Preserving the content and the network:
An innovative approach to web archiving

Amanda Spencer

The National Archives
Kew
Richmond TW9 4DU
Amanda.Spencer@nationalarchives.gov.uk

Abstract

Government's use of the Web has required new approaches to Web resource preservation. The National Archives' approach draws on its experience of Web Archiving, as well as expertise in the live Web arena. By harnessing these two elements, The National Archives hopes to deliver a truly innovative user-centric service predicated on preserving the content of websites as well as utilizing the value of the Web as a network.

Introduction

The proliferation of websites in the workplace has touched every sector, and Government has been no exception. Since the early 1990's the Government has been using websites to present information: official reports, papers, transcripts of speeches, guidance, announcements, press statements, regulations and advice. The benefits offered by these new technologies means that services are increasingly being delivered via electronic means and through digital channels.

The evolution of websites, coupled with the size and ever-changing nature of Government, mean that these sites are vulnerable to technological problems, such as documents 'falling off' sites, or links being broken between resources.

The prevalence of broken Web links impacts negatively on the reputation of government because it is perceived that government is managing it information poorly; a frustrating user experience on line also has the potential to reduce public confidence, and parliamentary scrutiny of government is impaired by its inability to refer to key government documents.

This state of affairs was brought into sharp relief by Cabinet Ministers looking for documents on government websites, only to find that these documents had been moved or removed. On 19 April 2007 the leader of the House of Commons wrote to the incumbent Cabinet Office Minister expressing concerns over the issue of documents and information disappearing from websites, concerns supported by a sample survey of URLs (Uniform Resource Locators) cited in Hansard in response to parliamentary questions[1]. It was noted that such links in the parliamentary record often failed to resolve.

As a consequence the Archiving Digital Assets and Link Management working group was formed in May 2007. The working group was comprised of members drawn from The National Archives of the UK, the British Library, Information Services at the House of Commons (formerly the Parliamentary Library), the Parliamentary Archives, and the policy unit at Central Office of Information.

Research

The working group identified a number of interrelated issues, supported by a number of pieces of applied research, which were all contributory factors to the loss of significant official information over time.

The current situation

The working group identified that the government has effective strategies in place for ensuring that all information laid before Parliament is published appropriately and flows through to The British Library, for long-term access and preservation. In the UK the British

[1] Preliminary research conducted by the then House of Commons Library (now Information Services at the House of Commons)

Library is custodian of a body of official government publications/information which has been built up over past centuries, for historical access and long-term preservation of government activity. Researchers and historians expect such long-term access to official information through the preservation work of British Library. These government strategies rely on the existence of a printed rendition and a degree of centralisation and control of these official publishing arrangements under the auspices of Her Majesty's Stationery Office (HMSO), which operates from within The National Archives. Where material falls outside such practices, there are no agreed procedures for ensuring that online information is preserved over historical periods and made accessible. At present e-government information is received and preserved by the British Library on a voluntary basis. An understanding of how Government handles this body of information was required.

Broken links

The first issue, and the primary driver for the establishment of the working group, as described above, was that of a breakdown in online access to information through links, highlighted by the preliminary findings of the Information Services at the House of Commons and confirmed by the research on Hansard conducted by The National Archives. A longitudinal survey of URLs cited in response to Parliamentary Questions and recorded in Hansard revealed that 60% of links in Hansard, cited between 1997-2006, had since broken, resulting in '404 Page Not Found' errors, suggesting that many government departments do not consider the issue of long-term access to government information. [1] And yet ministers and other government officials assume that the information situated at any given URLs cited in response to a Parliamentary Question will remain available in perpetuity.

Government's use of the Web

This issue is compounded by the fact that much of this information is increasingly only available electronically [e.g.s], not in print, and even then is not always filed in electronic document and records management systems (EDRMS) making the integrity of Web links crucial to the business of government. Some government departments,

tend to post documents and information on websites in HTML, rather than PDF or Word, making it more difficult to extract and archive the stand-alone documents from the websites. Additionally, as our understanding of the potential use of the web has developed, there are powerful arguments in favour of using HTML instead of document formats such as PDF on the web. In terms of data mashing and the semantic web, HTML can yield far greater benefit than PDF, which can lock-in information and prevent its reuse for other purposes. Further, some Web-based database-driven content is only available via a Website's search interface. As a consequence any solution to the problems identified needed to take account of the changing nature of and potential uses of the Web.

