



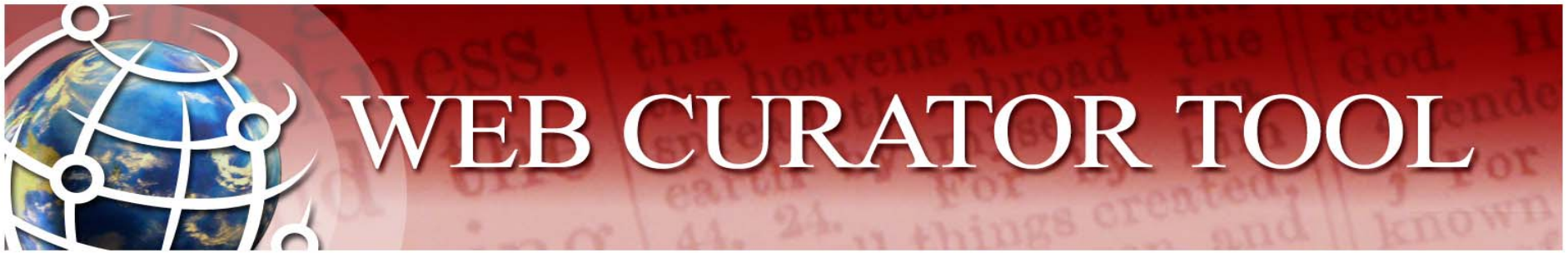
WEB CURATOR TOOL

The IIPC Web Curator Tool:

An Open Source Solution for
Selective Web Harvesting

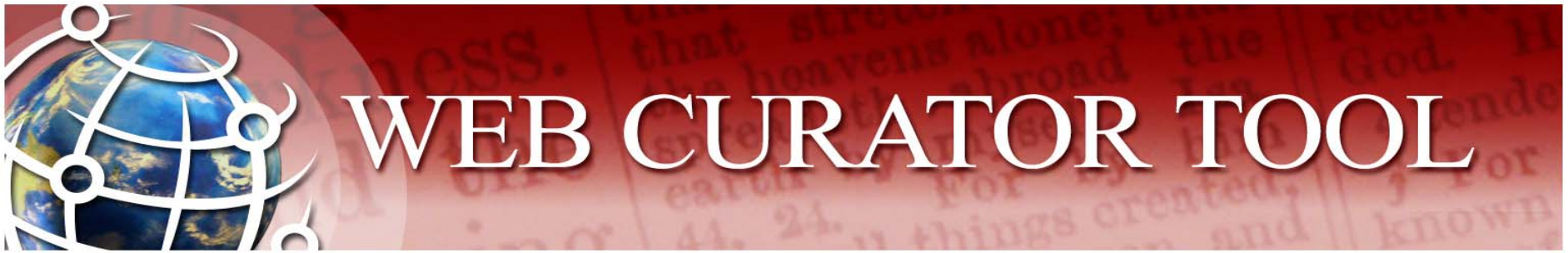
<http://webcurator.sourceforge.net>

Philip Beresford and Ravish Mistry
The British Library November 2007



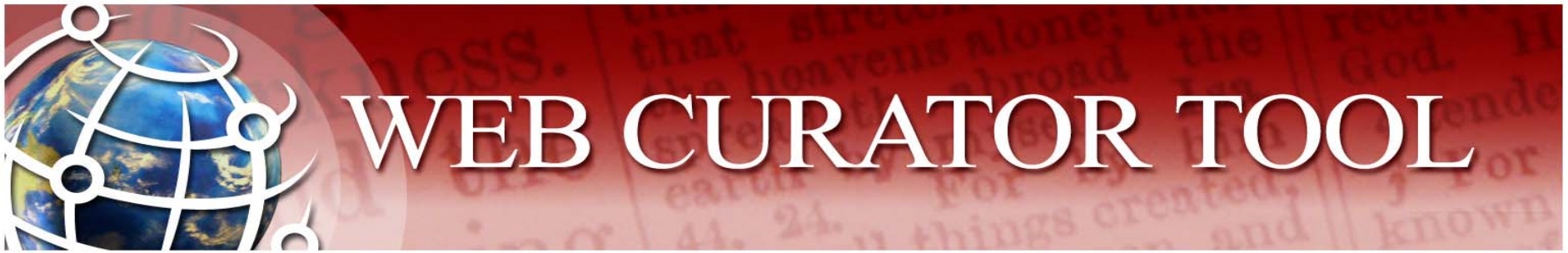
What is the WCT?

- **The Web Curator Tool (WCT) is a tool for managing the selective web harvesting process.**
- It is designed for use in libraries by non-technical users.
- It aims to manage the workflow for curators collecting web materials for addition to a digital repository.
- It is open-source software available for anyone to download and use free, and to contribute to its future development.



