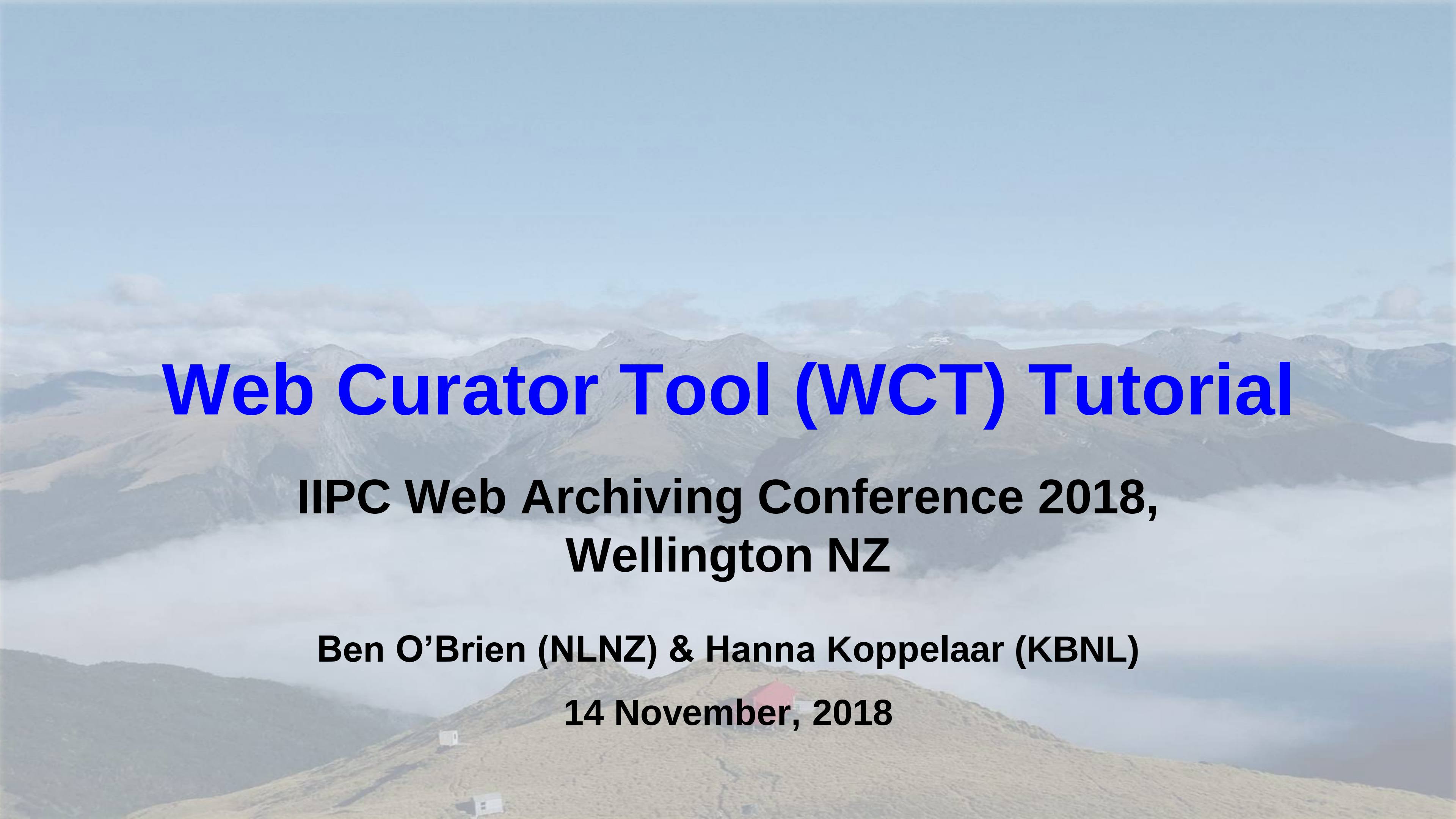


Web Curator Tool (WCT) Tutorial

IIPC Web Archiving Conference 2018,
Wellington NZ

Ben O’Brien (NLNZ) & Hanna Koppelaar (KBNL)

