Using ArchiveWeb.page to capture the Carmichael Watson project

Sara Day Thomson & Anisa Hawes

@sdaythomson
@AnisaHawes

Web Archiving & Preservation Working Group: General Meeting, 9 December 2021
Digital Preservation Coalition
• Largely from Centre for Research Collections

• Mainly the papers of pioneering folklorist Alexander Carmichael (1832-1912)

• Includes a rich corpus of materials from Gaelic-speaking areas of Scotland:
  ◦ Stories
  ◦ Songs
  ◦ Customs
  ◦ Beliefs
  ◦ Artefacts
Background

• A £750,000 investment by the Leverhulme Trust and AHRC

• Digitised archival and museum materials with transcriptions and other unique functionality

• Significantly improved user experience

• Popular resource with a very engaged stakeholder group (esp. teaching)
Imminent Risk of Loss

- Hosting infrastructure obsolete, security risk
- Code could not be ported to new server
- Taken offline in 2018
- Moved onto a Virtual Machine (RedHat 5) now itself obsolete and incurring annual licence fee
- UK Web Archive capture incomplete; unable to fix in timescale
Preservation Strategy

• VM secured behind VPN

• Manual Webrecorder tools for capture

• No in-house capacity, contract a web archivist (work began July 2021)

• Sustainable set of archival materials in a preservation standard format

• Opportunity to trial the preservation of web content

• Increase our capacity in web archiving
During the past few months, I've been collaborating with the University of Edinburgh to preserve the Carmichael Watson Project web resource.
I use ArchiveWeb.Page, a free, open-source web archiving tool created by Webrecorder. It's available as a stand-alone desktop app, or as a browser extension in Chrome.
The screenshots and reflections I am sharing today, represent a web archiving project-in-progress.
Before I begin creating any web archive, I make a detailed map. My knowledge and understanding of a website's structure, determine how I approach its capture.
The map informs how I plan my capture sessions, helps me to plot milestones against which I will track my progress, and provides a way-finder for retracing back to review my work.
Only by establishing the relative positions of a site's sub-sections, and charting their spatial dispersal, will I be able to be certain (at the end of the archiving process) that I have traversed its full extent.
An initial close-to-the-surface scan of the site revealed >3000 catalogued objects, across three collections. The Carmichael Watson Project brings together holdings of the West Highland Museum, the National Museum of Scotland and the University of Edinburgh.
Additionally, with a rough tally, I counted almost 4000 entries in the indexes, which cross-reference the names of people, places, subjects and ideas that are embedded within the material.
As I began mapping in more detail, I realised that the Carmichael Watson Project web resource has an unusually complex construction.
I found multiple URL sequences running in parallel.
In fact, I identified between six and ten different URL sequences intersecting each record. I observed that around half of those variations resolved to identical pages, but knew that all of the sequences would need to be captured to facilitate intuitive browsing in the archive.
The website's convoluted architecture compounded my task ahead.
The scale of this capture would far exceed our initial expectations.
I used ExCel to plot out each sequence, and began visualising the website as built in a series of tall columns,
<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
<th>Date</th>
<th>URL 1</th>
<th>URL 2</th>
<th>URL 3</th>
<th>URL 4</th>
<th>URL 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Silver coin</td>
<td>13-01-15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>British coin</td>
<td>13-01-15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Copper coin</td>
<td>13-01-24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Silver coin</td>
<td>13-01-26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Pewter communion bowl</td>
<td>13-01-26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>British coin</td>
<td>13-01-27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Scottish coin (1600 - 1700)</td>
<td>13-01-28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Silver coin (1600 - 1700)</td>
<td>13-01-29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Tin container (17th century)</td>
<td>13-01-30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Clip plate bowl</td>
<td>13-01-31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Silver Scottish pub coin</td>
<td>13-01-31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Two-truncups</td>
<td>13-01-32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Silver coin</td>
<td>13-01-33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Silver coin</td>
<td>13-01-34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Silver coin</td>
<td>13-01-35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Small spear or arrow</td>
<td>13-01-36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Square truss lam (1800</td>
<td>13-01-37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

noting anywhere I found gaps or hops in the sequences.
I worked steadily,
reviewing my progress in ArchiveWeb's index of Pages.
I expand ArchiveWeb’s information panel (upper right) so I can check the number of pages I’ve captured.
and use the catalogue reference numbers as milestones to delineate my sessions.
But, as my collection grew larger, I noted that the index became less reliable. Pages I'd captured seemed to be missing...
I perform a direct search to confirm that the 'missing' URL has been captured.
I realise that when building an archive of this scale, my quality-assurance and capture need to be simultaneous.
I change the colour of the URLs I've captured, to keep track of where I am.
After capturing a full notebook, I routinely perform a search of the URL index to retrieve and review the full sequence of pages.
I mark my progress with each back-cover scan.
One of the fundamental tenets of archiving using the Webrecorder tool suite, is the principle of performativity.
Any actions you want to be performable in the archive, must be performed during capture.
There is a necessary mirroring between the functionality activated by the archivist, and the functionality an end-user of the archive will be able to access.
And, the explorable, spatial extent of the archive is delineated by the archivist’s definition of its boundary during capture.
Something that I haven’t needed to articulate within this framing until now, but which has come to the fore in this case,
is that the intuitive browsing potential within the archive, is also dependent upon an archivist's acute attention to all the possible pathways a user might take through it.
I hadn't expected that the direction from which a user navigates their pathway through a website would matter. But in this case, I found that it did.
The Carmichael Watson catalogue records are richly appended with tags. These enable researchers to cross-reference the people, places and ideas which link different parts of the collections together. Each of these tags represents a link to an entry in one of several detailed indexes.
Each index entry offers a portal of associated notebooks and objects that researchers can browse.
and through which they can choose to navigate back into the collections.
The unusual thing I noticed was that these paths which lead from tags or indexes back to the same object in the collection, don't resolve to the same URL. Rather, they each retain a reference of their journey.
To capture these indexes so that they are functional as discovery tools in the archive, would involve plotting out and performing all possible paths of navigation. The scale and complexity of achieving this manually is unfeasible.
I've initiated several tests using Webrecorder's automated Browsertrix Crawler, but with limited success so far.
I watch the number of pages the crawler has identified grow, in inverse proportion to its progress through the task. Meanwhile, the VPN connection periodically shuts off and timeout errors tick up.
If we instead take the decision to omit the URLs that retain their complex pathways between indexes, tags and objects, then we will need to annotate the archive – explaining to future users that some routes through the collections are accessible, and others are not.
ArchiveWeb doesn't currently offer the facility for annotation that its previous iteration did. I frequently used List descriptions in the (now phased-out) Webrecorder Desktop app to signpost problems and give guidance. I've fed-back to ArchiveWeb's developers, but this time we'll need to devise an alternative method.
For now, the work continues!
Thank you.

Sara Day Thomson & Anisa Hawes

@sdaythomson
@AnisaHawes

https://webrecorder.net/
https://archiveweb.page/