National Records of Scotland Web Continuity
Service Model

Introduction

1. The National Records of Scotland (NRS) Web Continuity Service (the Service) enables us to archive websites as part of our statutory collecting responsibilities.[1]

2. Users can freely access all content archived by the Service within the NRS Web Archive. Further information on the context and collection policy of the NRS Web Archive can be found in our Selection Policy.

3. This document describes the technical processes on which the Service is based, along with the service model within which it operates. The wider opportunities and benefits to stakeholders that the Service will provide are also highlighted.

Web Archiving

4. The Service is based upon the technical process of web archiving. Web archiving involves capturing content that has been made available via the Web, permanently preserving this as archived content, and making it accessible to users.

5. The Service archives websites using web crawler software. Our crawler visits a selected website, or 'seed', and explores the seed via its hyperlinks, copying content as it goes. Captured content is quality assured to a defined standard, and then preserved within an ISO 28500:2009 WARC (Web ARChive) file ('WARC file'). The WARC file constitutes an archival record of the website as captured at that point in time.

6. Archived content differs from a back-up of a website, given that good quality metadata is captured simultaneously, enabling a high percentage of the 'look and feel' and behaviour functionality of the site to be rendered to the user in a web browser. When viewed in a browser, archived content is clearly identifiable to the user via a banner, to avoid confusion with live websites.

7. Archiving websites with crawler software brings technical challenges, most notably in capturing dynamic content such as Javascript, client-side script, database-driven content etc. Nevertheless, quality assurance (QA) can go some way towards mitigating these risks and NRS's service has a standardised process for this. Furthermore, website owners should embed good design principles into their site(s) to help control the structure, content and 'archivability' of a site, so that preservation and long-term access begins at source: please contact the NRS Web Continuity Service for further details.

Web Continuity

8. Web Continuity builds on this web archiving working model by providing website owners with the option to enable a redirection service on their website(s). This redirection service will take users seamlessly from missing pages on these live site(s) to the NRS Web Archive, where a search for an archived version of the missing page will be made, and served if found. This means users will see many fewer '404 page not found' error messages when visiting these live sites.

---

[1] Please see the Selection Policy for the National Records of Scotland Web Continuity Service for details on NRS's statutory and strategic collecting scope.

Preserving the past | Recording the present | Informing the future

9. Archived content will be clearly identifiable to the user via a page banner, and the URL of archived content will be rewritten to include http://webarchive.nrscotland.gov.uk/, followed by the date and time of capture and the particular URL in question e.g. http://webarchive.nrscotland.gov.uk/201612011714/http://www.gov.scot. Preserving the chain of online information through this mechanism can help improve user experience through continued access, and in turn support the website owner's public accountability and transparency.

10. Website owners who wish to avail themselves of the Web Continuity redirection capability must install a small piece of code on their own web server to enable this service: contact the NRS Web Continuity team for further information.

NRS Web Continuity Service

11. The NRS Web Continuity Service will operate as a free service to owners (and, in turn, users) of all websites selected for capture by NRS. The contents of the NRS Web Archive will be freely available online to all users.

12. NRS has entered into a contract with Internet Memory Research (IMR) to carry out the technical aspects of the Web Continuity Service. IMR's technical service enables NRS to select and schedule capture of websites, quality assure captured content, and make archived content publicly available online. As a matter of principle, NRS will use the QA process to ensure that each website is archived to as high a degree of quality and completeness as possible, before content is made publicly available. WARC files containing archived content will be securely transferred to NRS for permanent preservation.

13. Information on the scope of the NRS Web Archive, our selection criteria, and the frequency at which we will capture selected sites can be found in the Selection Policy for the NRS Web Continuity Service.

14. The Web Continuity Service will be staffed by the NRS Digital Records Unit, who will support the collecting branches across NRS. The roles of this team are outlined in Appendix 1.

15. A working group will meet quarterly to discuss issues including stakeholder engagement, website selection, and service delivery. The makeup of this working group is described in Appendix 1 of the Selection Policy.

16. NRS will always gain permission from website owners to capture content ahead of time. This will ensure a better chance of success for capture, help inform timing for capture, and permit NRS to accrue valuable collections' knowledge.

