



Web Curator Tool renewed

OPF Conference 2020 (online)

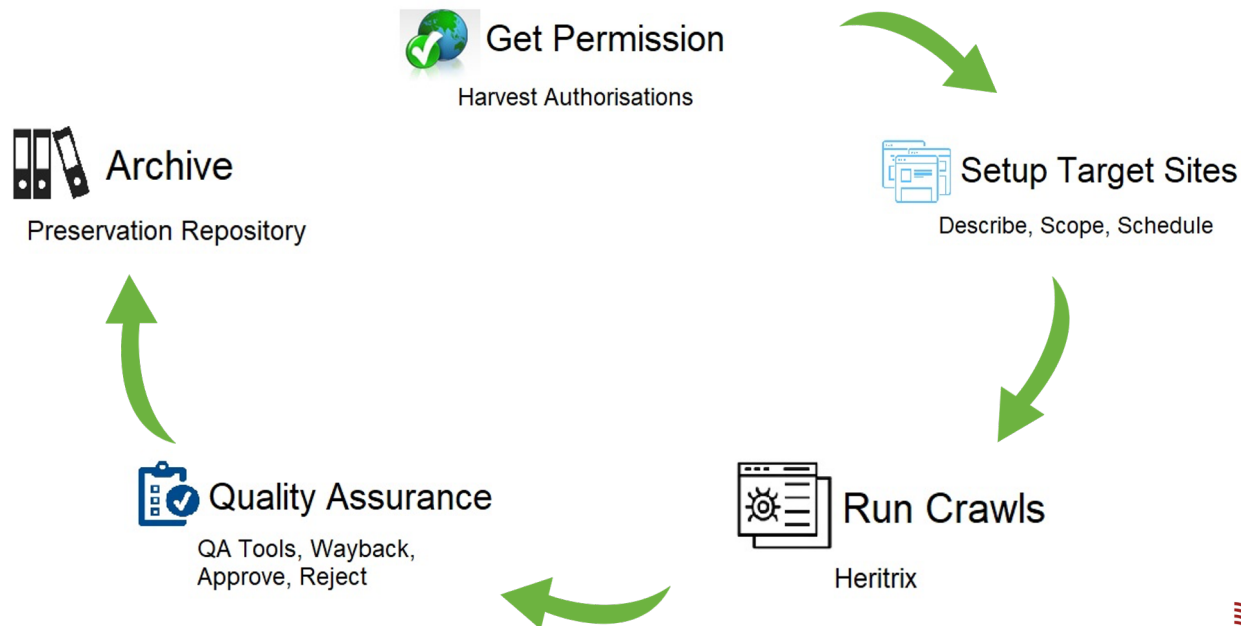
Lightning talk by Jeffrey van der Hoeven

National Library of New Zealand & National Library of the Netherlands

10 June, 2020

What is the Web Curator Tool (WCT)?

An open source workflow management tool for selective web harvesting. It supports selecting, crawling websites, performing QA and preparing websites for ingest to archival storage.



Highlights:

- Supports Heritrix 1 & 3
- Modular & extendable
- GUI-based
- Doesn't require deep technical knowledge to operate...
- ... but can be configured completely tailor-made for the job

About us



- Archiving the Web since 1999
- Selective Web archiving approx. 35,000 web instances in Rosetta
- 8 whole-of-domain domain crawls since 2008
- Legal deposit legislation since 2003

KB } national library
of the netherlands

- Selective Web crawling since 2007
- 18,000 sites as of Q2 2020
- No legal deposit, but..
- Preparing domain crawl of Dutch domain

Recent & upcoming upgrades

Version 2 (released end of 2018)

- Complete [Heritrix 3](#) integration
- Updated [documentation](#) and improved [installation](#) process

Version 3 (currently in test; to be released end of June 2020)

