Vanished:

Preserving the Carmichael Watson Project Website Offline Using Webrecorder

|  |  |
| --- | --- |
| **Sara Day Thomson** | **Anisa Hawes** |
| *University of Edinburgh* *UK**Sara.Thomson@ed.ac.uk**0000-0002-3896-3414* | *Freelance Researcher & Web Archivist**UK**anisa.hawes@icloud.com*  |
|  |  |

**Abstract – In 2021, the Carmichael Watson project website — a highly valued resource of Gaelic culture, culminating in an investment of over £750,000 — faced imminent termination. This case study details how this project website, only online from 2013 until 2018, came to imminent risk of permanent loss. It then presents the strategy undertaken to transform it into a more sustainable format through web archiving and to revive its public accessibility.**

**Keywords – Web Archiving, Webrecorder, Technology Obsolescence, Disaster Recovery, Heritage Stewardship**

**Conference Topics – Community; Resilience.**

# Introduction

Due to security issues with underlying infrastructure, in 2021, the Carmichael Watson project website [1] — a highly valued resource of Gaelic culture, culminating in an investment of over £750,000 in external grant funding — faced imminent termination. The Digital Archivist at the University of Edinburgh, in collaboration with the Digital Library Development team and freelance web archivist Anisa Hawes, intervened to capture and preserve this resource using Webrecorder’s ArchiveWeb.page [2]. Ultimately, the University of Edinburgh aims to embed good practice in digital preservation at the outset of creating digital resources, but like every steward of digital data or heritage resources, the university is subject to practical constraints. This case study details how this project website, only online from 2013 until 2018, came to imminent risk of permanent loss. It then presents the strategy undertaken to transform it into a more sustainable format through web archiving and to revive its public accessibility.

# Background

The Carmichael Watson Project, based on archives held by the Centre for Research Collections (CRC) at the University of Edinburgh [3], revolves around the papers of the pioneering folklorist Alexander Carmichael (1832-1912) and brings to life customs, stories, songs, and beliefs from the Gaelic-speaking areas of Scotland. It offers fundamental insights into the creation of Carmichael's greatest work *Carmina Gadelica* [4] but also supports interdisciplinary cooperation between local and scholarly communities for collaborative research in history, theology, literary criticism, philology, placenames, archaeology, botany, and environmental studies [5]. Through cataloging, indexing, transcribing, translating, digitizing, and conserving, this project opened up and made accessible this important collection to the academic and broader community. In particular, this web resource was developed to make erratic and multilingual notebook entries more readable and accessible, with transcriptions and detailed notebook entry-level descriptions. The project team also created EAC (Encoded Archival Context) records for many of the individuals from whom Carmichael collected material, giving prominence to people regarded as ordinary folk and marginalized groups. Linking to APIs to give geographical context to each individual item within a notebook provided increased usability and engagement with the resource, an essential part of key research interests.

# A Series of Unfortunate Events

The project website has been unavailable to the public since 2018 when the original stewards turned it off due to security concerns. The original URL, <http://www.carmichaelwatson.lib.ed.ac.uk>, now points to a holding page but the actual resource is not accessible to the public. However, the Library Development team maintained access for internal users via Virtual Machine (VM) with restricted access. Unfortunately, this VM for internal access was on RedHat 5, a technology that had reached its end of life by 2021. As a result, IT Infrastructure (ITI) notified the Development team to terminate all services running on it. In addition to the security issue posed by running end-of-life technology, ITI had been paying an additional license fee for the VMs, so maintaining the resource had been incurring an on-going cost, despite being unavailable to the public.

The archives of the Carmichael Watson collections remain available through other dispersed access systems. When the website was taken offline in 2018, the project team ensured access to digitized versions of the notebooks and full archival descriptions through ArchivesSpace (the university’s archives discovery system) [6]. However, the project team envisioned and invested in the website to make the materials more accessible and usable by researchers and members of the communities represented in the collections. The content and functionality underpinning that wider accessibility exist only in the web-based resource, including the TEI [7], the geotools, and the handwriting guide.

The underlying code for the project website was saved, but it would require substantial funding to build a new website from scratch or integrate this resource into another, newer platform. Resource for development at this scale — which would essentially repeat a project already substantially funded in the recent past — was simply not available. Furthermore, re-developing the resource on a new platform would inevitably lead to the same situation: the need to redevelop the resource when its underlying infrastructure inevitably becomes obsolete.