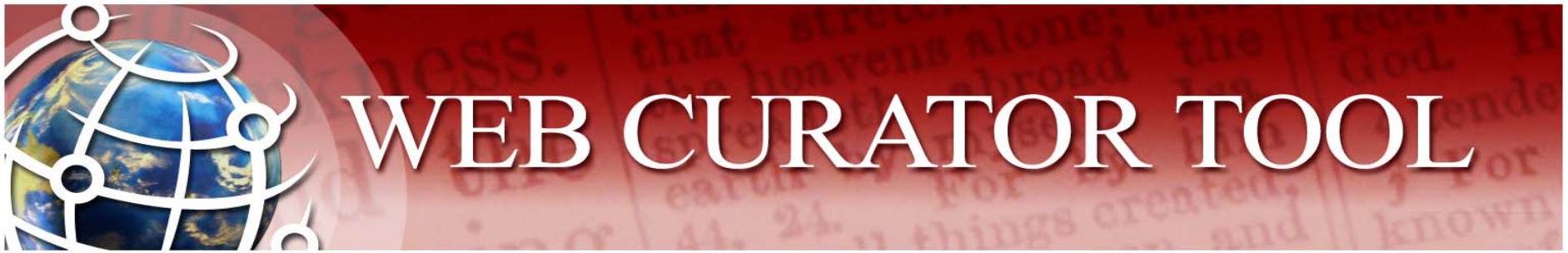
Project history

- **WCT was developed as a collaborative project between National Library of New Zealand and the BL**
- From initial joint requirements workshop through design to first release took 6 months (April - September 2006).
- Now implemented and operational at NLNZ.
- Target for implementation (on outsourced infrastructure) for use by UKWAC - Feb 2008.



What does it do?

- The WCT supports:
 - **Harvest Authorisation:** getting permission to harvest web material and make it available.
 - **Selection, scoping and scheduling:** what will be harvested, how, and how often?
 - **Description:** Dublin Core metadata.
 - **Harvesting:** Downloading the material at the appointed time with the Heritrix web harvester deployed on multiple machines.
 - **Quality Review:** making sure the harvest worked as expected, and correcting simple harvest errors.
 - **Submitting** the harvest results to a digital archive.



What is it 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
 - (But it does let you review your harvests)
 - You should use Wayback or WERA as access tools



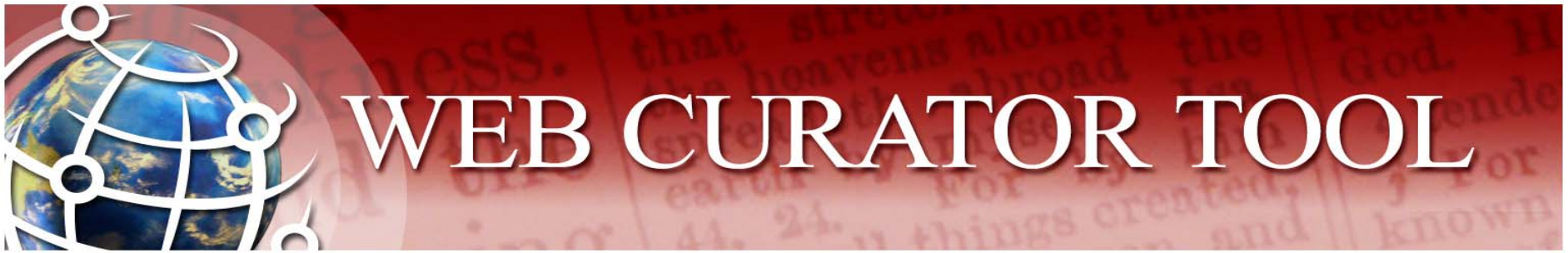
What is it NOT?

- 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



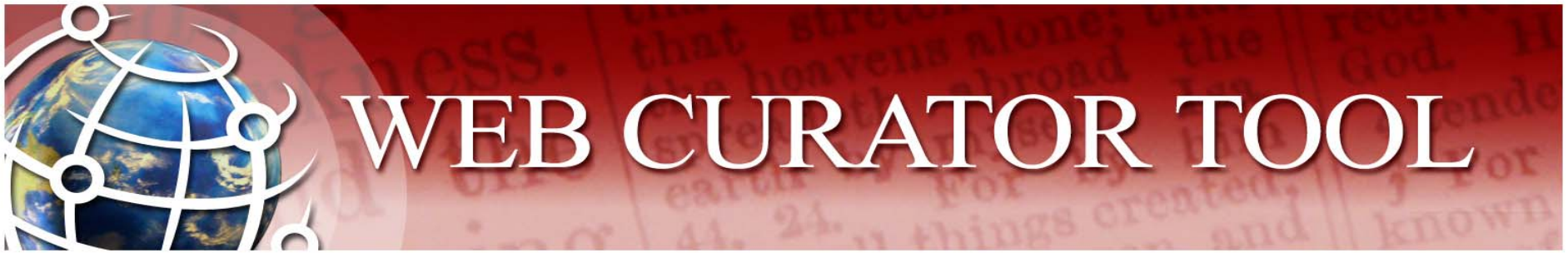
Technology

- Implemented in Java
- Runs in Apache Tomcat
- Incorporates parts or all of
 - Acegi Security System
 - Apache Axis (SOAP data transfer)
 - Apache Commons Logging
 - **Heritrix (version 1.8)**
 - Hibernate (database connectivity)
 - Quartz (scheduling)
 - Spring Application Framework
 - **Wayback**