Website Rationalisation

A further area of concern related to The Transformational Government Website Rationalisation programme. The Website Rationalisation programme, aimed at streamlining the Government's estimated 2,500 Websites[2], began in 2007 and is due to complete in 2011, and is concerned with delivering a better web user experience for the citizen seeking to access government information. Much citizen-focussed content will be converged onto the Directgov supersite. Other content may move to Departmental corporate sites and many websites will close in order to reduce government website proliferation. Although the issue of broken links is not a problem initiated or caused by the Website Rationalisation programme, there is a concern that it may exacerbate an already poor situation. It was agreed that there needed to be a policy and process for archiving and also for link management to ensure that information remains findable.

Given this state of affairs the working group concluded that all web-based information should be treated as an important contribution to the body of government information, and in particular that all online information that has been cited should remain available and accessible in its original form. This idea reflects an acknowledgement that the web has changed user behaviour in the way information is accessed.

The Options

The group explored a number of different options including improvements to existing practices, which would

---

[1]Unpublished research conducted by John Sheridan at the Office of Public Sector Information (part of The National Archives since 2005) on Hansard which revealed that 60% of Web links cited in Hansard 1997-2006 are now broken, suggesting that many government departments do not consider the issue of long-term access to government information.

---

[2] Government on the Internet: Progress in delivering information and services online, 29 April 2008, http://www.publications.parliament.uk/pa/cm200708/cmselect/cmpubacc/143/143.pdf

involve re-issuing guidance on legal deposit of e-only publications, and using Digital Object Identifiers (DOIs). Both these options were considered to have certain drawbacks: guidance alone was unlikely to have significant impact, and if it did improve voluntary deposit, it was not considered to be very user- nor web-centric.

Using DOIs was initially a very popular idea, as it has been successfully implemented in the scientific publishing world. However, it was acknowledged that this system requires a high degree of centralised management and control on an ongoing basis, which was recognised as being too complex to achieve within a UK central government context. In working through these options the working group arrived at a solution which encompassed both elements: further guidance on managing websites and web content, and the principles behind use of DOIs and the idea of identifiers. With the existing web archiving programme at The National Archives we knew that it was possible to capture website content, and with the Website Rationalisation programme underway, we had already made a commitment to capturing more content, as websites closed or content was moved. Each piece of information already has an identifier in its URL, and in the web archive the same piece of information has a predictable archive URL, based on the original reference. For example, the most recent available copy of http://www.mydepartment.gov.uk/page1.html becomes http://webarchive.nationalarchives.gov.uk/*/http://www.mydepartment.gov.uk/page1.html.We just needed a way of matching an original URL with the archive version in the web archive. Because the European Archive identifiers are predictable   we concluded that using a redirection component could enable us to run a DOI-type scheme using the web archive.

Following informal consultation of stakeholders, The National Archives (now leading the Knowledge and Information Function across Government) assumed responsibility for delivering the solution devised by members of the working group. This paper will outline the findings and outputs from this significant piece of work.

The National Archives and the Web Continuity Solution

The scale of the programme, the issue of trying to preserve both the content and the network, and the need for content capture to be as comprehensive as possible, required a truly innovative approach.

The project necessitated new thinking in web archiving to address a number of different, difficult elements:

1. How to capture significant levels of important Government information from possibly thousands of distributed, heterogeneous websites (including websites closing as part of the Website Rationalisation Programme);

2. Methods to ensure not only a greater capture of content, but also increase exposure of this content to the web harvesting crawler, from sites that vary hugely in nature;
3. Ensuring that links persist to ensure that users will always find the last available version of the page, whether it is on a live site, or in the web archive.