14 November, 2018

Today’s agenda

- Introduction
  - About us
  - What is the WCT?
- Then
  - A brief history of the WCT
  - The WCT state until recently
  - Why did we stick with the WCT?
- Now
  - Heritrix 3
  - Collaborative development
  - Building a work plan
  - Demos
- Future

A bit about us

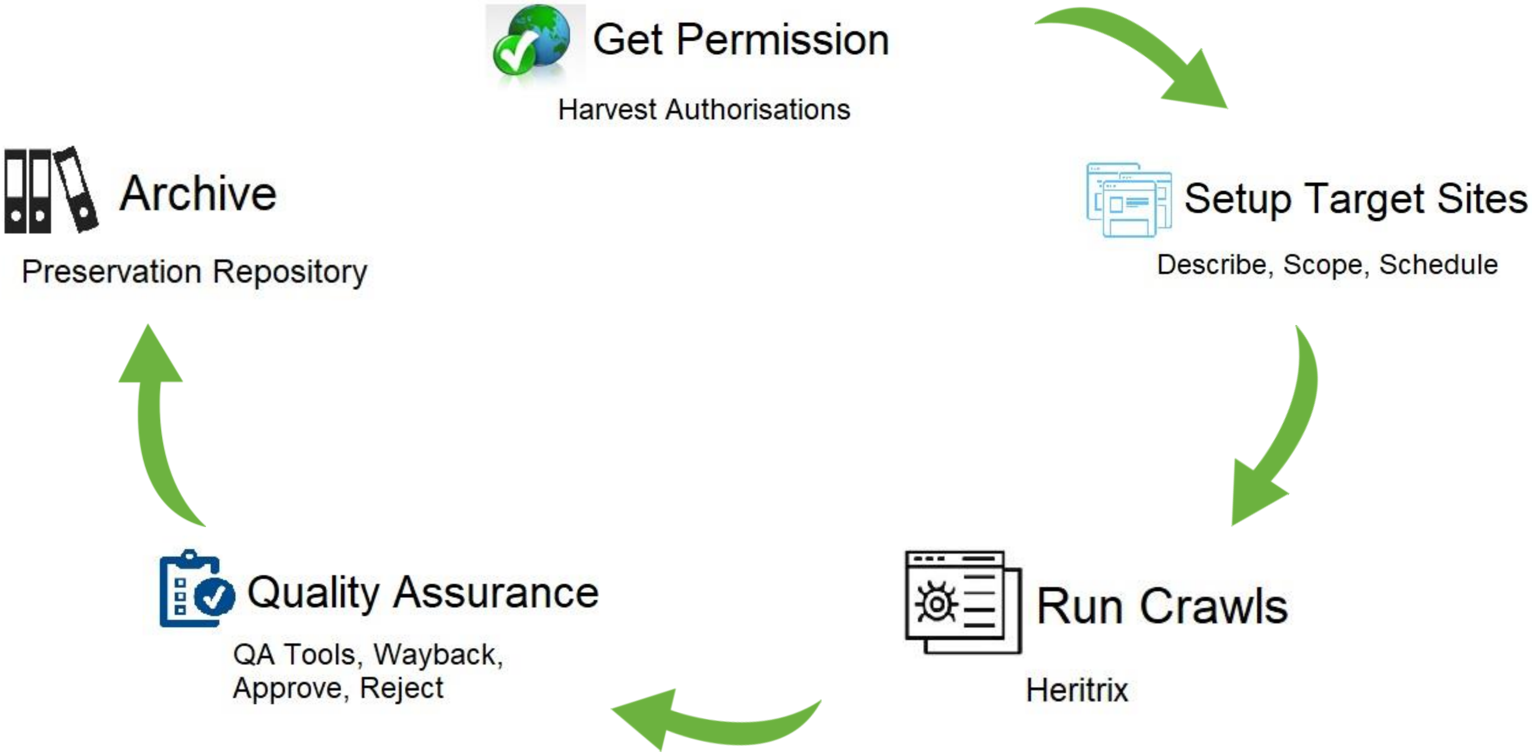
Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY
OF NEW ZEALAND

- Archiving the Web since 1999
- Selective Web archiving and domain crawling
- Legal deposit legislation since 2003

KB

- 13800 sites as of June 2018
- Selective Web crawling since 2007
- No legal deposit

What is the WCT?



