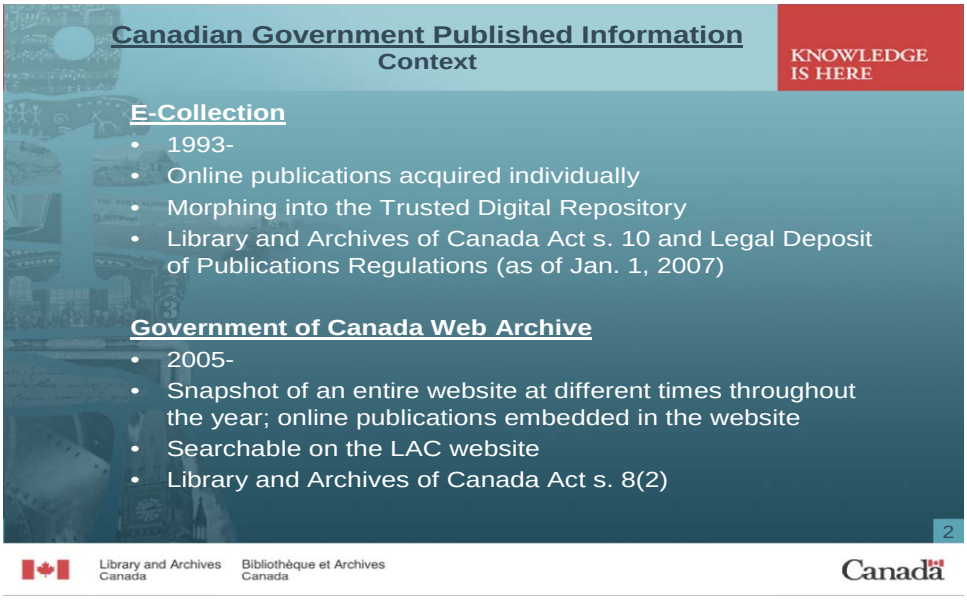
Rendering access to government information at any level of government in any jurisdiction is indeed a 21st Century global challenge. Governments are generally prolific web publishers and, very similar to the commercial or private website creators, generally update and/or remove their web content as quickly. In Canada, how much government information is served via website, either on its own, or with a parallel conventionally published source is unknown. However, it is not uncommon to hear Canadian government web masters agreeing with Internet Archive's estimate that the average life span of a web page is 44-75 days[1]. Whether this range statistically holds for Government of Canada web pages is uncertain, but anecdotally or experientially it does seem to fit. Part of the overall solution to the maintenance of access is some form of permanent archiving that seeks to preserve web content indefinitely. In Canada, the work of preserving government information is being undertaken by Library and Archives Canada (LAC).

To this end, this paper describes two closely intertwined initiatives undertaken by Library and Archives Canada: the **E-Collection**[2] and the **Government of Canada Web Archive**[3]. However, it is primarily the latter on which the focus will be placed as it holds certain inherent powers that lend admirably to the goal of optimal capture and archiving of government information.

The topic fits nicely into this session's theme: "Globalization of government information: creating digital archives for increased access". However, it should be understood that other LAC programs, also contributing to the control of government information, form the context in which this particular initiative finds itself – these are the broadening of existing legal deposit legislation to include the online publications (the application to federal government publishers having already been in place for many years), the building of a digital collection, and the development of systems to support it. Combined, this represents the investment that Library and Archives

[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)

Canada has made to acquire, preserve and maintain access to government information today and well into the future.