- Technical Uplift including up-to-date libraries and deployment

Version 4 -> in development

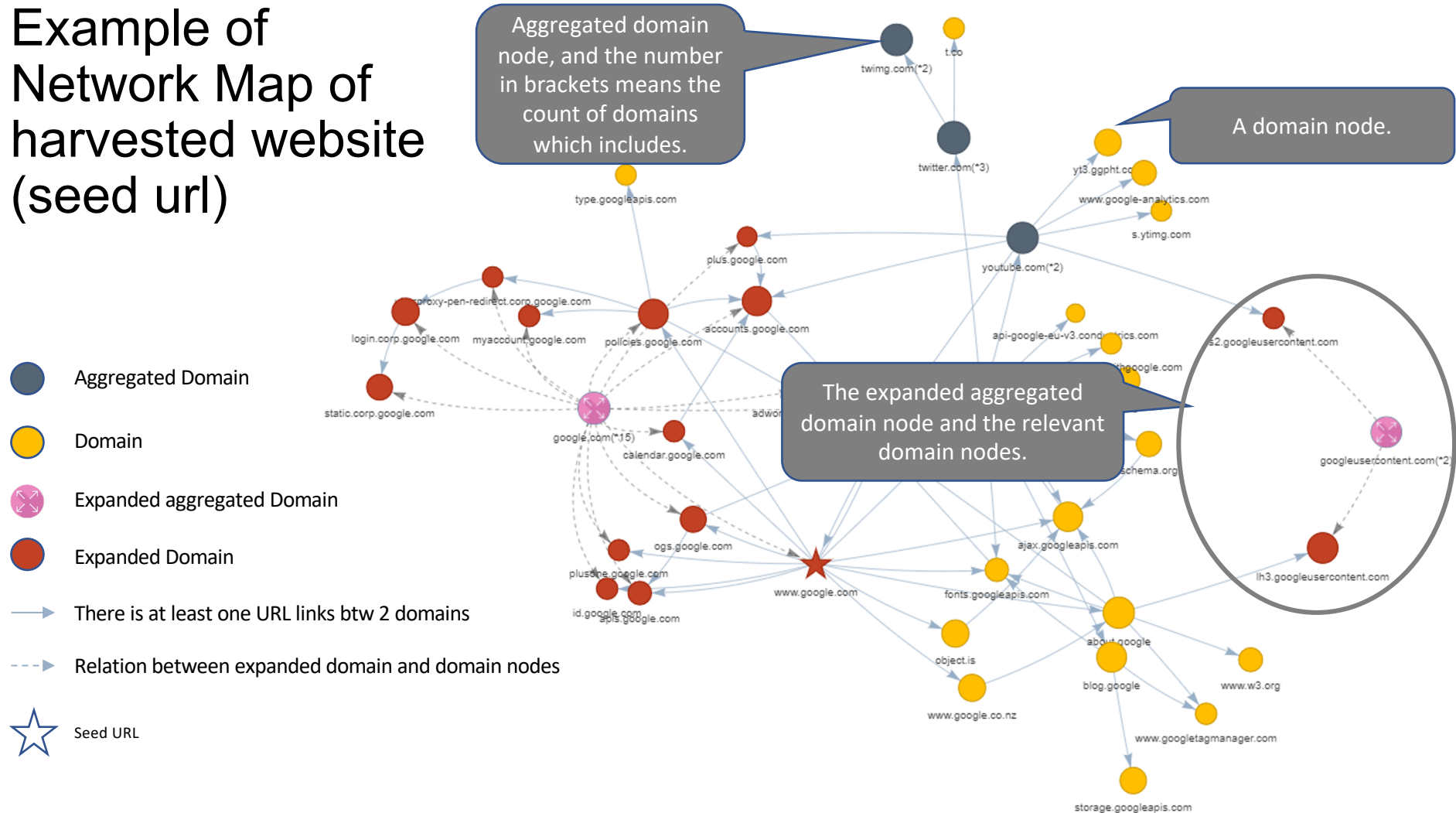
- Functional uplifts -> see next slide
 - Better support for Quality analysis
 - Integration webrecorder for patching

WCT version 4: focus on quality analysis

Bringing improvements to three core areas:

1. **Crawl patching using Webrecorder** -> this will add the ability to repair missing content in addition to the existing WCT import and prune functionality. For this, we offer visualisation of harvest results for inspection and harvest modification
2. **Screenshot generation** -> capture screenshots of live websites being crawled and the resulting web harvest for comparison
3. **Integration with Pywb viewer** -> the best options available for web harvest replay and review

Example of Network Map of harvested website (seed url)



THANK YOU

Feel free to contact us:

Slack - webcurator.slack.com/

Github - <http://webcuratortool.org/>

On behalf of the team:

Ben O'Brien

Hanna Koppelaar

Charmaine Fajardo

Frank Lee

Trienka Rohrbach

Andrea Goethals

Steve Knight

Jeffrey van der Hoeven