The Web Curator Tool Relaunch

The Next Generation of Web Crawling

Jeffrey van der Hoeven
National Library of the Netherlands
The Netherlands
Jeffrey.vanderhoeven@kb.nl
https://orcid.org/0000-0002-2534-8017

Ben O'Brien
National Library of New Zealand
New Zealand
Ben.O'Brien@dia.govt.nz
https://orcid.org/0000-0002-4290-2972

Abstract – This poster will highlight new features of the Web Curator Tool (WCT), added from January 2018 onwards through a collaboration between the National Library of New Zealand (NLNZ) and the National Library of the Netherlands (KB-NL). One of the themes from the collaboration has been to develop a modern fit-for-purpose WCT. This involves learning the lessons from previous developments, responding to recent trends in the web archiving community and completing a technical uplift. On this foundation a new, revamped WCT has been developed and released as version 2.x. As well as highlighting the latest developments the poster outlines the roadmap and community building planned for the WCT in the coming years.

I. Introduction & background

In 2006 the NLNZ and the British Library developed the WCT, a collaborative open-source software project conducted under the auspices of the IIPC. The WCT managed the web harvesting workflow, from selecting, scoping and scheduling crawls, through to harvesting, quality assurance and archiving to a preservation repository. The NLNZ has used the WCT for its selective web archiving programme since January 2007. However, the software had fallen into a period of neglect, with mounting technical debt: most notably its tight integration with an out-dated version of the Heritrix web crawler. While the WCT is still used day-to-day in various institutions such as the KB-NL, it had essentially reached its end-of-life as it has fallen further and further behind the requirements for harvesting the modern web. The community of users have echoed these sentiments over the last few years.

During 2016/17 the NLNZ conducted a review of the WCT and how it met their business requirements, and compared the WCT to alternative software/services. The NLNZ concluded that the WCT was still the closest solution to meeting its requirements - provided the necessary upgrades could be made to it, including upgrading to the Heritrix 3 web crawler. Serendipitously, the NLNZ discovered that another long-time WCT user, the KB-NL, was going through a similar review process and had reached the same conclusions. This led to collaborative development between the two institutions to uplift the WCT technically and functionally to be a fit for purpose tool within these institutions' respective web archiving programmes.

II. WCT next gen uncovered: version 2.x

The objective of the joint effort of NLNZ and KB-NL is to get the WCT to a platform where it can keep pace with the requirements of archiving the modern web. The first step in that process was to decouple the integration with the old Heritrix 1.x web crawler, and upgrade to the more modern Heritrix 3.x version. Improved ability to configure the crawling variables were also realised. Apart from the technical side, the documentation has been given a major uplift, including updated instructions on installing the new WCT and migrating from older versions to the latest 2.x. The new version of WCT was launched at the end of 2018 [1] and is available on Github[1] as open source.

[1] Web Curator Tool made available on Github, https://github.com/DIA-NZ/webcurator

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

III. Continuing the work, together

Both NLNZ and KB-NL are working to jointly improve the WCT even further and have drawn up a roadmap with further milestones to be delivered in 2019 and beyond. This includes better support for various ways of using the WCT in web archiving by adding predefined user journeys, better support for quality assurance and making WCT suitable for crawlers other than Heritrix. Virtualising WCT by containerizing it is also on the agenda.

With our effort in revamping the WCT we hope to encourage existing WCT-users to upgrade their install base to the latest version and inspire others to start using it and take part in a growing upcoming community dedicated to improving the way we archive the web for generations to come.

References

[1] K. Teszelszky, "Web Archiving Down Under: Relaunch of the Web Curator Tool at the IIPC conference, Wellington, New Zealand," 2018. https://www.kb.nl/en/news/2018/web-archiving-down-under-relaunch-of-the-web-curator-tool-at-the-iipc-conference-wellington-new

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.