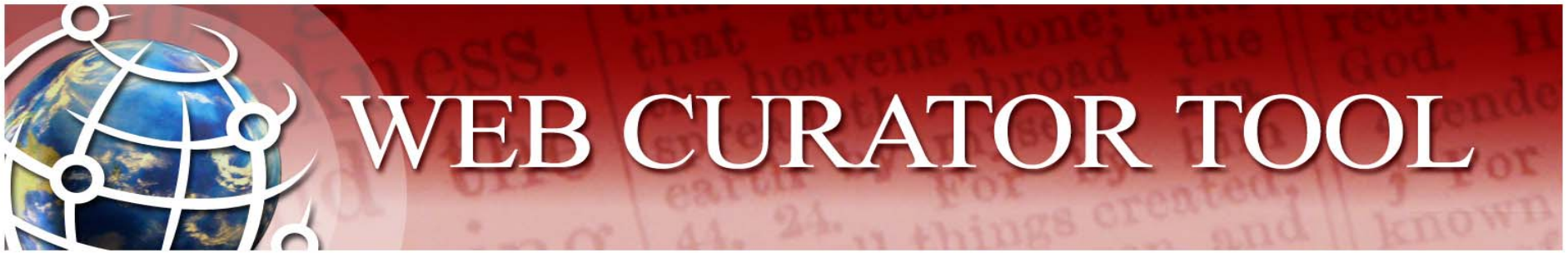
More technology

- Platform:
 - Tested on Solaris (version 9) and Red Hat Linux
 - Developed on Windows
 - Should work on any platform that supports Apache Tomcat
- Database:
 - A relational database is required
 - Tested on Oracle and PostgreSQL
 - Installation scripts provided for Oracle and PostgreSQL
 - Should work with any database that Hibernate supports
 - Including MySQL, Microsoft SQL Server, and about 20 others



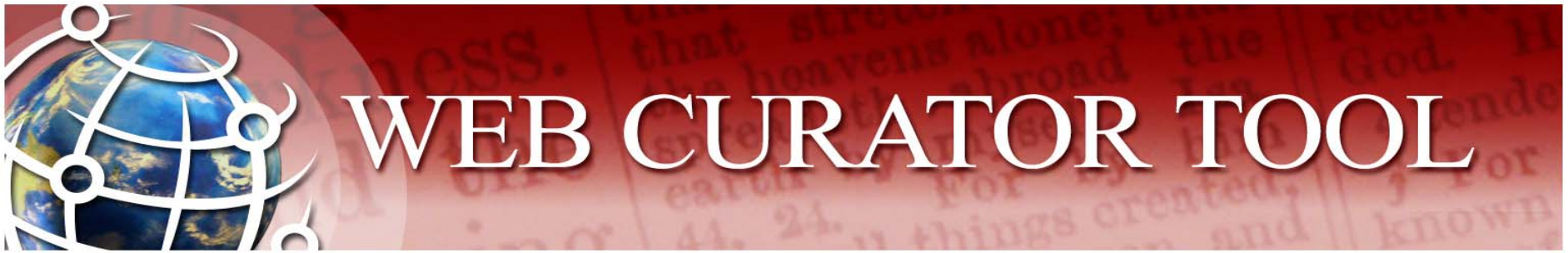
Current status

- In testing at the British Library as a harvesting tool for UKWAC (to replace PANDAS)
- Software in development to convert existing UKWAC archive (about 10,000 instances in nearly 2 Tb) to ARC format files
- Likely to form the basis of a selective web archiving service offered by the BL - initially to UKWAC and then potentially to other institutions



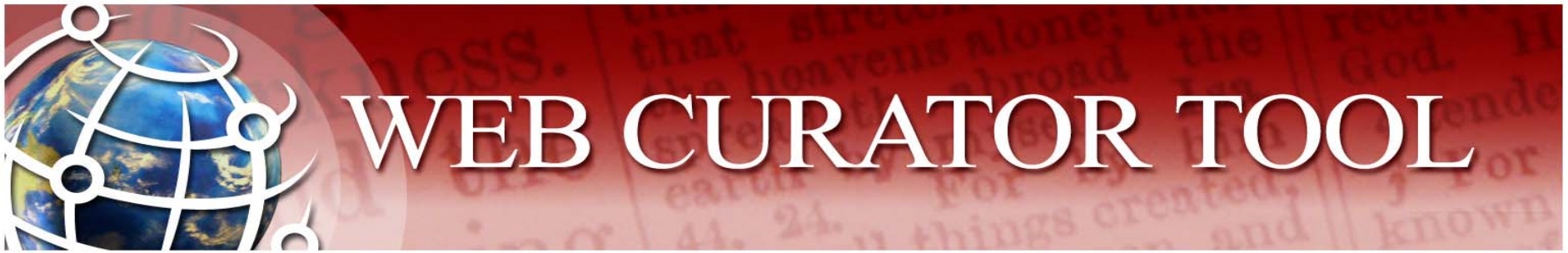
Immediate priorities for development

- Performance optimisation
- Streamlining user interface and workflow
- Develop additional quality review tools
- Facilitate interaction with later versions of Heritrix
- Improve support for other platforms and databases
- Integration with open-source Wayback Machine for access