17. As part of our Service, NRS will advise website owners on the content, structure, and management of their site to maximise the benefits derived from this Service.

18. In turn, NRS requires website owners to inform us of major changes to their website(s) e.g. major content change/restructure; creation or closure of a micro-site for a purpose or event; change of URL etc. All such information will enable NRS to make more informed decisions on selection and scheduling of capture. Ideally such information should be communicated to NRS at least **2 months** prior to the change(s) taking place, so that we can potentially capture the site(s) before this occurs.

Capture, QA and Access

19. The web archiving process takes approximately **8 weeks** to complete, from scheduling of the capture through to QA, before the archived content is made publicly available. This workflow is illustrated in Appendix 2. During that time it is important that the live website being archived remains accessible to support our QA process.

20. How long our crawler spends on a website will be depend upon the size of the site in question, but our upper time limit for this is 4 weeks. We aim to capture content in a manner that minimises the effect this has on live sites.

21. In order to capture a complete copy of a website, we will discuss with website owners whether their site contains any robots.txt protocol, which may block our crawler from entering certain parts of the site.

22. Each time NRS archives a website, our crawler will identify itself to the website manager via a specific user-agent string, and provide contact details.

23. During QA, a host of tests and mechanisms are used to capture and replicate the complete content and experience of the live website as fully as possible.

24. Once QA is complete, the archived content, or 'instance' is made available via a permanent URL on the NRS Web Archive. The web archive can be accessed via the NRS website, via Memento, or in the case of broken links on the owners' websites, via the Web Continuity redirection service.

25. NRS aims to include a description of each archived website in our online catalogue.

26. Guidance for users on searching and using the NRS Web Archive can be found on our website.

Benefits

27. The NRS Web Continuity Service can potentially offer a host of benefits to website owners and the wider user community. Some key examples include:
    - preserving a representative record of how NRS's stakeholders interact with the citizen online, from the context of the user experience;
    - helping maintain access to official online information, and so assist the Scottish Government in its stated aim for openness and accountability to the citizen, as well as improve business efficiency;
    - supporting organisations who deposit archival records at NRS with their own recordkeeping responsibilities;
    - affording website owners the opportunity to rationalise their old web content and enhance the functionality and structure of their live sites by archiving historic content;
    - creating a new online public resource for the study of Scotland, providing access to content that may disappear from the live web.

    Website owners are encouraged to contact the Web Continuity Service team to discuss how our Service may assist them in these aims.

28. NRS will adopt the principles of the Digital First Service Standard for all Web Continuity Service operations.

29. Initially, all questions relating to the NRS Web Continuity Service should be addressed to the website owner's responsible NRS Client Manager or key contact. For technical queries or issues, the Web Continuity Service can be contacted directly on: digital_records@nrscotland.gov.uk.

30. This document will be reviewed every two years.

February 2017

National Records of Scotland

Appendix 1: The NRS Web Continuity Service Roles and Responsibilities

NRS Web Continuity Service- Our Roles and Responsibilities

Work with NRS staff, website owners and external stakeholders to select and manage websites for capture

Liaise on access with users and maintain catalogue of archived websites

Advise website owners on best practice in site design and web continuity redirection

Work with contracted service provider on technical process of capture and monitor service performance

Appendix 2: The Web Continuity Workflow Diagram

Stakeholders – Website owners, archive colleagues

The User Community – website owners, citizens, researchers

1. SELECTION NRS Web Continuity Service creates list of selected websites to be archived, in collaboration with stakeholders

4. ACCESS Archived content is made available online and can be accessed via the NRS website, or via web continuity redirection. NRS creates a catalogue entry for each website archived.

2. INSTRUCTION Using our service provider's tools, NRS arranges for the selected websites to captured on agreed dates. Our service provider then launches the web archive technical process

3. TECHNICAL PROCESS

Crawler visits website(s) and copies content from hyperlinks

Once capture is completed, service provider undertakes minimum QA and passes to NRS

NRS tests/QA's captured content and seeks extra remedial actions if required

Once QA is complete, the content is then accepted, or rejected to seek further remedial action

Stages 2 to 4 takes approximately 8 weeks in total