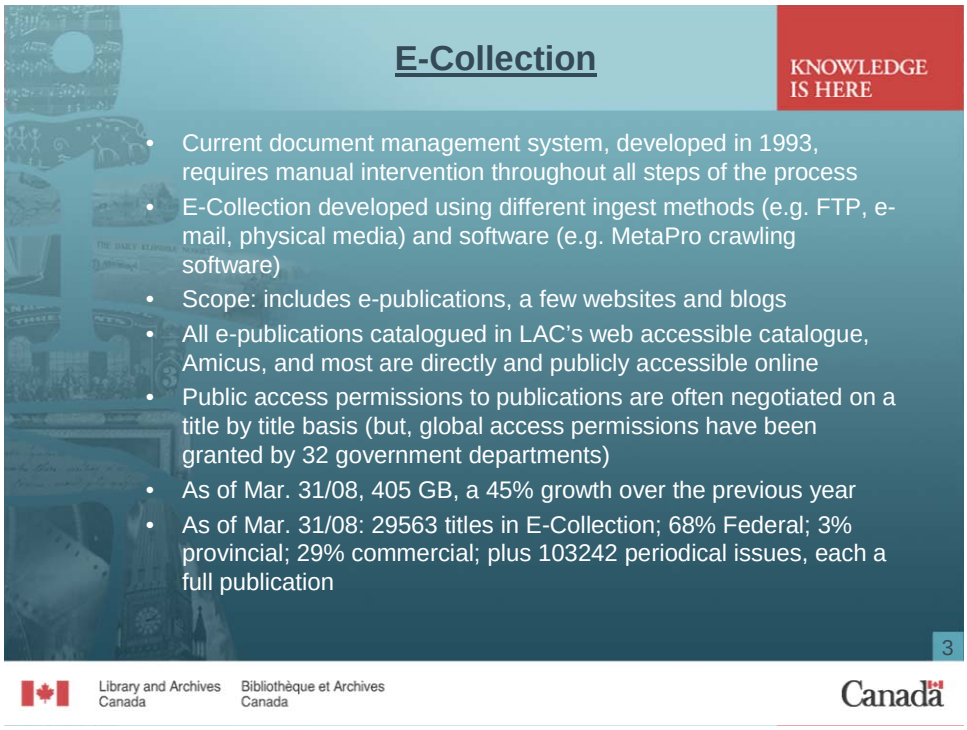


Canadian Government Published information – the context

To understand what the **Government of Canada Web Archive** is all about, it is important to understand the context in which it was originally developed and finds itself today.

By way of a quick overview, Library and Archives Canada (LAC) currently has two systems which together enable us to capture and make accessible web-based government information. The first system, the **E-Collection** is a digital archive dating back to 1993. Although it houses online publications from many domestic sources, a large portion is devoted to publications produced by the Canadian federal government. As this is not an ideal or true preservation environment, development is underway to replace it with a bona fide Trusted Digital Repository. The acquisition of domestic online publications which are loaded and stored in this archive is supported through LAC's recently enacted expanded Legal Deposit legislation that now covers online publications as well.

The **Government of Canada Web Archive** on the other hand is of more recent vintage. While the purpose of the E-Collection is to archive individual web-based publications, the Government of Canada Web Archive instead captures entire websites of all departments of this government. Naturally, online publications which are archived individually in one system, will also turn up as embedded publications in the websites archived by the other system. Everything – publications, general information, forms, etc – housed in this website is accessible to the public via the Internet. And, as the Library and Archives of Canada Act supports Legal Deposit, the Act also supports online web harvesting.



First, let's go back and look in greater detail at the base system, the **Electronic Collection: a virtual collection of monographs and periodicals** (in this paper referred to as the **E-Collection**). For nearly 15 years, this digital management system has been LAC's only source through which domestically produced Canadian online publications have been acquired and archived.

Library and Archives Canada, and its predecessor, National Library of Canada has always considered it vital to build as comprehensive a collection as possible of Government of Canada publications. This role has broadened even more with the amalgamation in 2004 with National Archives, the sister institution, a portion of whose mandate includes responsibility for management of the federal government's business records. Since that date, Library and Archives Canada sees its overall role with respect to the management of government information at the federal level, possibly best stated in the Preamble to its founding Act: that it "serve as the continuing memory of the government of Canada and its institutions".

Although by no means intending to give short shrift to the management of government records which in itself is no mean feat, this paper focuses on the published half of government information. That is anything that a department of government produces meant for dissemination (hence "publication") to the public. Although there are still a considerable amount of conventional or analog publications being produced, the web presence or digital delivery of the same publications and other types of web information (ie. published) of the Government of Canada increase steadily from year to year.

As mentioned previously, LAC has been in the business of collecting published government information a long time, easily over five decades. In the early years, LAC (and its predecessor) gathered thousands of conventionally-produced government publications each year. However in 1993 it began to expand this intake

on an experimental basis to include publications produced on Government of Canada websites.