The extensive scope of the project has required a mechanism for auditing the Government web estate, for identifying and controlling the number of Government websites in operation, and for seeding the harvesting process. As a consequence new processes and tools have been developed. A central SQL Server database has been built for use as a registry of all UK Central Government websites. Originally intended solely as a means of seeding the harvesting process, discussions with other government stakeholders identified a need for a single source of up-to-date information about the live government web-estate, details of all websites, current and inactive, any schedules for content closure or convergence as part of the Website Rationalisation programme, and evidence of compliance with government web standards guidelines (such as the accessibility standard). The database will be available to all website managers in central government and the responsibility rests with them to keep their information current. Appropriate access controls have been applied so that website managers can only edit their own departmental records. In respect of Website Rationalisation, only the Transformational Government team at Central Office of Information, with responsibility for new government domain registration and as the Data Quality Officer for reporting on progress of the programme, will have 'update' access to the scheduled website closure and convergence dates.

The archiving of government websites is to be carried out using the most popular method of capture for large-scale programmes, remote harvesting using a Heritrix web crawler. The National Archives web archiving is carried out under contract to the European Archive, and the Web Continuity Project has meant a significant increase in the number of websites captured, moving from a selective archiving programme to a comprehensive programme involving all websites of central government departments, agencies and Non-Departmental Public Bodies (NDPBs). The research conducted by the Digital Assets working group which highlighted that often websites are the only source for particular documents, has required that the archiving programme recognize that the partial archiving of websites, often a result of the limitations of current remote harvesting technology, is not an adequate solution. As a consequence The National Archives has explored the possibility of using the XML sitemap protocol[1], to ensure

---

[1] http://www.sitemaps.org/protocol

that capture of the Government web estate is comprehensive.

The widespread adoption of XML sitemaps by government departments will have other associated benefits, most notably relating to web resource discovery using search engines.

Citizens increasingly use search engines to look for information hosted on websites on a wide variety of subjects. Government information forms a part of the enormous mass of information available, but if it is not exposed somehow to search engines indexing, it can be 'buried' among the mass of other, less relevant information, or worse, remain completely undetected, and therefore, unable to reach its intended audience. Search engine providers build indexes of available (i.e. exposed or linked to) information available on the World Wide Web. They are unable to include unlinked to or 'hidden' content (the so-called 'hidden' or 'deep' Web). Hidden content not only includes databases which can only be interrogated by queries, but also content which is essentially generated 'dynamically' or 'on the fly.' XML Sitemaps enable website owners to expose hidden content if appropriate, and moreover, allow website owners to have better control over what parts of their website they expose to search engines.

Software used to 'crawl' websites remotely in order to take archival snapshots operates in a similar way to search engine software. This type of crawling is the most efficient, robust and therefore widely used in large-scale crawling programmes. However, fundamentally, it can only crawl (and capture for archiving) content which is linked to, or exposed in some way. XML Sitemaps also enable website owners to expose hidden content if appropriate, to web archiving crawlers

The pan-government search group has recognised that more action is needed to ensure that current government information on websites is findable for citizens. It has also been recently recognised that action is needed to ensure that continued access to information over longer periods of time is also required. The National Archives through the Web Continuity project is developing a solution to the latter, which involves more comprehensive archiving of websites within the central government domain, and a method of links persistence so ensure that instances of 'broken' links to government information (acutely represented by 'broken' links in Hansard) are reduced.

The close relationship between searching for live context and capturing greater archival content and the recognition that both situations can be greatly improved through the adoption across government organisations of XML sitemaps, has meant that The National Archives has assumed responsibility for the Sitemap Implementation Plan across government.

Some people operating in the government Website arena already understand and use sitemaps, some know little or nothing about them, while others may understand what they are, but have little knowledge or experience of how to set about using them. Given that current knowledge and understanding of sitemaps and their practical uses is varied across government, The National Archives approach to sitemaps implementation is three-fold and is detailed below.