Future directions

- After Legal Deposit regulations are implemented (2009?), Copyright Libraries will start to take 'wholesale' snapshots of UK webspace (using 'Smart Crawler' extensions to Heritrix)
- Selective archiving will still have its place, so WCT will continue to be used and will evolve to meet future UKWAC requirements
- Need to develop fuller technical metadata capture for future preservation actions
- Need to improve resource discovery (access)
- Will consider other 'smarter' ideas - e.g. capturing change rate, and usage data to govern adjustment of harvesting schedules



Further information:

WCT can be downloaded from SourceForge at:

<http://webcurator.sourceforge.net/>


Here you can also find installation and user documentation,
and lists of software issues and potential development
requirements



Sample screen-shots:

....





Login v1.1GA

username

password



Edit View History Bookmarks Tools Help

http://194.66.226.132:8080/wct/curator/home.html

Getting Started Latest Headlines Mozilla Firefox Start P...

sable Cookies CSS Forms Images Information Miscellaneous Outline Resize Tools View Source Options

WEB CURATOR TOOL


[Home](#) | [Queue](#) | [Help](#) | [Logout](#)
User njohnson is logged in.



In Tray

8 tasks, 2 Notifications

[open](#)



Harvest Authorisations

2 harvest authorisations

[open](#) [add new](#)



Targets

1 Targets

[open](#)



Target Instances

0 Scheduled instances, 11 ready for Quality reviews


[open](#) [queue](#)



Groups

1 Target Groups

[open](#)



Permission Request Templates

[open](#) [add new](#)




Reports

[open](#)



Harvester Configuration

[general](#) [bandwidth](#) [profile](#)



Users, Roles & Agencies

Users: [open](#) [add new](#)

Roles: [open](#) [add new](#)

Agencies: [open](#)



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

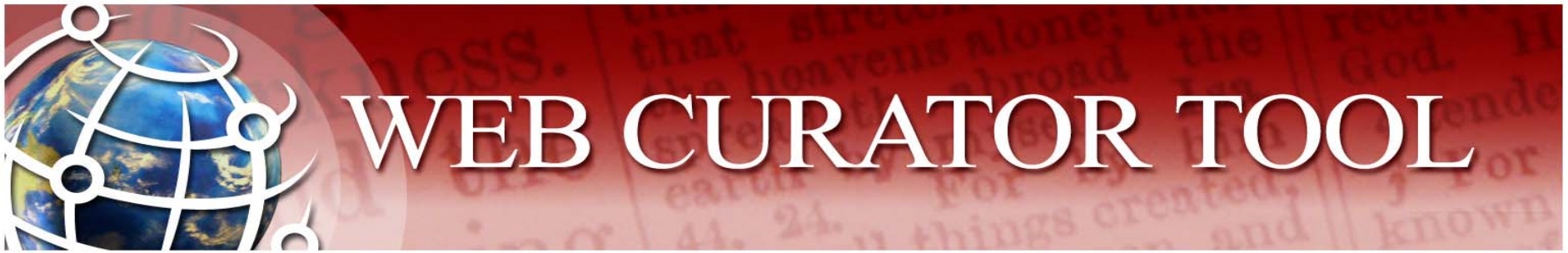
In Tray Harvest Authorisations Targets Groups Target Instances Reports Management



In Tray

Tasks

| Date | Subject | Owner | Action |
|-------------------------|---------------------------|------------|--------|
| 2007-04-17 10:38:39.873 | Endorse Harvest '2424832' | Unclaimed | |
| 2007-03-16 14:59:02.358 | Endorse Harvest '1867791' | N. Johnson | |
| 2007-03-16 11:05:01.132 | Endorse Harvest '1867789' | N. Johnson | |
| 2007-03-16 10:47:25.234 | Endorse Harvest '1867787' | N. Johnson | |
| 2007-03-13 17:40:58.619 | Endorse Harvest '1867786' | Unclaimed | |
| 2007-03-13 17:33:26.825 | Endorse Harvest '1867785' | Unclaimed | |
| 2007-03-13 17:24:58.607 | Endorse Harvest '1867784' | Unclaimed | |



Harvest Authorisations

- WCT Harvest Authorisation is concerned with:
 - Permission to harvest web material
 - Permission to make web material accessible to users
 - Any and all special conditions that apply



[Home](#) | [Queue](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray **Harvest Authorisations** Targets Target Instances Groups Management



Harvest Authorisations

Search

ID: Name: Authorising Agent:

Order No: Show Disabled: ☐

Results

[create new](#)

| Id | Name | Auth Agent | Order No | Action |
|---------|---------------------|---------------------|----------|--------|
| 2195456 | Phil test | | | |
| 32768 | Web Archive Testing | The British Library | | |

Results 1 to 2 of 2

Page 1 of 1



Harvest Authorisations

Opal Coast Tours

General

URLs

Authorising Agencies

Permissions

Id: 2162689

Name: Opal Coast Tours

Description: Authorisation to harvest the website granted by site owner

Order No:

Published: ☐

Enabled: ☒

Annotations

add

Date

User

Notes

No annotations are available for this Harvest Authorisation

save

cancel



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray **Harvest Authorisations** Targets Groups Target Instances Reports Management



Harvest Authorisations

Opal Coast Tours

General **URLs** Authorising Agencies Permissions

New URL Pattern:

add

URL Pattern

http://www.fencott.com/Clive/OpalCoast/*

Action



save

cancel

In Tray | **Harvest Authorisations** | Targets | Groups | Target Instances | Reports | Management



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray **Harvest Authorisations** Targets Groups Target Instances Reports Management



Harvest Authorisations

Opal Coast Tours

General URLs **Authorising Agencies** Permissions

search

create new

Authorising Agency

Contact

Action

Opal Coast Tours

Clive Fencott



save

cancel

In Tray | **Harvest Authorisations** | Targets | Groups | Target Instances | Reports | Management



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray **Harvest Authorisations** Targets Groups Target Instances Reports Management



Harvest Authorisations

Opal Coast Tours

| | |
|--------------|--|
| Name: | <input type="text" value="Opal Coast Tours"/> |
| Description: | <input type="text" value="owner of the website"/> |
| Contact: | <input type="text" value="Clive Fencott"/> |
| Phone: | <input type="text" value="01642 384540"/> |
| Email: | <input type="text" value="p.c.fencott@tees.ac.uk"/> |
| Address: | <input type="text" value="Department of Computing, University of Teesside, UK"/> |

save

cancel

In Tray | **Harvest Authorisations** | Targets | Groups | Target Instances | Reports | Management



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray **Harvest Authorisations** Targets Groups Target Instances Reports Management



Harvest Authorisations

Opal Coast Tours

General URLs Authorising Agencies **Permissions**

[create new](#)

| Status | Authorising Agent | From | To | URL Patterns | Action |
|----------|-------------------|------------|----|--|---|
| Approved | Opal Coast Tours | 17/04/2007 | | http://www.fencott.com/Clive/OpalCoast/* |    |

[save](#)

[cancel](#)

In Tray | [Harvest Authorisations](#) | [Targets](#) | [Groups](#) | [Target Instances](#) | [Reports](#) | [Management](#)




Harvest Authorisations



Opal Coast Tours

Authorising Agent: Opal Coast Tours 


Dates: 17/04/2007 to dd/mm/yyyy

Status: Pending 

Special Restrictions: no restrictions apply  

Copyright Statement: crown copyright does not apply to this site  

Copyright URL:

Access Status: Open (unrestricted) access 

Open Access Date:

Quick Pick: ☐

Display Name:

Urls: ☒ <http://www.fencott.com/Clive/OpalCoast/>*

File Reference:

Assign Approval Task: No 

Exclusions

URL

Reason



WEB CURATOR TOOL

Targets

- A Target is a portion of the web you want to harvest.
- The Target is the “unit of selection”:
 - If there is something you want to harvest and archive and describe, then it is a Target.
- You can attach a Schedule to a Target to specify when (and how often) it will be harvested.
 - But you can’t harvest until you have permission to harvest, and you can’t harvest until the selection is approved.



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray Harvest Authorisations **Targets** Groups Target Instances Reports Management



Targets



Search

Name:

Seed:

Agency:

User:

State:

☐

Pending

☐

Reinstated

☐

Nominated

☐

Rejected

☐

Approved

☐

Cancelled

search

reset

Results

[create new](#)

| Id | Name | Agency | Owner | Status | Seeds | Action |
|---------|-----------------------------|---------------------|----------------|------------|---|--------|
| 1900547 | Border Reivers 2 | The British Library | Nicola Johnson | Reinstated | http://www.borderreivers.co.uk/ | |
| 2064384 | BT Group | The British Library | Nicola Johnson | Approved | http://www.btplc.com/ | |
| 1900546 | Cape Farewell 2 | The British Library | Nicola Johnson | Approved | http://www.capefarewell.com/ | |
| 1900548 | Human Scale Education | The British Library | Nicola Johnson | Approved | http://www.hse.org.uk/ | |
| 1966080 | Marine Conservation Society | The British Library | Nicola Johnson | Approved | http://www.mcsuk.org/ | |
| 2326528 | Opal Coast | The British Library | Nicola Johnson | Pending | http://www.seaham.i12.com/oc/ocindex.htm | |



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray Harvest Authorisations **Targets** Groups Target Instances Reports Management



Targets

General Seeds Profile Schedule Annotations Description

Id:

Name:

Opal Coast Tours

Description:

Website of Clive Fencott

Reference Number:

Run on Approval:



Owner:

Nicola Johnson

State:

Pending

save

cancel

In Tray | Harvest Authorisations | Targets | Groups | Target Instances | Reports | Management



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray Harvest Authorisations **Targets** Groups Target Instances Reports Management



Targets

Opal Coast Tours

General **Seeds** Profile Schedule Annotations Description

Seed:

Authorisation:

add

import

| Seed | Primary | Harvest Auth | Auth Agent | Start | End | Status | Action |
|---|-------------------------------------|---------------------------|------------------|------------|-----|----------|--------|
| http://www.fencott.com/Clive/OpalCoast* | <input checked="" type="checkbox"/> | UK Webspace Authorisation | UK Government | 15/01/2007 | | Approved | |
| | | Opal Coast Tours | Opal Coast Tours | 17/04/2007 | | Approved | |

add

save

cancel

In Tray | Harvest Authorisations | Targets | Groups | Target Instances | Reports | Management



Targets

Opal Coast Tours

General Seeds **Profile** Schedule Annotations Description

Base Profile

Base Profile: Default - The British Library

Profile Overrides

| Profile Element | Override Value | Enable Override |
|------------------------|----------------|--------------------------|
| Robot Honouring Policy | classic | <input type="checkbox"/> |
| Maximum Hours | 0 | <input type="checkbox"/> |
| Maximum Kilobytes | 0 | <input type="checkbox"/> |
| Maximum Documents | | <input type="checkbox"/> |
| Maximum Path Depth | | <input type="checkbox"/> |
| Maximum Hops | | <input type="checkbox"/> |
| Exclude Filters | | <input type="checkbox"/> |
| Force Accept Filters | | <input type="checkbox"/> |
| Excluded MIME Types | | <input type="checkbox"/> |



Targets

Opal Coast Tours

[create new](#)

| Schedule | Owner | Next Scheduled Time | Action |
|----------|-------|---------------------|--------|
|----------|-------|---------------------|--------|

[save](#) [cancel](#)



[Home](#) | [Queue](#) | [Help](#) | [Logout](#)
User njohnson is logged in

Tray Harvest Authorisations **Targets** Target Instances Groups Management



Targets

I Beatrice

From Date:

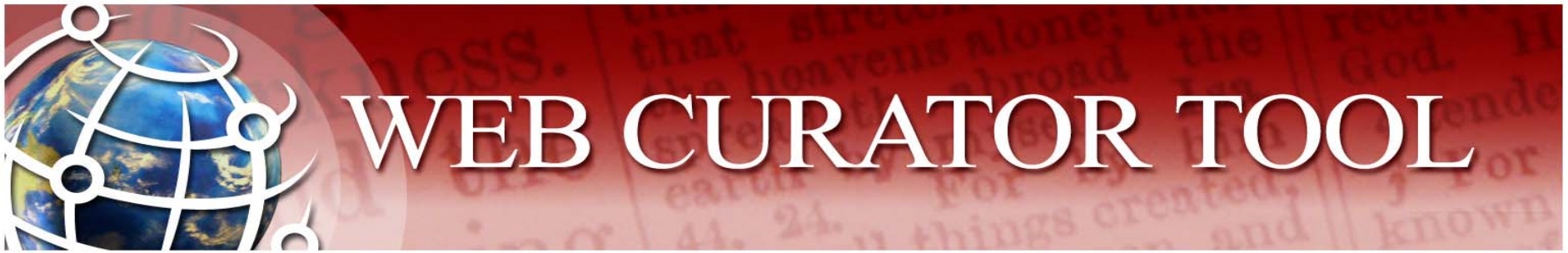
To Date:

Type:

Time:

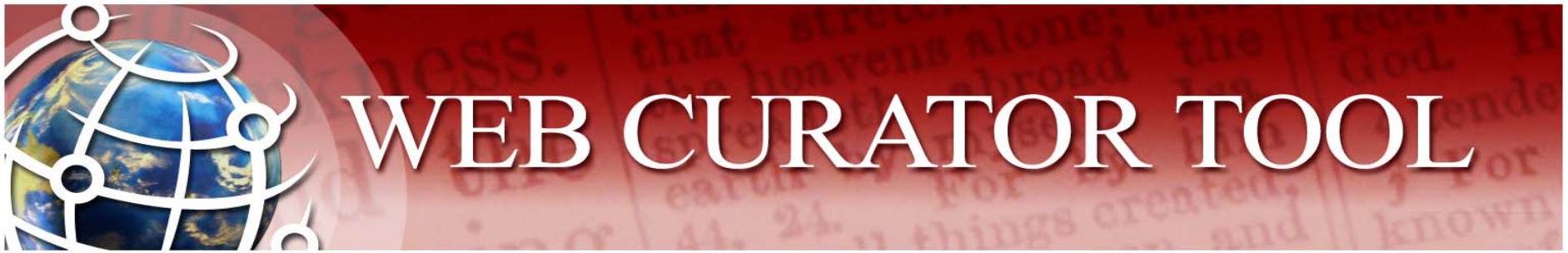
Day of Month:

In Tray | [Harvest Authorisations](#) | [Targets](#) | [Groups](#) | [Target Instances](#) | [Reports](#) | [Management](#)



Target Instances

- Target instances represent individual harvests that
 - Are scheduled to happen, or
 - Are in progress, or
 - Have finished.
- Target Instances are created automatically for a Target when that Target is Approved.
 - A Target Instance is created for each harvest that has been scheduled.