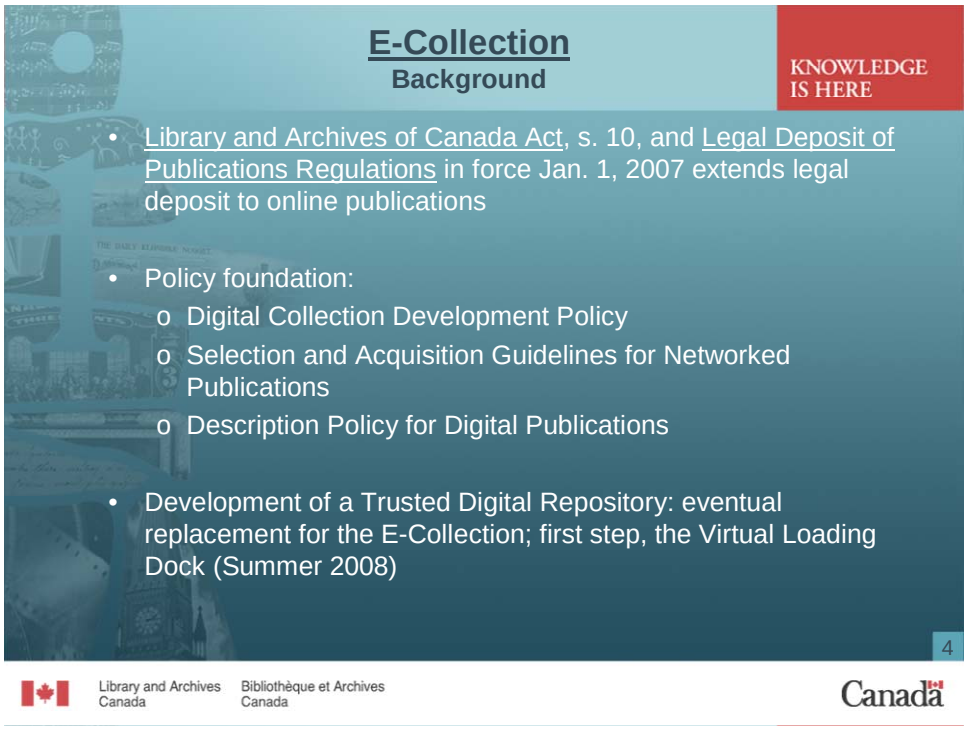
These government online publications were acquired and archived in the E-Collection using a variety of ingest methods. Staff negotiated this ingest and archiving of, and accessibility to Government of Canada online publications (eg. Reports, journals, gazettes, annual reports, and very experimentally, a few websites and blogs) on a title by title basis. Originally voluntary, this changed effective Jan. 1, 2007 when the existing legal deposit legislation was expanded to include e-publications including those of the Government of Canada. However, their acquisition was and remains essentially a title by title negotiated process.

In the E-Collection, all publications have been fully described similar to the treatment accorded conventionally produced publications. To access the digital publications in the E-Collection, users search LAC's web accessible online catalogue, Amicus. From there they are directed to the publication which is stored in the E-Collection. Whether the full content of a publication is accessible to the user is determined by the publisher. Access to each publication is negotiated with the publisher, even with government publishers. Although LAC tries to negotiate the best possible terms for its clients on a global basis (eg. One permission for all publications, for all time) as we have with a number of departments, this may not always be the case. Where the publisher deposits but wishes to limit public accessibility to the E-Collection copy, access is set to "restricted". This means that researchers wishing to read the publication can see it at specially designated terminals at LAC headquarters. There the publication can be read, but neither downloaded nor printed.

The system has admirably withstood the test of time and continues to witness a robust annual growth. As of Mar. 31, 2008, the publication count stood at nearly 30,000. The composition of the publications in this database consists of 68% federal government, 3% provincial government, and the remainder, 29% from commercial and non-commercial sources.

However, there are limits to the E-Collection's capacity. For example, with the existing staff complement of eleven, ingest of titles is capped at around 6000 new titles and 15000 serial issues added annually to the archive. Ironically, considering we are working with digital publications, much of the labour that is involved in organizing the collection in various directories and file structures, is still largely manual. And, above all, from a preservation stand point, it is not an ideal preservation environment or trusted digital repository, the attributes of which have been described elsewhere[4]. So, while it has withstood the test of time, it is about to be retired and will morph into LAC's Trusted Digital Repository on the latter's full implementation.

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



As mentionned earlier, all of the online publications archived in the E-Collection have been acquired voluntarily. Until Jan. 1, 2007 LAC's legal deposit legislation did not extend to online publications. However, with the advent of Canadian legal deposit legislation[5] to include online publications, conditions are changing. The legislative part of the puzzle is in place.

Policies such as the **Digital Collection Development Policy**[6], and **Selection and Acquisition Guidelines for Networked Publications**[7], a second crucial piece of the overall tool kit are in place as well. A third policy, **Description Policy for Digital Publications**, to be released later this year, is setting the stage for a different view of bibliographic access to e-resources. It takes into account descriptive measures to handle vast amounts of web data. Together, these round out the legislation and show the direction in which LAC intends to go.

LAC now awaits a more robust ingest capacity to start applying legal deposit systematically and seriously. This comes in the form of a **Trusted Digital Repository** (TDR) which is in the development stages. The first significant part of that system, the Virtual Loading Dock or ingest portion of the TDR is scheduled for release in Summer 2008.