Online Instruction Packages (Breezos)

These packages will raise awareness of sitemaps and are designed to reach a wide non-technical audience. They have been written by a Third Party provider and comprise three separate modules:
- Introduction – why sitemaps are important and why you should have one
- Detail – what a sitemaps is
- Practical – How you can create a sitemap

The practical module will be complemented by research, testing and guidance organised by The National Archives (outlined in the following section).

Software

The National Archives has contracted a third party to evaluate a survey of the sitemap generation software market, against a set of pre-defined minimum functional requirements. Software vendors have been approached to validate their software against a rigorous set of technical and assurance-related claims, using the government-endorsed CESG Claims Tested (CCT) Mark scheme. It was intended that this exercise would be repeated every two years to ensure that the market-place evaluations remain timely, that new releases of software would be validated appropriately, and that only software which had been successfully validated using the scheme would be recommended to government organisations. However, there have been certain limitations with this aspect of the project, the primary issue is one of supplier incentivisation. The Claims tested scheme relies on suppliers financing the claims testing process, which costs around £20,000 per product tested. Few third party software suppliers seem interested in investing such a significant sum when the returns are likely to be relatively small. These products seldom cost more than £30.

Guidance

Guidance will be made available, which provides all the information necessary for installing sitemaps generation software, creating a sitemap, and deciding what to include or exclude from a sitemap.

Practical Implementation

The National Archives is working with its web archiving partner, the European Archive to ensure that where sitemaps are deployed, maximum effectiveness in capture

of content is achieved. After some discussions it was considered inappropriate to use the organisation's sitemaps as the primary mechanism for seeding the crawl. European Archive were concerned that if the sitemaps were out of date, or incomplete then the quality of the archived website instance would be impaired. Instead, agreement was reached that the sitemaps would be used to complement and enhance the content captured via the initial gather by the Heritrix crawler.

The relatively large-scale nature of the programme also favoured an automated approach to both the seeding of the crawls and the capture of preservation copies of the archived websites, which The National Archives receives under contract from the European Archive. As a consequence two interconnected workflows have been designed: one for the harvesting lifecycle which drives the crawling process and ends with the production of publicly accessible copies of the archived websites, made available by European Archive at a TNA IP address, and another which begins with the preservation copies being made available for ingest into The National Archives Digital Object Store (DOS)[1].

Automated harvesting begins with series level creation for cataloguing purposes, which provides another means of identifying websites and instances of websites within our wider collection, and enables websites to be situated within the context of the other digital and paper records of the creating department. The cataloguing information is captured in the website database, and is passed through to the European Archive so that it can be provided, alongside other metadata at ingest. The step-by-step harvesting process is dependant on a series of messages exchanged via FTP between the European Archive and TNA. Once this process is complete, a further message triggers the start of the ingest workflow and allows for the ingestion of multiple preservation copies in a single process step. Apart from the obvious benefits of having preservation copies stored separately from the presentation copies and their back-up versions, ingest into the DOS allows for active preservation of websites alone side the active preservation and migration of other digital records at The National Archives.

The final element of the new web archiving process concerns the use by government organisations of a redirection software component. Installation of this component will ensure the persistence of Web links and creates a different purpose for the web archive. The components, to be supplied by The National Archives,

_____

Seamless Flow Newsletter Issue 1,
http://www.nationalarchives.gov.uk/documents/newsletter_issue1.pdf

following configuration by The Stationery Office (TSO), utilise open-source software and have been designed to work with Microsoft Internet Information Server versions 5 and 6, and Apache versions 1.3 and 2.0, which are the platforms most commonly used in UK Central Government[2]. The IIS component is produced by Ionics www.codeplex.com/IIRF and the Apache component is the mod-rewrite module: http://httpd.apache.org/docs/1.3/mod/mod_rewrite.html

The component works by redirecting the user to the UK Government Web Archive in the event that the page could not be found (a 404 error). It does not replace any existing redirections on the live website. The component is also installed on the web archive site. Here its role is to rewrite the URL for the original department website, if it is not found in the archive. In this case, a further 301 is sent back to the requester. The departmental server will be reconfigured to recognise this URL as indicating that the archive has been checked, and will therefore be able to issue the appropriate custom error page.