# Web Archiving … Offline

The first step to developing a strategy for archiving this project website was to check for a copy in a national web archive. The UK Web Archive [8], as a national program with a remit to capture UK websites, provides a first port of call when searching for legacy web resources from the University of Edinburgh. The university itself does not have a web archiving program, but since 2020 has been collaborating with the UK Web Archive [9] to better look after its valuable (and vast) web estate. The UK Web Archive had, in fact, crawled the original URL on 17 occasions since 2013, noting on the backend that the Live Site Status was ‘Vanished’. Unfortunately, under closer examination, none of the successful captures contained the original formatting and many pages were missing images. The ‘QA Status’ in the UK Web Archive’s system W3ACT [10] was listed as ‘none’, which perhaps provides some explanation for why the crawls, even later ones, lacked the underlying style sheets and images (a problem affecting many of the university’s web pages in the UKWA). Neither the Wayback Machine run by the Internet Archive [11] nor any other copies discovered through querying Memento Time Travel [12] held more than the top-level pages.

In the absence of a complete archived copy, the next option was to test if it could be archived offline — from the internal VM — using manual web archiving tools. Conifer [13] was able to capture a high-quality copy of a selection of pages, including authentic formatting and images. The Webrecorder tool used by the Conifer service [14] was built to preserve websites, even complex and interactive websites, to a high degree of fidelity (or accuracy). However, a manual tool like Webrecorder requires the user to click every link and activate every function in order to capture content. It would take a full-time archivist several weeks, if not months, to archive the Carmichael Watson project website in its entirety using this approach.

There are, however, important benefits to using a Webrecorder-driven approach to preserving this resource, namely the longevity of the output format and resulting persistent access. Using Webrecorder tools, the resource can be captured and transformed into a warc or wacz file, ingested into a digital preservation system, and accessed with appropriate user guidance through local systems. Perhaps most importantly, the tool is incredibly accessible and enabled the team, with no institutional web archiving infrastructure, to start capturing the resource immediately.

# Resurrecting Carmichael Watson with ArchiveWeb.page

In collaboration with Anisa Hawes [15], the University of Edinburgh opted to pursue the preservation of the Carmichael Watson project website using up-to-date Webrecorder tool ArchiveWeb.page, released in January 2021. ArchiveWeb.page allows users to systematically capture a web resource, mainly through a Chrome extension which enables capture through normal interactions with the browser.

The active capture of the website was only one part of the rescue workflow. To ensure no pages were overlooked, the website was first mapped and scoped to understand the boundaries and extent of the resulting archival resource. Due to the restricted access to the resource (via VPN), Anisa Hawes mapped the resource manually, documenting the URLs of each of the many navigation pathways in spreadsheets. Once capture commenced, the index of captured URLs in Archiveweb.page had to be cross-referenced to check all relevant pages had been captured and the quality of each capture (text, formatting, and functionality) assured. In order to support access, the archived resource has been annotated and documented. Annotation is particularly important to explain to target end users where the archived website ‘ends’ and when errors or missing content derive from the original and when from failures of the capture tool.

Capture commenced in summer 2021, though from the outset a number of challenges posed serious impediments to this planned workflow. The first and perhaps most prohibitive challenge was the design of the website itself. Rather than individual pages resolving to a persistent URL, individual pages had been duplicated in multiple user navigation pathways, including multiple access points and browsable indexes. Therefore, the same identical page exists at six, seven, eight, or more unique URLs culminating in nearly 20,000 URLs. This challenge highlighted the importance of resource mapping to establish a realistic scope and timescale. Ultimately, the team decided it was simply not feasible, in the available time, to capture the entire website manually. This decision drastically increased the importance of annotation and documentation to clearly communicate which pathways and access points are archived and which are excluded.

In addition to the time-consuming task of manually capturing so many pages, technology constraints also created barriers. As mentioned, the obsolete infrastructure underpinning the VM created security concerns, requiring VPN restrictions. While providing a layer of security, the VPN also slowed down the power and speed of ArchiveWeb.page and the machine used for capture. The VPN also created a hindrance to experimenting with Browsertrix [16], a Webrecorder tool with automation to support capture. Though in theory Browsertrix could have made it feasible to capture the website in full, the VPN caused the tool to time out before crawls could be completed.

The functionality of ArchiveWeb.page, only released months before capture commenced, presented challenges as well. As the number of captured URLs accumulated, the ArchiveWeb.page index stopped displaying them all, making it difficult to check that all pages from a capture session had been successfully archived. Though the deprecated predecessor of ArchiveWeb.page (Webrecorder desktop) had functionality for adding metadata and annotation directly into the archived resource, this early version of Archiveweb.page does not, though a request for improved curatorial functionality has been submitted to Webrecorder. Therefore, annotation and descriptive metadata has been created separately, which will need to be maintained over time through the CRC’s digital preservation program. A fuller presentation of the methodology and challenges of capturing an offline resource this way can be found on the Digital Preservation Coalition event page for the December 2021 Web Archiving & Preservation Working Group [17].