Features of the Virtual Loading Dock, include:

[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)

- 4 ingest channel possibilities: a Web form, email, FTP, and CD-ROM or other storage devices sent by regular mail

- Metadata extraction

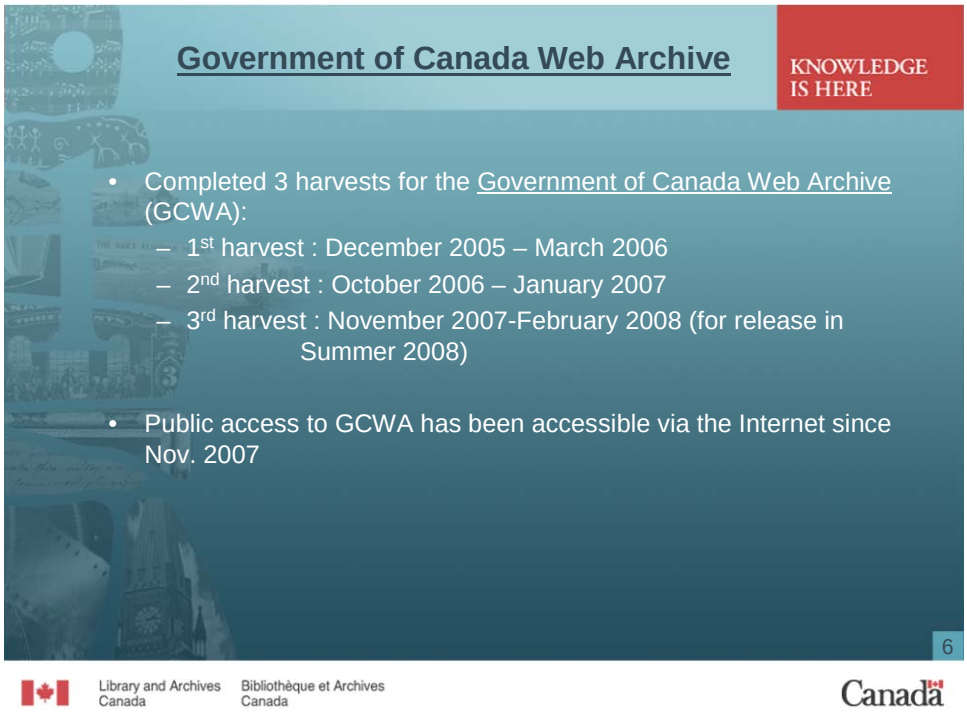
Once this ingest application is established and has been tested under full-load conditions, it is planned that other parts of the TDR will follow in the coming year:

- Development of the document management and public access portions of the TDR

- Migration of all publications from the E-Collection to TDR

- Possible extraction of e-publications from the Government of Canada Web Archive harvesting program into TDR

- Ingest Government of Canada departmental e-records



Now, let's turn to the second and more recent initiative, the **Government of Canada Web Archive**[8].

What exactly is the **Government of Canada Web Archive**? As the name implies, it's an archive composed of copies of websites of the departments, agencies, commissions and the like that belong to the federal Government of Canada. Approximately twice yearly, LAC launches a "crawl" of all of these websites. To date LAC has run three crawls. The crawler takes a "snapshot" of the entire publicly accessible contents of a website and returns it to the archive. As each site is live at the time of the crawl, the copy is a faithful rendition of the content, the look and feel of the site, as well as the links both internal and external, on the day, hour or minute the crawler makes a copy of it. As you might imagine, over time, a user can access successive generations of websites of a department.

LAC began experimenting with emerging web harvesting technologies in 2005. Early attempts at website harvesting seemed fairly positive indicators that this was going to be an extremely efficient method for gathering lots of web content in very little time. Having resolved most difficulties in the experimental stages, LAC began a first pass at the Government of Canada web domain in late 2005. A fairly tightly contained corps of websites, and relatively easy (or so we thought at the time!) to identify because of its characteristic address termination - .gc.ca – LAC finished its first harvest ever a few weeks later.

Although how many layers of a website to encompass in this snapshot can be programmed into the crawler, in LAC's case we instructed it to go as deep as possible. The result: a full and comprehensive snapshot of the publicly accessible web presence of each government department's website. It should, however, be noted that the crawler cannot gather all web content. For instance, the crawler will not penetrate a firewall to access content found on a department's Intranet. It is also stopped by registration points requiring user input (eg. A search screen to a database of reports; web information for which users must pay and which requires registering evidence of payment).

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

As the slide above shows, the Government of Canada Web Archive has three basic indexes: full-text search, departmental name, and URL. Anecdotal evidence suggests that users are quite satisfied at this point with these indexes, basic though they may be, to find and locate a wealth of information that they never had access to before.

The next slide illustrates another key feature of the Archive, that of differentiating archived web content from the web content on the department's current website. Because segregating archival from current web content is a major issue for users, not just in this Archive but also in other archival situations and search engines in general, some years ago LAC developed a banner by which to distinguish an archived page from current content. Without this bright green banner announcing the precise time the page was crawled, issuing some other caveats (eg. That links to external websites may not function), and offering other options (eg. Viewing earlier/later versions of the same page), a user may easily confuse live with dated web content.