- The user requests a URL e.g. http://www.mydepartment.gov.uk/page1.html
- If the URL is resolved, it is served back to the user in the normal way
- If it is not resolved, the web archive is checked to see if the page exists there. This is achieved by parsing the live URL into the predictable URL pattern used by the European Archive, e.g. http://webarchive.nationalarchives.gov.uk/*/http://www/mydepartment.gov.uk/page1.html
- If it the page is found in the web archive, the user is served with the latest version held there
- If the page does not exist in the archive, the user is served a "custom 404" from the original department website, stating that the page was not found on the original site, or in the archive.[3]

_____

[2] Research conducted in December 2007, surveying 1101 Central Government Websites identified by Central Office of Information (COI) as part of phase 1 of Website Rationalisation, revealed the following usage:
644 uses of Microsoft IIS (of which 257 were using IIS 5.0 and 455 were using IIS 6.0); 287 users of Apache (of which 92 were using 1.3 and 76 were using 2.0)

[3] For a diagrammatic expression of this, see Appendix 1, created by Brian O'Reilly at The National Archives in April 2008.
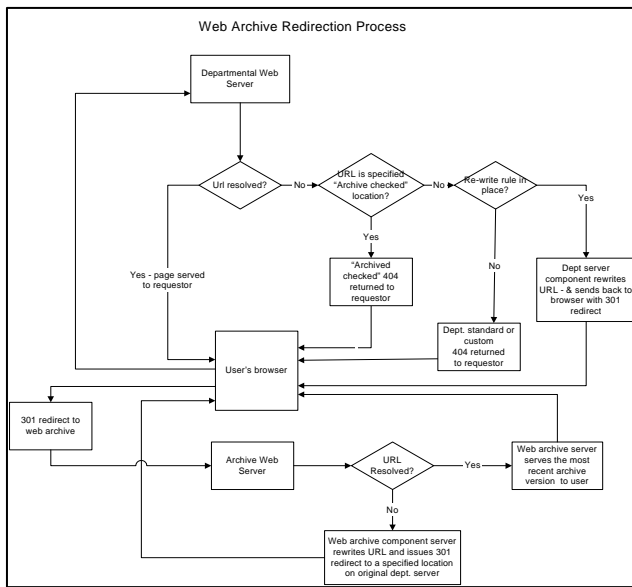
Figure 1. Diagram of Web archiving process,

The components configured, are, of course, only one means by which departments can choose to implement the required behaviour. However, they should be suitable for most government web server platforms. In order to provide guidance to those areas not using one of the two most popular environments, or where the configuration of their environment is atypical, documentation will be provided which describes the behaviour required as well as the technical means of achieving that behaviour. The technology developed, will however, only work successfully if government always uses Content Management Systems which allow the content to be published with persistent URLs. If URLs are randomly generated each time, or if Session IDs become part of the URL itself, then such URLs are irretrievable and will not match anything captured in the Web Archive.

The components are currently being tested at The National Archives, who run a load-balanced Microsoft IIS 6.0 web environment, and the Ministry of Justice, who use an Apache 1.3 environment. Load testing simulating up to 60 concurrent users has been applied, with favourable results. In order that the European Archive can also cope with increased demand for the web archive due to a much greater number of redirections, EA have introduced a mirrored infrastructure capable of failover, with a primary datacentre in Paris and a secondary datacentre in Amsterdam. The European Archive have also developed new indexing techniques to ensure that user requests for

the last archived instance of a website is locally cached, in anticipation of a greater number of calls for the latest snapshots predicated on the increased number of requests to the web archive arising from redirection.

The Web Continuity project and its use of the redirection component signals a marked departure from traditional web archiving programmes in the sense that it is not only concerned with preserving websites for their historical value, but also for their value as recently published information, and for their value in preserving the integrity of the network as a whole. The use of redirection software to persist links to the web archive implies the bringing together different audiences - the archival researcher and the user of current or semi-current information, and in doing so introduces the web archive to new communities of users, and introduces a temporal dimension to the web, which has implications not only for web archiving, but for the wider web more generally, ensuring a greater longevity of web pages than commonly experienced.