Target Instances - the Queue

- Scheduled Target Instances are put in a queue
- When their scheduled start time arrives
 1. The WCT allocates the harvest to one of the harvesters
 2. The harvester invokes Heritrix and harvests the requested material
 3. When the harvest is complete, the User is notified
- Examining the Queue gives you a good idea of the current state of the system
 - The WCT provides a quick view of the instances in the Queue, including Running, Paused, Queued, and Scheduled Instances



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray Harvest Authorisations Targets Groups **Target Instances** Reports Management



Target Instances



Search

From: 16/03/2007 00:00:00 To: 17/07/2007 23:59:00 Agency: The British Library Owner: Nicola Johnson

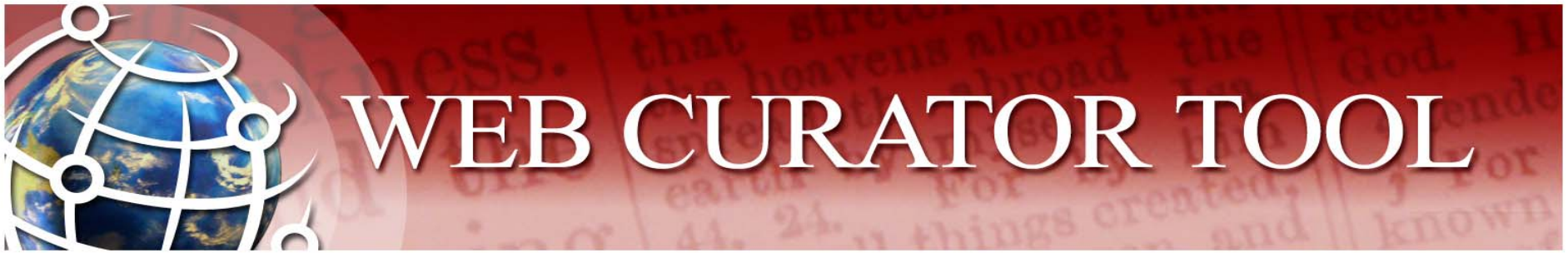
Name: State: ☐ Scheduled ☐ Queued ☐ Running ☐ Paused ☐ Harvested ☐ Aborted ☐ Endorsed ☐ Rejected ☐ Archived

search

reset

Results

| Id | Name | Harvest Date | State | Owner | Run Time | Data Downloaded | Action |
|---------|------------------|---------------------|-----------|------------|-------------|-----------------|--------|
| 1867792 | V&A | 21/03/2007 11:28:02 | Running | N. Johnson | 01:23:27:23 | 12.94 GB | |
| 2129920 | BT Group | 11/04/2007 10:06:17 | Running | N. Johnson | 00:01:03:34 | 935.01 MB | |
| 2424833 | Opal Coast Tours | 17/04/2007 23:00:00 | Scheduled | N. Johnson | | | |
| 1867787 | Seniors Network | 16/03/2007 10:44:32 | Harvested | N. Johnson | 00:00:02:21 | 9.36 MB | |
| 1867789 | Cape Farewell 2 | 16/03/2007 10:58:32 | Harvested | N. Johnson | 00:00:05:57 | 3.88 MB | |
| 1867790 | Border Reivers 2 | 16/03/2007 11:01:02 | Harvested | N. Johnson | 00:00:03:19 | 5.64 MB | |



Target Instances - the User's view

- When a harvest is complete, its Owner is notified
- The Owner (or another User) then has to
 1. Quality Review the harvest result to see if it was successful
 - Browse Tool: Browse the harvest result to ensure all the content is there
 - Prune Tool: Delete unwanted material from the harvest
 2. Endorse or Reject the harvest
 3. Submit the harvest to an Archive (if it has been endorsed)
- The User view of the Target Instances shows all the instances that the user owns.