In comparing LAC-initiated web harvesting to the deposit of individual titles approach reflected in the E-Collection where the onus to deposit falls on the publisher, some interesting observations can be made. Our experience has made us consider web harvesting as an important tool for inclusion within the legal deposit toolkit, as a way of optimizing the amount of government information captured, and reducing the overall human resources for both publisher and archive.

In a comparison of the two initiatives, the reader is easily drawn to noticing their symbiotic relationship. Whereas the Government of Canada Web Archive crawls can ingest enormous amounts of web content there are some areas it cannot reach (eg. Information in databases; information for which the user must pay to access). In such a case, Legal Deposit can be brought to bear to see that the “locked down” information is deposited and made accessible at an appropriate time. Similarly, because e-publications, the same that government publishers ought to be depositing on a title by title basis, are embedded in the web content the crawlers pick up with each pass they make, LAC has found that the harvest of e-publications alone by far outstrips the deposit of individual titles. As well, a number of government publishers have requested LAC to harvest exclusively as doing that removes the resource pressures from them to ensure each new publication is deposited.

However, it should be also noted that there are drawbacks involved which it is possible to imagine can be resolved with more system development. For example, deposit in the E-Collection also ensures e-publications that are collected are

organized in tried and true traditional library ways (eg. Each report is individually catalogued; serial issues are all arranged chronologically in one place), whereas in the Government of Canada Web Archive, the user has to search across all the archived websites for instances of the report in question. Reports are not isolated easily, they may be duplicated, and in the case of serial issues are not stacked in a sequential, easy to use, easy to find way. Yet, this is more than outweighed by the fact that so much web content has been brought together in one place and is accessible immediately to the user who is already familiar with website navigation. Simply put, users are grateful to find so much government web content at their fingertips in the Government of Canada Web Archive.

Though these illustrate a number of points where both approaches differ, it's obvious that when the powers of both are harnessed together they constitute an even greater advance in ensuring the ultimate accessibility of government information.

Government of Canada Web Archive : background

The legal foundation for harvesting Canadian websites in the interest of posterity lays in the relatively recent **Library and Archives of Canada Act** [9] that came into force in Apr. 2004. Under Section 8(2) of the Act, Library and Archives Canada is permitted

"… for the purpose of preservation, the Librarian and Archivist may take, at the times and in the manner that he or she considers appropriate, a representative sample of the documentary material of interest to Canada that is accessible to the public without restriction through the Internet or any similar medium."

[9] Ibid. s. 8(2)

As discussed previously, Legal Deposit and harvesting are two separate provisions in the Act.

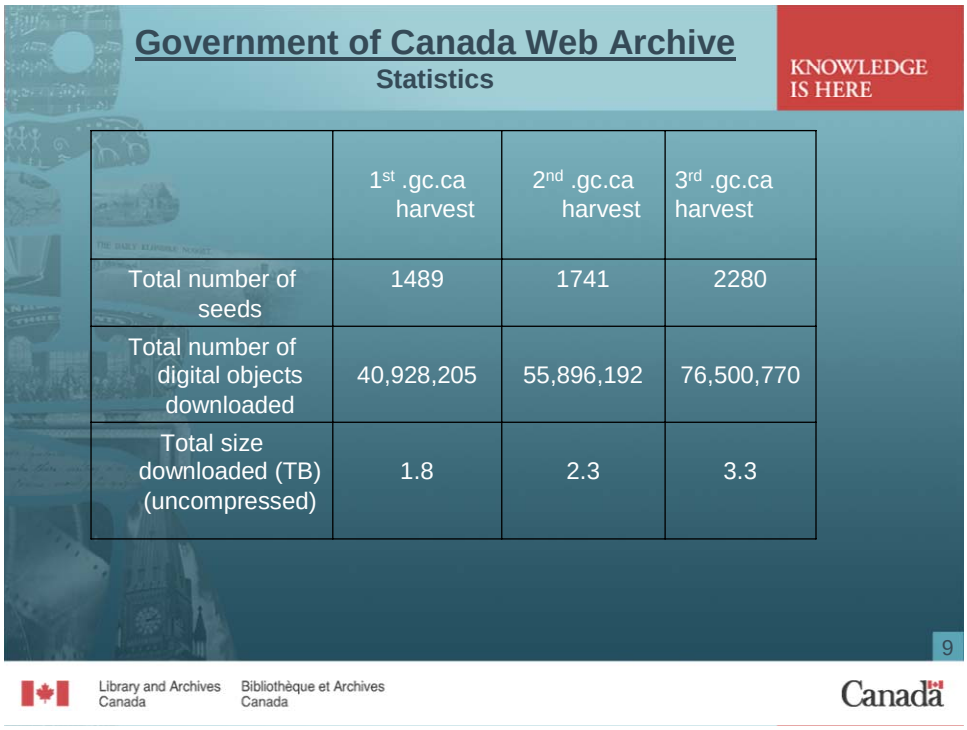
Given the state of infancy of harvesting knowledge in Canada at the time, were the Act to be reworked today, it would likely be reworked in such a way that harvesting would become part and parcel of legal deposit as well. In a way the legal deposit legislation is a trifle anachronistic in that it conceives of an online "publication" much in the same way as publications published in a conventional way. When this is also taken in conjunction with the benefits of a mass ingest methodology and technology of web content that harvesting represents, it is enticing to hypothesize how legal deposit legislation may be shaped in the future. Could we be looking at a time when deposit is a combination of LAC harvested web content mixed in with a more traditional approach where the onus is on publishers to deposit?

