Archiving the Web
Web archiving is the process of collecting parts of the Web to preserve so that it can be accessed by current and future users, especially after the content no longer exists on the Web. The collection is typically done by automated crawlers such as Heritrix 3, which is integrated into the Web Curator Tool (WCT).

The Web Curator Tool (WCT)
The WCT is designed for non-technical users who want to:
• selectivity capture websites or documents on the Web
• describe them to help users discover them
• prepare them for export into a preservation system

Most current web archiving activities rely heavily on the technical expertise of the harvest operators. The Web Curator Tool, on the other hand, makes harvesting the responsibility of users and subject experts (rather than engineers and system administrators) by handling automatically the technical details of web harvesting. The tool is designed to operate safely and effectively in an enterprise environment, where technical support staff can maintain it.

The WCT is free, open source software, available under the terms of the Apache License, Version 2.0.

Top 5 Reasons to Collect the Web
1. Much of our cultural heritage is only available online.
2. Content on the Web is at risk of being lost unless its captured and preserved.
3. We preserve books and works of art - this is just another form of human culture.
4. You may be mandated to collect it.
5. They can be used for research today and in the future.

How to Get Started with the WCT
Read the WCT documentation at webcuratortool.readthedocs.io including a Quick Start Guide and User Manual.

Try WCT 2.0
Follow along with a series of tutorials created for WCT 2.0. A Virtual Box image containing pre-installed instances of WCT, Heritrix 3 and OpenWayback is available for performing these tutorials.
webcuratortool.readthedocs.io/en/latest/guides/tutorials.html

Have a Question about WCT?
Ask it on our Slack channel: webcurator.slack.com
Workflow
The Web Curator Tool Relaunch

Administration of WCT
WCT provides extensive user management capability, through a flexible system of users, permissions, roles and agencies. Each user belongs to an agency, and has a number of roles that define the user’s access to WCT functionality.

Users
Each user has a WCT account, which includes basic identifying information and preferences. WCT can be configured to allow user authentication through an external Directory.

Roles
A role is a way of capturing a set of privileges and responsibilities that can be assigned to sets of Web Curator Tool Users. Each role has a set of privileges attached. Users who are assigned the role will be given permission to perform operations.

Agencies
An agency is an organisation involved in harvesting websites using WCT. Users and roles are defined for an agency scope. Targets, Groups and Harvest Authorisations are also owned at the Agency level, providing a convenient way of managing access to the tool for multiple organisations.
Workflow
The Web Curator Tool Relaunch

Harvest Authorisations
A Harvest Authorisation is a record of requested permission from copyright holders (e.g. website owner) to harvest a website. This supports the recording of the legal context, which may become important when preserving or providing access to the content.

Descriptive Metadata
Description fields are available for storing Dublin Core metadata in each Target harvest record. This is not used in WCT, but is included when any harvests are submitted to a digital archive.

Targets
The desired entity that is being harvested. Contains information relevant to the harvest including some metadata. A Target could be:
- a single document
- a part of a website
- a website distributed across several different hosts
- a collection of related websites
- an HTML serial issue located on a website
- any combination of these

Crawl Profiles
The WCT profile contains settings that control how a harvest behaves. The settings for WCT profiles are based on Heritrix profiles. Profiles can be created to crawl particular kinds of websites, such as blogs.
Scheduling Crawls
Crawls are based on configured schedules. They can be automatically or manually assigned to Harvest Agents, which perform the crawling.

Controlling a crawl
The WCT performs a crawl using the Heritrix 3 web crawler. The crawl can be started, paused, stopped and aborted from within WCT. Also a Heritrix 3 Console can be launched to execute scripts against running crawls.

Monitoring Crawls
Monitoring crawls is a core activity in WCT, and is essential to obtain successful web harvests.

The performance and health of a crawl can be assessed through real-time statistics, logs and reports. At a higher level the Harvest Agents which run the crawls are also monitorable.
Workflow
The Web Curator Tool Relaunch

Harvest Quality Assurance
WCT provides tools and techniques to support quality review of the harvest results. Reviewing a harvest is a key step in determining whether a crawl was complete and successful. Heritrix crawl logs/reports can also be used for quality review, and troubleshooting problematic harvests.

Harvest Patching
WCT provides a tree-like view of the harvest as a visualisation tool, but it can be used to delete unwanted material from the harvest or add new material. Patching the harvest enables minor problems to be resolved without restarting an entire crawl.

Browse Harvest Results
To see a simulated experience of the captured version of the online material, browse the harvest results using OpenWayback.