Get Permission

Harvest Authorisations

Setup Target Sites

Describe, Scope, Schedule

Run Crawls

Heritrix

Quality Assurance

QA Tools, Wayback,
Approve, Reject

Archive

Preservation Repository

What the WCT is not

- It is NOT a digital archive or document repository
- It is NOT appropriate for long-term storage
  - It submits material to an external archive
- It is NOT an access tool
  - It does not provide public access to harvested material
- It is NOT a cataloguing system
  - It does allow you to record external catalog numbers
  - And it does allow you to describe harvests with Dublin Core metadata
- It is NOT a document management system
- It does NOT store all your communications with publishers
  - But it may initiate these communications
  - And it does record the outcome of these communications

A brief history of the WCT

Development began in 2006 as a collaborative open source project between the British Library and the National Library of New Zealand, created to solve the challenges of capturing online content using Heritrix

Initial objectives:

- Meet the needs of the BL and the NLNZ
- Modular & extendable
- Use for managing permissions, selection, description, scoping, harvesting, quality assurance
- Not require deep technical knowledge to archive Web content

The WCT state until recently

Running on outdated and unsupported libraries and frameworks

Tight integration with Heritrix v1

Convoluted installation process

Outdated documentation

Why did we stick with the WCT?

Three options:

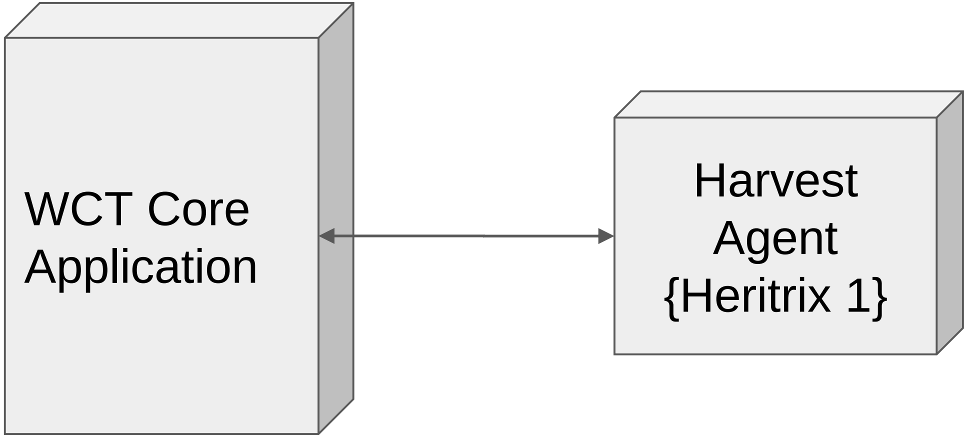
1. Build a new tool
2. Switch to an alternative tool
3. Upgrade the WCT

Why did we stick with the WCT?

Three options:

1. Build a new tool → too expensive
2. Switch to an alternative tool → no complete replacement could be found →
3. Upgrade the WCT would be possible to meet our requirements