Policy is always required to complete the entire picture. In this case, LAC has developed two key policy documents, the **Digital Collection Development Policy** [10](noted earlier) and the **Selection and Acquisition Guidelines for Canadian Web Sites**[11] which are worth studying. To this we will soon be adding a web harvesting strategy which will round out LAC's direction.

As an aside LAC has also undertaken the harvesting and archiving of provincial and territorial government websites. These particular crawls are only done on an annual basis. The archived websites are "deep archived", thus not publicly or even internally accessible until LAC can broker agreements with the respective governments or find ways to collaborate in making them publicly accessible.

Finally, it's worth noting that LAC has also added to its harvesting experimentation with selective harvesting of smaller domains, special events, and assorted other web developments (eg. Blogs, FaceBook, Wikipedia, YouTube). The latter especially are cutting edge archiving experiments that are appearing even in government. Again, how these are to be effectively captured, presented for internal access or even, with negotiation, rendered accessible to the Internet is still under discussion.

[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)



Web Archiving Statistics

Although the statistics seem fairly self-explanatory, they deserve a few side comments.

Total number of seeds (Internet addresses)

As you can see, each successive crawl shows a growth pattern. This speaks to the fact that even in a relatively "finite" universe of seeds/urls, as time moved on we kept discovering new government websites that we hadn't know about before. This was partially because both the general public as well as authoring departments keen on ensuring that their entire site, as they knew it, should be represented, alerted us to missing related sites. Needless to say, in any organization that is as big as government, in an environment where web site management is pretty "wild-west", departments don't necessarily always act in a concerted way. Departments may also lose track of who in their departments create sites. Websites seem seldom managed centrally.

Total number of digital objects

This figure too has grown over time. It's growth attests not only to the increased effectiveness of our crawls with experience, but also to the absolute wealth of web content that LAC picks up in these crawls. Although duplication of web content from one crawl to the next exists, there is still a significant net growth of web content that is occurring. A testimony perhaps to the burgeoning use that governments are making of the Internet as a method of communicating information to citizens.

Also to be noted are the potential publications embedded in this web content. As you know, a government web site is composed of general information as well as

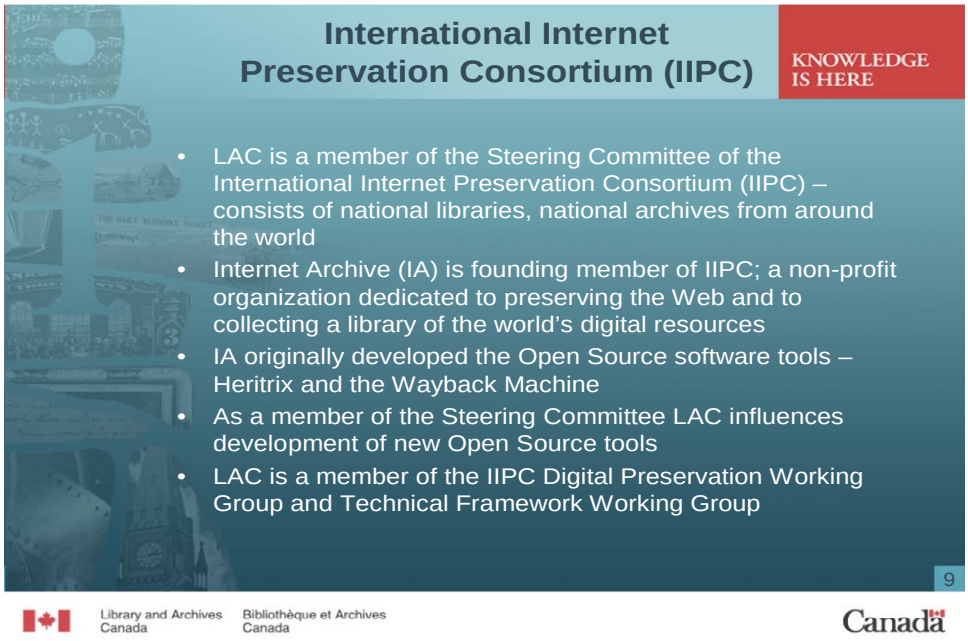
publications embedded in the site. While we have no precise figures on how many publications exist in each crawl, experimentation on some sites has led us to discover that from a "publication" perspective there are literally thousands of them embedded in a website, far more than are brought to our attention by the title by title deposit required by our legal deposit legislation. Further on I will say a few words about the potential that we see in using harvesting as another legal deposit ingest technique.