This has both benefits and drawbacks: the persistent of links, naturally has an immediate user benefit in that the user journey is less likely to be abruptly ended by a 404 message, but even more than this, the network of interlinking pages that makes up the World Wide Web is preserved, ensuring greater findability of content. However, the persistence of links to semi-current or even out-of-date information also has potential risks, and signals the need for information to be managed differently. Some government organisations have raised the issue of the potentially harmful effect that obsolete information could have if, for example, advice or guidance is revised and moved to a new location. Web users who still have the 'old' URL could potentially unwittingly access information that is no longer current, or which is actually completely inaccurate, for example, where new medical thinking has emerged. To mitigate the risk of archived information being mistaken for live information, The National Archives is working with the European Archive to develop a stripe located above the archived Web page. This stripe will be red in colour, will bear The National Archives logo, and will contain wording to the effect that the Web page is an archived snapshot, taken on a particular date. Various approaches to achieve this design have already been developed, the first of which used Frames. The use of Frames was considered inappropriate because of the issues it poses for accessibility. An iFrames solution was developed by Web developers at TNA, but testing with the European Archive revealed that the iFrame sat awkwardly with the layout of some of webpages already captured in the collection:
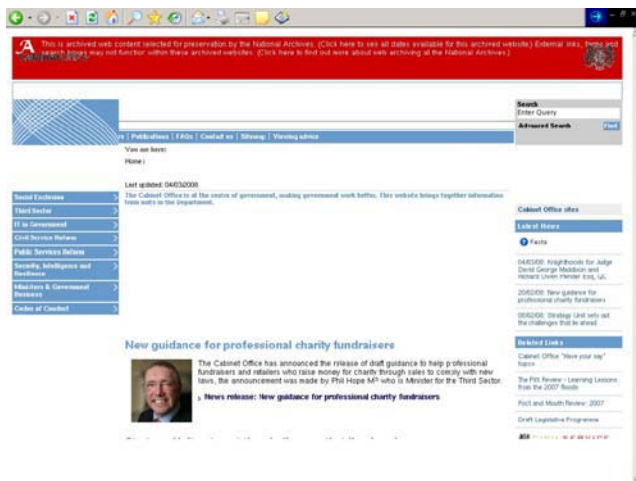
Figure 2. Example of archived webpage where the original layout was adversely impacted by the inclusion of an iFrame.

Currently the use of a server-side include to achieve the desired effect is being developed. Other government website developers and Web publishers solve the problem of the potential dangers of access to inaccurate or out-of-date information by overwriting the information at a given URL as the old information is superseded by new legislation or guidance. While this has an obvious benefit to the organisation and to the user of current information, it causes enormous problems for the user of non-current information – the researcher hoping to compare previous and current policy, the parliamentarian wishing to scrutinize government, the government minister wishing to access documents previously cited by his or her department, the historian wishing to access the record of the past. This issue has proved to be especially problematic in the case of Hansard, which is intended as an historical record of the proceedings of Parliament. When users access a given URL cited in answer to Parliamentary Questions they would often (but perhaps not always) expect to see the information as it was at the time, within the context of the Question that was asked. If information at a URL is regularly overwritten, the answer could be meaningless.

The National Archives has discussed with the Information Services department of the House of Commons the possibility of both parties working on a 'bridging page' which would give the user a choice about whether they see historic or current information, but this was considered to be costly, resource intensive, and would require significant redevelopment of both Hansard, its complex publishing routines, and the Government Web Archive.

Both parties consider that the responsibility for the content at URLs must ultimately rest with government publishers and government Web teams. Government organisations do need to give appropriate consideration to the management of their own content and the user experience.