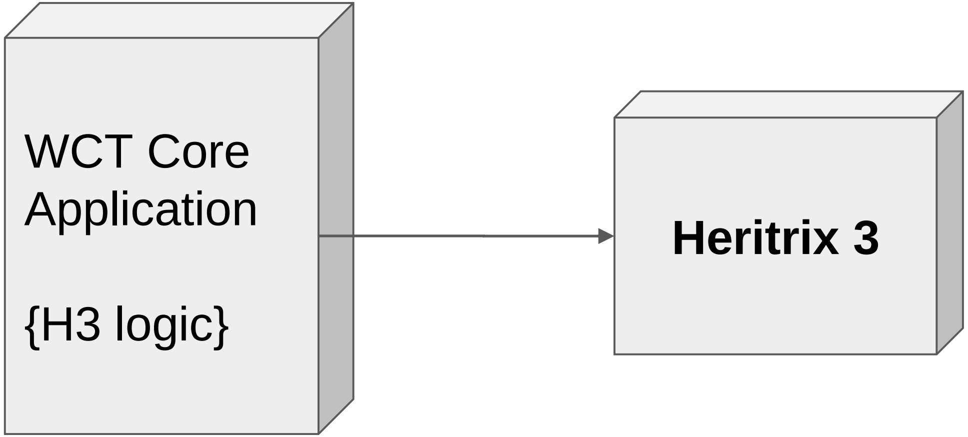
Heritrix 3



WCT Core Application

Harvest Agent {Heritrix 1}

Heritrix 3

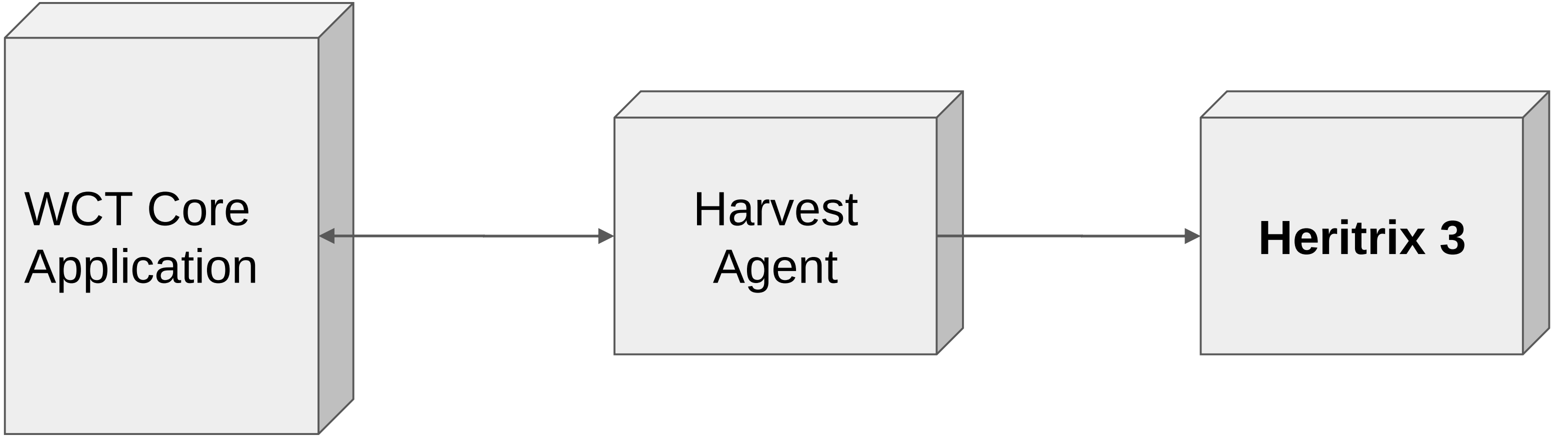


WCT Core Application

{H3 logic}

Heritrix 3

Heritrix 3

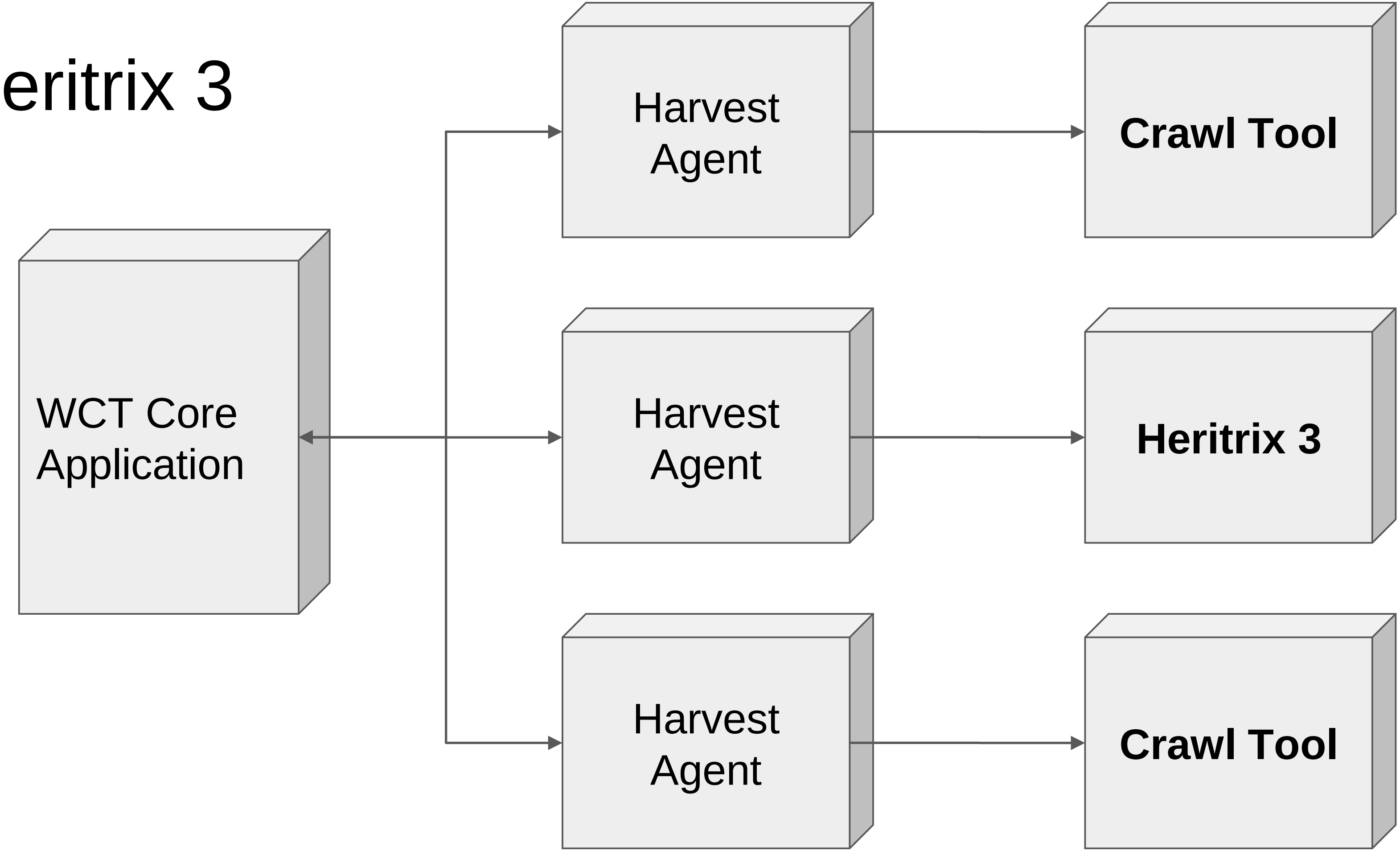


WCT Core Application

Harvest Agent

Heritrix 3

Heritrix 3


WCT Core Application
Harvest Agent
Harvest Agent
Harvest Agent
Crawl Tool
Heritrix 3
Crawl Tool

New collaborative development

National Library of New Zealand + National Library of the Netherlands

Coinciding independent reviews of WCT produced the same outcome

- Webex meetings
- Slack
- Email
- Google Docs
- Github

Development guidelines

Work plan

Milestone 1 - Complete Heritrix 3 integration

Milestone 2 - Update documentation and improve installation process

Milestone 3 - Technical uplift

Milestone 4 - Functional uplift

Milestone 5 - Abstraction from Heritrix

Work plan

Milestone 1 - Complete Heritrix 3 integration ✓

Milestone 2 - Update documentation and improve installation process ✓

Milestone 3 - Technical uplift

Milestone 4 - Functional uplift

Milestone 5 - Abstraction from Heritrix

WEB CURATOR TOOL

WCT 2.0 Demo

Future goals

Ensure that the WCT can keep up-to-date with Web and social media crawling techniques (plug-n-play third party crawlers and other tools)

Move users of older versions of WCT to WCT 2.0 (provide assistance and upgrade documentation)

The WCT is widely used and has a large development / support base (used by and developed by the community)

Questions?

Ben O’Brien

Ben.O’Brien@dia.govt.nz

Hanna Koppelaar

Hanna.Koppelaar@KB.nl

github.com/DIA-NZ/webcurator

webcurator.slack.com

webcuratortool.rtfd.io