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray Harvest Authorisations Targets Groups **Target Instances** Reports Management



Target Instances

Opal Coast Tours (2424837)

General Profile State Logs Harvest Results Annotations

Id: 2424837
Target Name: [Opal Coast Tours](#)
Schedule: 17/04/2007 12:34:39
Actual Start: 17/04/2007 12:35:01
Priority: Normal
Owner:
Agency: The British Library
State: Harvested
Bandwidth Percentage: Default
Allocated Bandwidth: 166 KB

save

cancel

In Tray | Harvest Authorisations | Targets | Groups | Target Instances | Reports | Management



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray Harvest Authorisations Targets Groups **Target Instances** Reports Management



Target Instances

Opal Coast Tours (2424837)

General Profile State Logs **Harvest Results** Annotations

| Date | User | Notes | State | Action |
|---------------------|------------|------------------|-------|---|
| 17/04/2007 12:35:48 | N. Johnson | Original Harvest | | Review Endorse Reject |

save

cancel

In Tray | Harvest Authorisations | Targets | Groups | Target Instances | Reports | Management



[Home](#) | [Queue](#) | [Help](#) | [Logout](#)
User njohnson is logged in

Tray Harvest Authorisations Targets **Target Instances** Groups Management



Target Instances

I Beatrice (2523136)

Quality Review Tools

Browse

<http://ibeatrice.blogspot.com/>

[Review this Harvest](#) | [Live Site](#) | [Archived Harvested](#)

Harvest History Tool

[View Harvest History](#)

Prune Tool

[Prune Tool](#)

done

In Tray | [Harvest Authorisations](#) | [Targets](#) | [Groups](#) | [Target Instances](#) | [Reports](#) | [Management](#)



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray Harvest Authorisations Targets Groups **Target Instances** Reports Management



Target Instances

Opal Coast Tours (2424837)

Harvest

- http://www.fencott.com/
 - http://www.fencott.com/Clive/
 - http://www.fencott.com/Clive/OpalCoast/
 - http://www.fencott.com/Clive/OpalCoast/cloughton.gif
 - http://www.fencott.com/Clive/OpalCoast/coast.jpg
 - http://www.fencott.com/Clive/OpalCoast/document.all
 - http://www.fencott.com/Clive/OpalCoast/document.layers
 - http://www.fencott.com/Clive/OpalCoast/dscript41/
 - http://www.fencott.com/Clive/OpalCoast/easington.gif
 - http://www.fencott.com/Clive/OpalCoast/filey.gif
 - http://www.fencott.com/Clive/OpalCoast/flamboro.gif
 - http://www.fencott.com/Clive/OpalCoast/tylingthorpe.gif
 - http://www.fencott.com/Clive/OpalCoast/hartlepool.gif
 - http://www.fencott.com/Clive/OpalCoast/hummersea.gif
 - http://www.fencott.com/Clive/OpalCoast/index.htm
 - http://www.fencott.com/Clive/OpalCoast/marske.gif
 - http://www.fencott.com/Clive/OpalCoast/nothing.html

Prune Single Item

Prune Single Item and Children

Provenance Note:



[Home](#) | [Help](#) | [Logout](#)
User njohnson is logged in.

Tray Harvest Authorisations Targets Groups **Target Instances** Reports Management



Target Instances

Opal Coast Tours (2424837)

General Profile State Logs **Harvest Results** Annotations

| Date | User | Notes | State | Action |
|---------------------|------------|------------------|----------|-----------------------------------|
| 17/04/2007 12:35:48 | N. Johnson | Original Harvest | Endorsed | Submit to Archive |

save

cancel

In Tray | Harvest Authorisations | Targets | Groups | Target Instances | Reports | Management



WEB CURATOR TOOL: Search Groups - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://194.66.226.132:8080/wct/curator/groups/search.html

Getting Started Latest Headlines Mozilla Firefox Start P...

Disable Cookies CSS Forms Images Information Miscellaneous Outline Resize Tools View Source Options

WEB CURATOR TOOL

Home | Queue | Help | Logout
User njohnson is logged in.

In Tray Harvest Authorisations Targets Target Instances **Groups** Management



Groups

Search

ID: Name: Agency: Owner: Member of:

Results

| Id | Name | Agency | Owner | Status | Action |
|---------|--------|---------------------|----------------|----------|--------|
| 2555905 | Poetry | The British Library | Nicola Johnson | Inactive | |

Results 1 to 1 of 1

Page 1 of 1

In Tray | Harvest Authorisations | Targets | Groups | Target Instances | Reports | Management

http://194.66.226.132:8080/wct/curator/groups/search.html

start <



File Edit View History Bookmarks Tools Help

http://194.66.226.132:8080/wct/curator/report/report.html

Getting Started Latest Headlines Mozilla Firefox Start P...

Disable Cookies CSS Forms Images Information Miscellaneous Outline Resize Tools View Source Options

In Tray Harvest Authorisations Targets Target Instances Groups **Management**

Reports

☒ **System Usage Report**
Start Date is inclusive.
End Date is exclusive

Start Date: dd/MM/yyyy

End Date: dd/MM/yyyy

Agencies: (Optional)

☐ **System Activity Report**
Start Date is inclusive.
End Date is exclusive

Start Date: dd/MM/yyyy

End Date: dd/MM/yyyy

Agencies: (Optional)

Users: (Optional)

☐ **Crawler Activity Report**
Crawler Activity Report

Start Date: dd/MM/yyyy

End Date: dd/MM/yyyy

Done

start | Inbox - Microsoft Out... | WEB CURATOR TOOL... | Microsoft PowerPoint ... | Desktop | 12:55