Total Size

We feel that the total size of 3.3 TB (uncompressed) of federal government web content is fairly staggering for a moderately sized country like Canada. As you can see, there continues to be a significant growth with each crawl. Where it stops we cannot and dare not project at this time.

In this coming year, we are looking at installing other software – the Smart Crawler – which may bring the size down with each crawl. As mentionned before, duplication of web content occurs from crawl to crawl. The software we are contemplating will, during a crawl, compare the current website with the website of the last crawl; unless a difference is noted, the crawler will not copy the current website. If something has changed on the website, users will still see the entire website the new content simply spliced in with the older and unchanged content captured by the previous crawl.

We believe that this will have the effect of reducing the actual size of what is archived for each crawl, probably increase performance, and as a by product also give us statistics on the rate of change of web content for each site. On analysis, the latter may lead us to decide to crawl websites that change frequently more often than those that hardly change at all.



International Internet Preservation Consortium (IIPC)

LAC started on this road of archiving government web content very much inspired by developments internationally that it was learning about through its membership in the **International Internet Preservation Consortium** (IIPC). Without the support, technical and information sharing, LAC would not be at the point at which it is today.

The cost involves developing certain areas such as standards on access as well as contributing software developments back to the IIPC membership. The cost has been well worth it.

Challenges and Points to Ponder

While engaging in this process of harvesting web content from government web sites for well onto two years now, we have certainly learned an awful lot. As well, though, we have all sorts of questions to which I'm certain we'll find answers as we work through the issues involved.

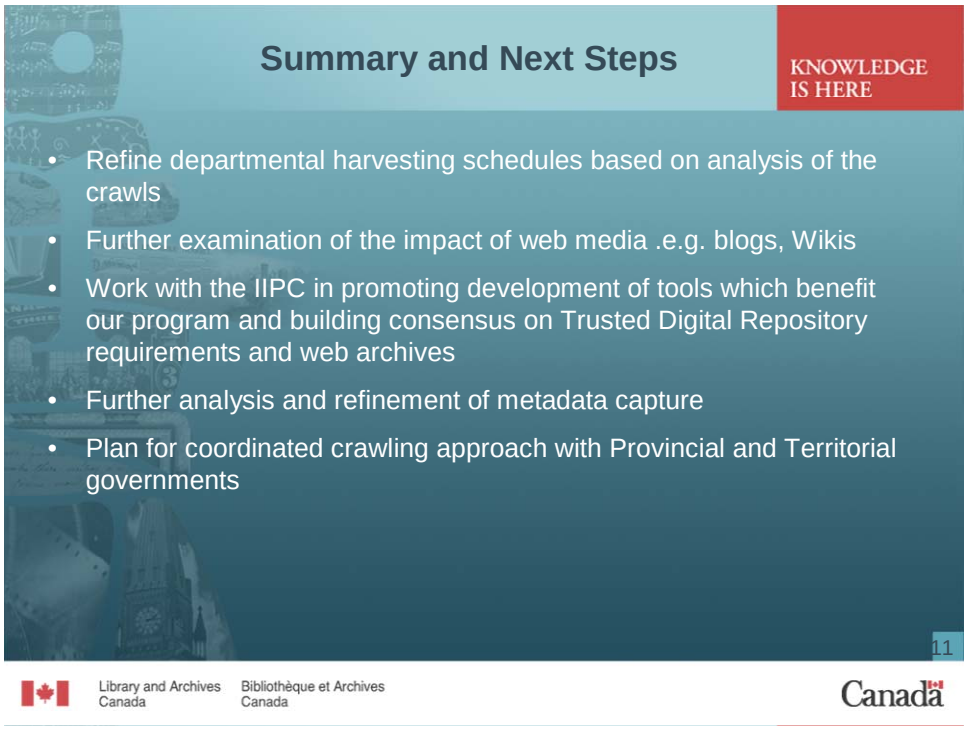
Technical challenges will always be present. Some such as elimination of duplicated web content (ie. Each crawl duplicates much of the previous crawls) will be eventually solved by the installation of a software solution created by one of the IIPC members. Finding clever ways to save computer space improves the speed at which the user can locate information in the archive. Other challenges with similar results are more internal in nature (eg. Available bandwidth, allocation of computer time). And, as users are not the only ones to have a vested interest in the contents of the archive, the creator departments also would like to see some changes made. One such department has asked if there is a way to distinguish in a Google search

between the a result set taken from the current version of the website and those versions LAC have archived. It is important to the department that today's user gets access to today's web content, and not that of the past. An important consideration. To date the contents of the Government of Canada Web Archive are not googled although the site home page per se is. Another consideration to users and to some extent by authoring departments, is whether users can get access to data kept in databases inside the website. It is a fairly commonplace occurrence to see that many departments install databases on their websites consisting of publications, reports, data on programs, etc. As the robot doing the crawls stops at the search screen heading a database, a registration point of sorts (there are many but each is characterized by the fact that a user has to enter search terms or register their presence), a lot of government information embedded in these databases is simply not archivable and thus remains inaccessible. There is much discussion internationally how feasible it is to penetrate databases; however, in late Fall 2007, when Google announced that it had found a way to get access to and index the content of these databases, it opened up the possibility that such functionality could be adapted to web harvesting.