A further aspect of the temporal expansion of the UK Government Web Archive is the fact that users who unknowingly access the archive through redirection may start to cite the archived URL as a reference in its own right. Consequently, it seemed appropriate to amend the existing contract with the European Archive in order to ensure that a National Archives-specific URL could be developed. The use of a National Archives subdomain, e.g. http://webarchive.nationalarchives.gov.uk/*/http://www.mydepartment.gov.uk rather than the previous European Archive-specificURL: http://collections.europarchive.org/tna/*/http://www.mydepartment.gov.uk, ensures a strong brand association between the UK Government Web Archive and The National Archives, and clearly identifies The National Archives as custodian of these records.

In terms of the practical implementation of the solution across government, the open source nature of the redirection component in particular, together with the intention to disseminate guidance and links to software downloads has called for a new method of distributing and sharing information. The move away from established IT supplier/department relationships and proprietary software implicit in this project will require government to be more flexible and resourceful in its approach to implementation. The National Archives will be unable to offer a high-level of support to individual organisations because of the resources implicit in such an arrangement. Aside from the technical considerations it has also been recognised that for the new processes, tools and guidance to be effective new groups of stakeholders need to be brought together: central Government website managers, e-communicators, IT staff and those involved in producing web standards for Government. For all these reasons, The National Archives has worked closely with COI, who, after identifying that there are many people working in Web-related fields in government and the wider public sector, has established a collaborative working platform called Digital People. It is intended that the Web Continuity and Sitemaps Implementation Plan sub-communities within this platform will serve as a forum for discussion, support and best-practice sharing for those responsible for website and records management across central government and the wider public sector as a whole. Launched at face-to-face project briefing sessions for central government, run by The National Archives, in May 2008, the forum is intended to complement and utilise existing relationships as well as helping to build new ones.

Within The National Archives, stakeholders from across a number of disparate areas are working together to develop the different elements of the Web Continuity solution: IT, Web, Network, Digital Preservation and Records Management and Cataloguing specialists. The outcome of this work is the potential to bring to the Government web

archive to a much wider raft of stakeholders; most notably academics, wider Government, and most importantly will facilitate better access to Government information for the general public.

The National Archives took forward the solution proposed by the working group in November 2007, and intends that the software, guidance and increased scope of its web archiving programme will be ready by November 2008. The implementation of various elements by Government is expected to take longer, although a system of monitoring will be in place by November 2008, serving as a means of encouraging take-up within government.

Conclusions

The way Government uses the Web has brought many benefits but has also posed questions about long-term access to important information. As a result innovative approaches to Web resource preservation have been required. Following research into the nature of the issues facing users of government information, the working group sought to provide a solution which was both user- and Web-centric. The approach which has been developed draws on The National Archives experience of selective archiving, as well as its expertise in the live Web arena.

The greater number of websites to be archived, as well as the need for the content capture to be as comprehensive as possible, led to the development of a registry database, an automated crawling process and the use of XML Sitemaps. The requirement not only to address the problem of disappearing documents from websites, but the issue of broken links and the implications for the user experience led to the development of the redirection concept. The open source nature of the redirection software, and the bringing together of a wide variety of government stakeholders has made it appropriate for government to harness new social networking tools in order to facilitate discussion and collaborative working. The project has also brought together different groups of stakeholders both within and without The National Archives.

The National Archives has been developing National Collection Strategies to address ellipses in archiving and preservation on a UK-wide basis, encompassing a number of themes and formats, and including Websites. The expansion of the scope of The National Archives Web archiving programme to include Websites of all Central Government departments, agencies and Non-Departmental Public Bodies (NDPBs) positions The National Archives as the source of archived Central Government Websites. In order for the concept of Web Continuity to be truly comprehensive across the UK, The National Archives has been involved in discussions with organisations responsible for the preservation of information pertaining to the devolved administrations of Wales, Scotland and Northern Ireland. The project

has also renewed discussions with other organisations in the Web archiving field, such as the British Library and the members of the UK Web Archiving Consortium on the subject scope and collecting remit of the respective organisations.

Redirection to the Government Web Archive has introduced a temporal dimension to the Web, raising important user considerations, which needed to be addressed through the careful labelling of archived material. Redirection will also brings enormous benefits to the user of the Web, with its potential to bring the Web Archive to a more diverse audience.