IQ Looks into the NZ-Led WEB CURATOR TOOL

Late last year, the International Internet Preservation Consortium (IIPC) , a coalition of 12 major archives and libraries in the US, Canada, the UK, Australia, New Zealand, France and other European countries based at the French National Library in Paris, announced that an IIPC group led by the National Library of New Zealand (NLNZ) had successfully created and initiated the Web Curator Tool (WCT). What is it? How does it work? *IQ* Contributing Editor Mike Steemson put our questions to NZNL's project leader STEVE KNIGHT

***IQ:* Steve, why and when was the IIPC formed?**

SK: The IIPC was formed in July, 2003 to find a way of 'preserving Internet content for future generations.' These founding charter nations have a combined population of almost 600 million .

***IQ:* Three of the IIPC partners, led by the National Library of New Zealand, developed the Web Creator Tool?**

SK: That's right. The NLNZ's supporting team comprised the British Library (BL), one of the world's oldest and largest libraries, and Australian TelstraClear's IT&T services subsidiary, Sytec.

***IQ:* How was it that this international project was led by New Zealand?**

SK: It was fairly serendipitous, really. In June 2005, I attended the IIPC meeting in Reykjavik, Iceland, as an observer. The National Library of New Zealand was not a member of the IIPC at that stage. Following a recent revamp, we are now on the Steering Committee. But we had been undertaking research in Web archiving for some time.

While there had already been some consideration amongst consortium members of the need for a desktop tool for the

managing of Web archiving, this began to be firmed during discussions at the Reykjavik meeting.

An agreement in principle to collaborate on the WCT was reached at Reykjavik and this was reaffirmed at the next IIPC meeting held in Washington in October, 2005.

***IQ:* What part did you play in the NLNZ being appointed to the lead role?**

SK: Following revision of its National Library Act in 2003, New Zealand became one of the first countries in the world to have a legislative mandate to collect digital materials, including Web sites. At this time The British Library already had in place voluntary deposit mechanisms for digital material as it worked through the legislative process to support legal deposit.

Because the NLNZ already had the legal deposit provisions in place, we were ready to go on the development of a desktop tool to support the Web harvesting process. It was something we were going to do anyway and we were ready to get started, so it seemed sensible at the time that NLNZ should take the lead in the project.

In my role of Manager Innovation Centre, it is my responsibility to be aware of emerging trends in the environment that might impact on NLNZ's activities in the digital space. We had already been doing some research using the HTTrack Web crawler, and it was very clear this was an opportunity for the Library to further its own needs in this space while contributing to the wider web archiving community.

We are very aware of the need to act as good and global citizens in our work on digital preservation, of which Web archiving is a component. It is still a very nascent activity for most of us, so the opportunity to lead the work on the WCT was attractive for a number of reasons.

***IQ:* What do you, the NLNZ and IIPC hope to get out of the WCT? How will it be better than what is currently available?**

SK: It's not really a question of how the WCT will be better than what is currently available. Web archiving is a very new activity and to a large extent all the communities involved in it are learning from each other how to optimise the range of activities involved.

The National Library of Australia (selective Web harvesting), the Royal Library in Sweden (harvesting whole-of-domain, for example), the Internet Archive in the United States (preserving the whole Internet, development of the Heritrix crawler) have been key contributors to a growing body of knowledge related to Web harvesting.

We hope that the WCT will become another piece of the puzzle in the development of end-to-end processes … a single work flow encompassing selection, permissions, harvesting, etc, right through to an appropriately controlled public interface to archived Web sites.

***IQ:* You and your BL counterpart, Philip Beresford, posed and answered the basic questions about the WCT in your presentation to the sixth International Web Archiving Workshop in Alicante, Spain, in September last year, when IIPC launched the programme. We'll throw those same question back at you. To begin with, what is the Web Curator Tool and what does it do?**

SK: The WCT is designed as a desktop application for managing a selective Web harvesting process. For example, specific sites to be harvested on a one-off or periodic basis and for event or theme-based harvesting, like the Rugby World Cup, or an election.

It is not designed for whole-of-domain Web harvesting - the whole of the .nz domain, for example. Although the underlying harvester technology, Heritrix, will be used for that purpose.

The primary purpose of the WCT is to provide an environment which can be put on the librarian's or curator's desktop which will allow them to do their work without the need to engage with the underlying engineering.

The WCT supports:

- Harvest authorisation - permission to harvest Web material and make it available.
- Selection, scoping and scheduling - what will be harvested and how often.
- Description (Dublin Core metadata).
- Harvesting - downloading the material at the appointed time.
- Quality review - making sure the harvest worked as expected, and correcting simple harvest errors.
- Submitting the harvest results to a digital archive.

***IQ:* What is it NOT?**

SK: The WCT is not a digital archive. It is not appropriate for long-term storage. It's not an access tool. It does not provide public access to harvested material. It's not a cataloguing system, although it does provide some base information about individual harvested websites. And it's not a document management system; for example recording of permissions from publishers must occur elsewhere.

***IQ:* Where is the WCT being used? Who's doing it, and what's been collected?**

SK: The tool was only released into the open source community in September last year, so it is too soon yet to gauge how successful it will be.

It is already in production at National Library of New Zealand, and the US Library of Congress has indicated that it will be moving to the WCT as its production environment for Web harvesting.

In the UK, the United Kingdom Web Archiving Consortium (UKWAC), comprising The British Library, the National Library of Scotland, the National Library of Wales, the National Archives of the United Kingdom and the Wellcome Trust Library, expects to move from its current Web archiving platform to the WCT.