As mentioned earlier, LAC has discovered the web harvesting of government web content metes us more than a surfeit of e- publications that we theoretically ought to have received deposited by authoring departments on a title by title basis. It is likely that even with the best of intentions harvesting web content outdoes individual publication deposit. In addition you have the context to publications plus all the surrounding information not classed as a publication. Is it possible then to ask if it's conceivable that future versions of legal deposit regulations will integrate the traditional requirement to deposit where the onus is on the publisher to deposit with harvesting where the initiative to acquire falls on the legal deposit agency itself thus removing the onus from the publisher to deposit?

There is also an intriguing relationship between harvests of public web content that the government creates and the records of behind-the-scenes business transactions that are the complement to public information. Like publications, records that were traditionally in paper format, are now changing to electronic. LAC plans on migrating the first batch of e-records this year into the same Trusted Digital Repository (albeit in a separate section) in which e-publications already appear. The presence of the two e-environments (ie. Government publications and records) in a single system will over time provide users with a fully integrated information package.

Finally, LAC is wondering about the use of the GCWA. Who actually uses this data, what is a typical user profile, will patterns of use change over time, will public use have a chain effect on the way governments will publish and package information to the public? It is too early to tell at this point but nevertheless remains an absorbing thought to explore. Perhaps at the next IFLA conference we will have enough evidence to report on user trends.



Summary and Next Steps

As I noted in the opening statements of this paper, "rendering access to government information at any level of government in any jurisdiction is indeed a 21st Century global challenge." What Library and Archives Canada has done in the creation of the two initiatives described in this paper is to have taken a simple and fairly rudimentary step toward meeting the goal of acquiring, archiving and rendering accessible government information of the Canadian federal government. I emphasize this as a "step" and not a solution for the simple reason that rendering access to government information will always remain an ongoing challenge however a clever solution we find to managing it. Governments continue to pump out an awe inspiring amount of information, it changes quickly, and in a digital environment can and will be expressed in so many different ways that we cannot yet imagine. No, this is simply the first step.

The points on this slide attest to the challenges that lay ahead of us in the near future. Through analysis of the web crawls, we will be able to better tailor the crawls so as to take more frequent snapshots of some government websites, whereas others that change less frequently we reduce the amount of crawling done. Will we be able to rise to the challenge of capturing new media expressions (e.g. blogs, Wikis) that governments are beginning to use, or even of those non-government media sources that comment on what the government does so that in future researchers may see a fully contextualized picture of how citizens felt about what they were reading.

LAC will definitely continue collaborating with the international community of web harvesting initiatives. It is this path that will eventually lead us closer to our goal of

"globalization of government information" – is it too dream-like to imagine a universal standard access to and long term preservation of government information? Will future legal deposit legislation morph quickly enough to accommodate new technologies?

Further analysis and refinement of metadata capture remains high on our list of priorities as well. In this regard every improvement in automated metadata extraction reduces the resource intensity of description, and aids the user in finding what they are looking for. The current indexes in the Government of Canada Web Archive work well, but could be considered to be rudimentary. Any additional way of enhancing those indexes will benefit users.

And why stop here? In Canada we are certainly pushing toward a more universal capture not only of government information at the federal level, but also a more systematic harvesting of government information from provincial and territorial governments. Next goals might be discovering collaborative arrangements with each of these jurisdictions to make the contents publicly accessible. Similarly, we are studying the possibility of harvesting the entire country's web presence.

In short, the two initiatives I have described here are indeed but a first tiny step toward the goal of this seminar: rendering government information accessible to the user. It is an evolving story, with basically no end in sight. Government publishers will like any other publishers in the 21st Century continue to evolve, and we will for ever follow in their wake, ever evolving ourselves to meet these new challenges.

Thank you.

Biographical information of author/presenters: Gillian Cantello, Director General, Published Heritage Branch, Library and Archives Canada has been involved with LAC since 2002. She has served in a variety of capacities including programs ensuring access to government information.

- or -

John Stegenga, has been associated in one form or another with government publications as well as with Legal Deposit in LAC since 1974. He currently works in the Published Heritage Branch.