A Harvest History
As more instances of a website are crawled, a history of previous harvest statistics are stored for comparative analysis.
Once a harvest has been quality reviewed, if the harvest was successful, it can be endorsed. If it was unsuccessful, it can be rejected. When rejected, a preconfigured reason is linked to that harvest, denoting why it was unsuccessful.

If endorsed, a harvest is ready to be archived. As WCT is not a digital archive or document repository, and not appropriate for long-term storage, it supports the submission to an external archive or preservation system.

Upon submitting a harvest to an external archive, WCT will build a Submission Information Package (SIP) containing the harvest files. This includes:

- WARC files
- CDX files
- Heritrix logs and reports
- METS.xml

WCT constructs a METS file during the archiving process, using system and user-generated data. The METS file includes sections for Dublin Core and DNX metadata.

The source of this metadata can be tailored for specialised web harvests using WCT plugins. For example the harvesting of online serial publications may require additional metadata from the user upon archiving.
About WCT
• The Web Curator Tool (WCT) is a workflow management application for selective web archiving.
• The tool is open-source software and is freely available for the benefit of the international web archiving community.

The Early Years (2006 - 2012)
• It was first developed as a collaborative project between the National Library of New Zealand (NLNZ) and the British Library (BL), conducted under the auspices of the IIPC.
• The WCT has been built with support and contributions from professionals at the NLNZ, the BL, the National Library of Australia, the Library of Congress, and many others.

NLNZ & KBNL Working Together (2017 -)
• National Library of New Zealand (NLNZ) and the National Library of the Netherlands (KBNL) have a long history in preserving digital collections, including websites.
• After extensive research in 2017 we determined that WCT is still the best solution matching our organizational needs.
• Since 2018 we have worked closely on improving WCT, updating documentation and offering online support for migrating organizations to the newest version of WCT.
• In November 2018, NLNZ and KBNL presented a tutorial on WCT at the IIPC Web Archiving Conference in Wellington, introducing version 2.
• In June 2019, NLNZ and KBNL held a WCT workshop at the IIPC Web Archiving Conference in Zagreb. This included a hands-on session for getting started with WCT, with accompanying learning materials and environment.

Timeline
• 2006 - WCT 1.0: first release
• 2008 - WCT 1.4: feature & performance improvements
• 2009 - WCT 1.5: Wayback review tool integration
• 2012 - WCT 1.6: workflow & auto QA features
• 2017 - WCT 1.7: Start of Heritrix 3 integration
• 2018 - WCT 2.0: major upgrade
• What’s Next?
Overview

The WCT has a flexible architecture, allowing the components to be distributed over multiple servers.

WCT Core
- An access-controlled web interface where users control the tool. Management of selected websites, and other settings, and maintains a harvest queue of scheduled harvests.

WCT Harvest Agent
- WCT Core assigns a harvest to an associated Harvest Agent when it is ready to start. The Agent acts as a proxy between the Core and the crawl engine. A WCT installation can have multiple Harvest Agents configured.

WCT Digital Asset Store
- Completed harvests are stored on the digital asset store, where a set of quality review tools allow users to assess the harvest results. Successful harvests can then be submitted to a digital archive for long-term preservation.

Technologies
- JSP + Spring + Hibernate + Quartz
- Java 8 + Maven
- Tomcat / Jetty
- Mysql / PostgreSQL / Oracle / HSQL
- Linux / Solaris / Windows

OpenWayback Integration
An instance of OpenWayback can be used as a review tool inside WCT. The Digital Asset Store component contains configuration for this integration, that allows completed harvests to be indexed and viewable through OpenWayback.

Repository Integration
The WCT can archive harvests to the Rosetta Digital Preservation System. A plugin for the Digital Asset Store component contains configuration for this integration, and custom deposit forms are available for specialised harvests.
What's Next?
The Web Curator Tool Relaunch

Technology Upgrade Plan

**Current Technology** (version 2.0.x)
- Support for Heritrix 3
- Simplified consolidated install
- Better documentation

**Improved Technology** (upcoming version 3.x)
- Modern software components
- Platform for future changes, including crawler abstraction and component APIs/REST APIs
- Cloud-ready

Roadmap 2019-2020

- Technical uplift of internal WCT structure (2019)
- Crawler abstraction bringing support for other web crawler engines
- Improved user interface & functional uplift
- Cloud-ready & REST API
- Quality assurance improvements

Why is Technical Uplift Important?
We are modernizing the underlying technologies so that we have:
- a stable foundation for adding on new functionality
- supported technologies
- a larger pool of potential maintainers
- an easier pathway to integrating 3rd party applications and services, including a REST API

How to Get Involved?
We are striving to grow a larger community of developers and users. If you are interested in contributing, please let us know at webcurator.slack.com