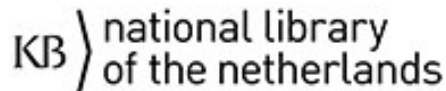


NOW EVEN BETTER

# Technical uplift of the Web Curator Tool

Ben O'Brien, Jeffrey van der Hoeven  
IIPC WAC 2019



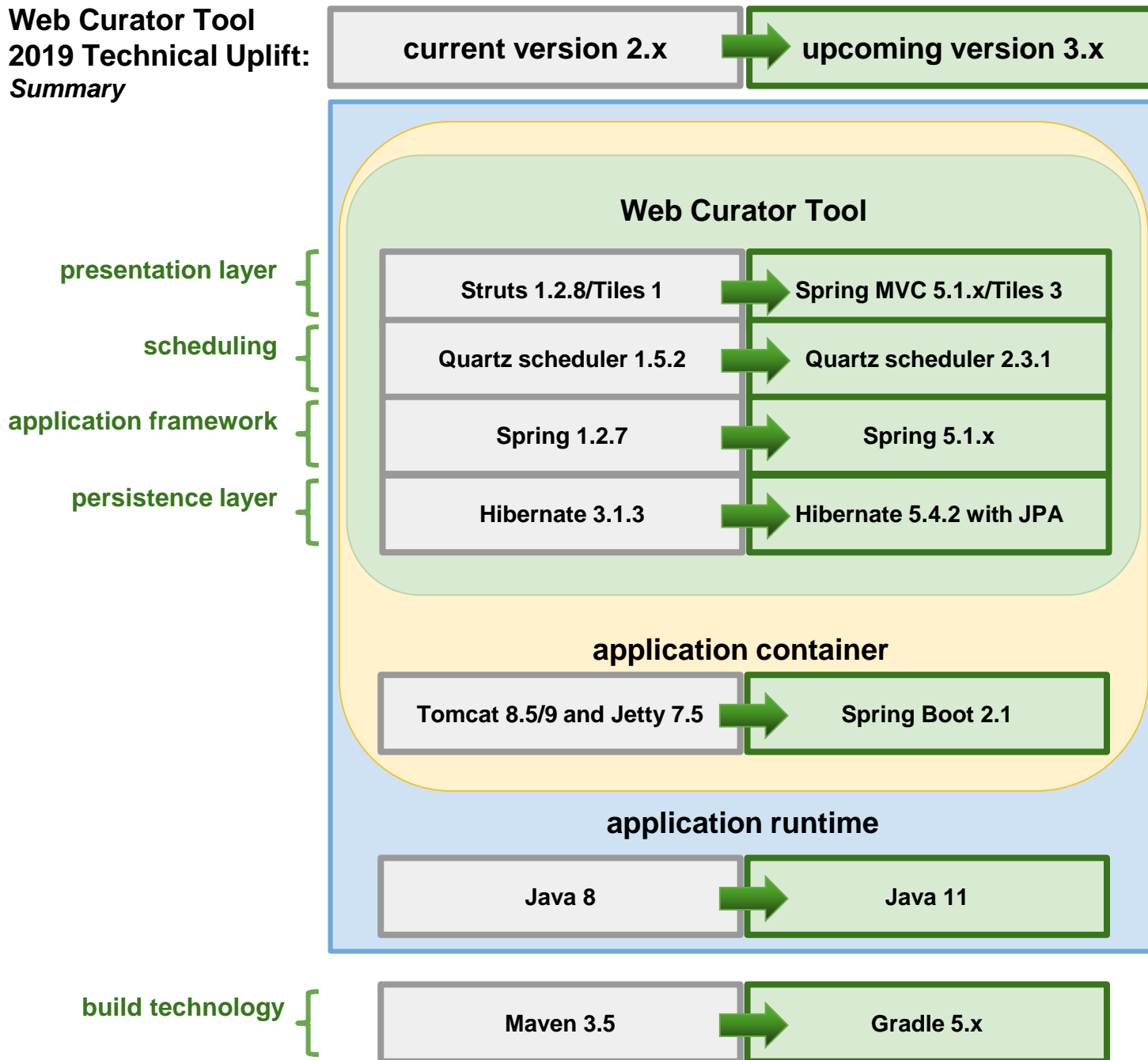
# WCT facts & figures

- The Web Curator Tool (WCT) is an open-source workflow management application for selective web archiving
- WCT 1.0: first release in 2006
- WCT 2.0: major upgrade, available since November 2018
- Currently 17 organizations worldwide are using it or considering start using

## Key features:

- Easy to use by any collection manager
- Embedded Heritrix H1 & H3 crawler engine
- Supports permissions, job scheduling, harvesting, quality review, and the collection of descriptive metadata.

# Web Curator Tool 2019 Technical Uplift: Summary



## Legend

### current technology

*current 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

# Check out our poster!

## Visit our website at:

<http://dia-nz.github.io/webcurator/>



V. 3.0 in Development

# WCT Technical Uplift

A Collaboration between the NLNZ and KBNL

## What is WCT?

The Web Curator Tool (WCT) is a free open-source workflow management application for selective Web archiving. It is designed for use in libraries and other collecting organisations, and supports collection by non-technical users while still allowing complete control of the web harvesting process. It is integrated with the Heritrix web crawler and supports key processes such as [permissions](#), [job scheduling](#), [harvesting](#), [quality review](#), and the [collection of descriptive metadata](#).

The WCT was first developed in 2006 as a collaborative effort by the National Library of New Zealand (NLNZ) and the British Library, initiated by the International Internet Preservation Consortium. In 2017, the NLNZ and the National Library of the Netherlands (KBNL) started on a new collaborative effort to uplift the technology and functionality of the WCT. Version 2.0 was released in November 2018.

## What is the WCT?



## What Comes Next for the WCT?

The NLNZ and KBNL will continue to improve the WCT to support their Web archiving programmes and to make available for use by the Web archiving community.

## Future Goals of the WCT

- Ensure that the WCT can keep up-to-date with Web and social media crawling techniques.
- Help users of older versions of WCT upgrade to WCT 2.x.
- Encourage wider use and development/support by the community

## 2018 Work Plan

- ✓ Heritrix 3.0 integration
- ✓ Improved and simplified installation process
- ✓ Updated and more accessible documentation

## 2019 Work Plan

- Technical uplift
- Crawler abstraction
- User journeys
- Functional uplift
- Cloud-ready
- QA improvements
- Let us know what else you would like to see in the WCT

## 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](http://webcurator.slack.com)

## 2019 Technical Uplift

We are modernising the underlying technologies so that we have:

- a stable foundation for adding new functionality
- supported technologies
- a larger pool of potential maintainers
- a clearer pathway to integrating 3rd party applications and services

## Lessons Learned

- Technical debt will catch up with you - we have to go through a series of intermediate version upgrades to get to the final upgrade, and the documentation for this old technology is scarce
- Automated testing would have reduced the amount of time to validate each stage of the upgrade

## 2019 Crawler Abstraction

We want to make the WCT less coupled with a single crawler so that it can work with a variety of crawl tools such as [Boulder](#) and [WebRecorder](#). We are migrating to a REST API and re-architecting the way crawlers are configured and managed in WCT.

## 2019 User Journeys

We are documenting how users interact with WCT to feed into automated tests, ensuring needed functionality remains across development changes.