So, you can see, there has been a substantial level of interest shown in it in a short amount of time. We hope that this will

WEB ARCHIVING FROM YOUR DESKTOP?: Continued from page 41

accelerate as more institutions trial the software to determine its fit for their programmes.

***IQ:* Who decides what WCT will harvest?**

SK: At the NLNZ, selection decisions are made by the responsible curatorial area, the New Zealand and Pacific Published Collections. This is part of the original rationale for the development of the WCT, to mask the engineering component, so that it can be operated within the business.

Clearly, though, this will depend on individual institutions workflows for Web archiving, and it will be up to each institution to determine the appropriate place for the actual harvesting activity in their own workflows

***IQ:* What must a WCT operator arrange with the website authors and owners regarding such matters as copyright, ownership and the like?**

SK: The NLNZ is very aware of its obligations relating to copyright, usage restrictions, etc, and is careful to ensure that legislation and other restrictions are honoured.

It should also be remembered that Web archiving is part of NLNZ's wider digital preservation activity, designed to ensure that

The Web Curator Tool Technology

The WCT is:
Implemented in Java
Runs in Apache Tomcat
Platform:

- Tested on Solaris (version 9) and Red Hat Linux
- Developed on Windows
- Should work on any platform that supports Apache Tomcat

Database:

- A relational database is required
- Tested on Oracle and PostgreSQL
- Installation scripts provided for Oracle and PostgreSQL
- Should work with any database that Hibernate supports including MySQL, Microsoft SQL Server, and about 20 others

Incorporates parts or all of

- Acegi Security System
- Apache Axis (SOAP data transfer)
- Apache Commons Logging
- Heritrix (version 1.8)
- Hibernate (database connectivity)
- Quartz (scheduling)
- Spring Application Framework
- Wayback

New Zealand's digital memory is kept safe for future generations of researchers. We do not want to put that at risk by failing to comply with restrictions or other ownership rights.

This will obviously vary from jurisdiction to jurisdiction but even for members of the public who might want to use the WCT, it is imperative that they are aware of their local requirements for the use and/or re-use of published materials.

IQ: Who built the WCT?

SK: The NLNZ contracted a local software development house, Sytec Resources now a subsidiary of TelstraClear, to build the software. The library has worked with Sytec for several years and we were really pleased to work with them on this project.

It is testimony to the skill and desire of the three teams involved, BL, NLNZ and Sytec, that the project was able to be completed on time and within budget across time zones and with only one face-to-face meeting for design specification workshops held here in New Zealand.

***IQ:* How easy is it to operate? How friendly are the graphic user interfaces?**

SK: We are very happy with the interface that has been delivered, and while we expect it will not always fit the internal workflows of every institution, we are comfortable that it provides the base level of both functionality and usability for other institutions to build on for their own programmes.

***IQ:* What has it cost to develop, and who paid? What will it cost to run annually?**

SK: As already noted, the project was delivered on time and within budget. The total cost of the project was around US$400,000, shared equally between the National Library of New Zealand and the British Library. Both organisations also contributed in kind services related to project management, requirements and design specifications.

As an open source project there is no annual licence or maintenance fees. The Library has no experience in running an open source project and we are still working through what that will entail. Costs related to that will be shared between BL and NLNZ.

***IQ:* Where can other institutions get it, and how much would it cost them?**

SK: It's free. The WCT was released as an open source project. Its source code, documentation, including user and administrator guides and FAQs, and an associated mailing list are available free at http://webcurator.sf.net.

The WCT is released under Apache License, Version 2.0. This does not mean anything in practical terms for normal users of the WCT. It is not a licence that one needs to have or buy in order to use the WCT.

By specifying the nature of the open source licence under which the tool is released, we are basically alerting the software development community to the rules that they would need to follow if they want either to contribute to the ongoing development of the tool or to re-use components of the tool for other purposes – for example, crediting the original developers.

***IQ:* What's next for WCT and the NLNZ-BL-Sytec team?**

SK: At present there are two concurrent streams of work going on with the WCT. NLNZ is undertaking a review of requirements which were not able to be developed in 1.1. For example, the increasing the functionality around the Quality Review tools for checking harvested websites for completeness, accuracy, etc.

A 1.2 version is planned later this year. At the same time, the BL is working on the requirements for a public access tool in order to build an end-to-end process within the WCT.

***IQ:* You must feel considerable satisfaction at the parts you and the NLNZ have played in the WCT's development.**

SK: As I mentioned, it is important to see the WCT development as part of a series of activities over time and underway at the moment that will, one day, result in a clearly understood, standards-based process for web archiving.

The National Library of New Zealand is very pleased to have been able to contribute to this ongoing global endeavour, to ensure that the Web is available to the future.

***IQ:* Thank you, Steve. And good luck with this important leading edge project.**

MORE ABOUT THIS KNIGHT OF NEW ZEALAND

Steve Knight is Manager Innovation Centre and Programme Architect, National Digital Heritage Archive, Digital Innovation Services, National Library of New Zealand, at Wellington.

ENDNOTES

International Internet Preservation Consortium web at http://netpreserve.org.

New Zealand population: 4.1 million.

IIPC founders (populations in millions): USA 300, France 61, Britain 61, Italy 58, Canada 33, Australia 20, Sweden 9, Denmark 5, Finland 5, Norway 5, Iceland 0.3.

HTTRack web crawler: see http://www.httrack.com/

The Heritrix Crawler, see http://crawler.archive.org/

Dublin Core metadata, see http://dublincore.org/

Sytec Resources, see http://www.sytec.co.nz/