Though these technology challenges created hurdles for archiving the Carmichael Watson project website, the strategy for access looks much more promising. Based on recommendations from Anisa Hawes, the team will be hosting the archived resource on a local server and providing a link from the CRC’s discovery record in ArchivesSpace. This approach, developed by Stanford University Libraries [18], will allow users to click on a link that takes them to the interface for the Webrecorder companion tool for replay — ReplayWeb.page [19] — to view and interact with the archived website through their browser.

The capture phase of the project has been completed and the VM turned off. The archive of the Carmichael Watson project website created by Anisa Hawes in Archiveweb.page is now the only, golden copy of the resource. Unfortunately, the large size of the warc file containing the archive exceeds the export limit of Archiveweb.page and, at the time of writing, an alternative export method is being explored with the developers. Arguably, the most difficult and important task remains: to clearly and effectively communicate this archived version of a beloved resource to its target users, many of whom will never have heard of web archiving.

# Lessons Learned

So far the team has learned some important lessons. The first lesson, of course, is to act sooner, acknowledging that the way resources are built or managed is not always in the control of the practitioner or team responsible for archiving. If web archiving had commenced before the resource came off the public web, automated tools like Browsertrix might have made the process much faster and cheaper. The CRC may have even been able to work with the UK Web Archive to improve automated capture, which would have secured a copy of the resource in one of the world’s most well-supported and well-known web archives.

It has become clear that in order to be responsible custodians of its public record on the web and its web-based collections, the university needs to actively engage with the UK Web Archive and build local capacity for web archiving. Immediate investment is required in the staff resource to QA the University of Edinburgh Web Estate in the UK Web Archive and build relationships with content creators and stakeholders (a business case is currently pending). These relationships would enable, over time, the transformation of development practices for web-based resources. These relationships could also enable archiving from the point of creation in a way that supports long-term requirements, whether through the UK Web Archive or alternative approaches like Webrecorder.

In summary, the University of Edinburgh, as perhaps with other organizations in the HE sector and others, continues to struggle to embed digital preservation early in the lifecycle of digital materials. There is a lot of work to do to raise awareness of and build capacity for digital preservation across the university. In the meantime, it will be a rewarding victory to make the archived Carmichael Watson website available to the researchers, teachers, and members of the community who will now be able to use it for years to come.

# REFERENCES

1. Original URL for the Carmichael Watson project website (2013-2018). <http://www.carmichaelwatson.lib.ed.ac.uk/cwatson>
2. Webrecorder Tools, ArchiveWeb.page. <https://webrecorder.net/tools#archivewebpage>
3. Centre for Research Collections, University of Edinburgh. <https://www.ed.ac.uk/information-services/library-museum-gallery/cultural-heritage-collections/crc>
4. Carmichael, Alexander. Carmina Gadelica [or, Ortha Nan Gàidheal]: Hymns and Incantations with Illustrative Notes on Words, Rites, and Customs, Dying and Obsolete, Translated into English by Alexander Carmichael. Edinburgh: Floris Books, 2006.
5. List of publications referencing the Carmichael Watson archives. <https://www.ed.ac.uk/information-services/library-museum-gallery/crc/research-resources/gaelic/carmichael-watson/publications>
6. ArchivesSpace. https://archivesspace.org/
7. Text Encoding Initiative. <https://tei-c.org/>
8. UK Web Archive. <https://www.webarchive.org.uk/>
9. UK Web Archive blog. ‘University of Edinburgh’s Collecting Covid-19 Initiative — Collaborative Collection Building with the UK Web Archive’ (12 March 2021), <https://blogs.bl.uk/webarchive/2021/03/university-of-edinburghs-collecting-covid-19-initiative-collaborative-collection-building-with-the-u.html>
10. UK Web Archive, W3ACT. <https://github.com/ukwa/w3act/wiki/W3ACT-User-Guide>
11. Internet Archive, Wayback Machine. <https://archive.org/>
12. Memento Time Travel. <http://timetravel.mementoweb.org/>
13. Conifer. <https://conifer.rhizome.org/_faq>
14. Webrecorder Desktop. <https://github.com/webrecorder/webrecorder-desktop>
15. Anisa Hawes. <https://anisahawes.github.io/about/>
16. Browsertrix Crawler. <https://webrecorder.net/tools#browsertrix>
17. Web Archiving & Preservation Working Group, General Meeting December 2021. <https://www.dpconline.org/events/past-events/wapwg-meeting-dec2021>
18. Stanford Libraries, ‘Archiving Instagram posts’ (2021-09-09), <https://library.stanford.edu/blogs/stanford-libraries-blog/2021/09/archiving-instagram-posts>
19. Replayweb.page. <https://webrecorder.net/tools#replaywebpage>