The International Journal of Digital Curation
Issue 1, Volume 4 | 2009

UK Government Web Continuity: Persisting Access through Aligning Infrastructures

Amanda Spencer, John Sheridan, David Thomas
The National Archives (UK)

David Pullinger,
Central Office of Information

July 2008

Abstract

Government's use of the Web in the UK is prolific and a wide range of services are now available though this channel. The government set out to address the problem that links from Hansard (the transcripts of Parliamentary debates) were not maintained over time and that therefore there was need for some long-term storage and stewardship of information, including maintaining access. Further investigation revealed that linking was key, not only in maintaining access to information, but also to the discovery of information. This resulted in a project that affects the entire government Web estate, with a solution leveraging the basic building blocks of the Internet (DNS) and the Web (HTTP and URIs) in a pragmatic way, to ensure that an infrastructure is in place to provide access to important information both now and in the future[1].

[1] This paper is published with the permission of the Controller of HMSO and the Queen's Printer for Scotland.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

Government's use of the Web in the UK is prolific. Since the early 1990's the Government has been using websites to communicate and present information: official reports, papers, transcripts of speeches, guidance, announcements, press statements, regulations and advice. A wide number of transactional services have also been added, for example submitting tax returns and renewing vehicle licensing. The benefits offered by the Web are its widespread availability and easy access 24/7. This means that services are increasingly being delivered via electronic means and through digital channels. While such developments have brought enormous benefits in terms of making information more widely accessible and readily available to citizens, businesses and professional audiences, there are also risks associated with relying so heavily on the Web as a means for the dissemination of government policy and information.

The Web was initially conceived as a communication channel, to deliver rapid distribution of high-energy physics results to collaborating groups around the world. However the intuitive visual browser interface developed by Mosaic in 1993 led to many other services being accessed through the Web, including access to bodies of material such as library catalogues and academic journals. The Web allows for the dissemination of transient and ephemeral information in a cost-effective way to millions. Conversely, the Web is increasingly used as an interface to repositories of information that support access over long periods. A tension can arise within organisations between optimising their websites for effective communication, with only topical and relevant information available, and ensuring that all information published is maintained and available for future reference. Central government bodies seek to optimise topicality and relevance to their audiences by regularly moving or removing content, while other parts of government treat their websites as a document store and expect to be able to refer to Web content published years earlier. This may not matter in a commercial context, but when it comes to mandatory guidance and legislative matters, it becomes critical that there is a reliable historical record. It is important that people reach the latest mandatory guidance rather then earlier copies; on the other hand, earlier versions may have guided critical decisions and there needs to be access to them, for example in a court of law.

To give a more specific example, both Members of Parliament (MPs) and government officials assumed information situated at Web addresses[2] and referred to in response to Parliamentary Questions would remain available, and the former in particular have been frustrated at not being able to access documents they had previously viewed as part of their work in Select Committees. In April 2007, the Leader of the House of Commons wrote to the incumbent Cabinet Office minister, the Chancellor of the Duchy of Lancaster, expressing grave concern that documents were disappearing from government websites and, in particular, that URL links recorded in Hansard no longer resolved to return the relevant information.

In response, the Cabinet Office established a working group to find a strategic way forward to ensure that MPs would be able to access information referenced by URL in perpetuity. This was led by policy officials from the Central Office of Information with members drawn from The National Archives of the UK which

[2] technically called Universal Resource Locators (URLs)

includes the Office of Public Sector Information and Her Majesty's Stationery Office, the British Library, Information Services at the House of Commons and the Parliamentary Archives.

Disappearing Documents and Broken Links

At the start of the group's work, the primary focus was considered to be a problem of archiving government documents published online, to ensure nothing was lost. By document, we mean information brought together in a coherent way for the purposes of its dissemination. For government, typically such documents are conceived of as printed publications even if they are only published online. The initial exploration was therefore on the need for material to be stored at the British Library in order that it could be accessed, irrespective of whether a referenced piece of information was online or not. This proved to be insufficient for a number of reasons; the most important being that there is no statutory requirement in force for Web-only publications to be deposited with the British Library.

In the UK the British Library is custodian of a body of official government publications that has been built up over centuries, for historical access and long-term preservation of government activity. The government has effective strategies in place for ensuring that all information laid before Parliament is published appropriately in print and flows through to the British Library. Researchers and historians expect such long-term access to official information through the preservation work of the British Library.

However, these government strategies rely on the existence of a printed rendition and a degree of centralisation and control over these official publishing arrangements under the auspices of Her Majesty's Stationery Office (HMSO). Where material falls outside such practices, there are no widely adopted procedures for ensuring that online information is preserved and made accessible. At present Web-only publications are received and preserved by the British Library on a voluntary basis. To meet the objective of persistent links, a solution is needed that addresses all kinds of online information, and one that could be put into place without legislative action.

Investigation

So what was the extent of the problem, and what was likely to be affected by the main issues that arose from it? The working group undertook some small pieces of applied research to find out.

Link Analysis from Hansard: A Longitudinal Study

With the growth of Public Sector Information on the Web over recent years, central government departments have increasingly answered Parliamentary Questions by including a URL citation. Hansard provided a good data source to understand long-term degradation of URLs and documents on government websites.

A bespoke algorithm ("crawler") was developed to "crawl" Hansard to automatically acquire data to this end. The application was written in the C# programming language, using the Microsoft .NET platform. Unusually, URLs cited in Hansard were, at this point, not marked up as links but appeared in plain text. Therefore "regular expressions", which provide a means for recognising strings of text,

were used to identify and help capture URL references. The source of each URL citation and the URL referenced was captured.

To check that the URL references harvested pointed to a resource that matched the context of the answer, a tool was developed to assist with a process of manual testing of each reference. Content was gathered from Hansard for the period 1997 through to mid-2006 – stopping where the file naming convention changed. All 4,000 URLs harvested were manually checked and marked as either succeeding (i.e. the URL resolved to the right Web page), or failing (the URL either failed to resolve or resolved to a 404 error, indicating the page was not found). Automatic testing was not possible given some departments' habit of returning "page not found" with a HTTP 200 message which on the contrary says that the request was successful.

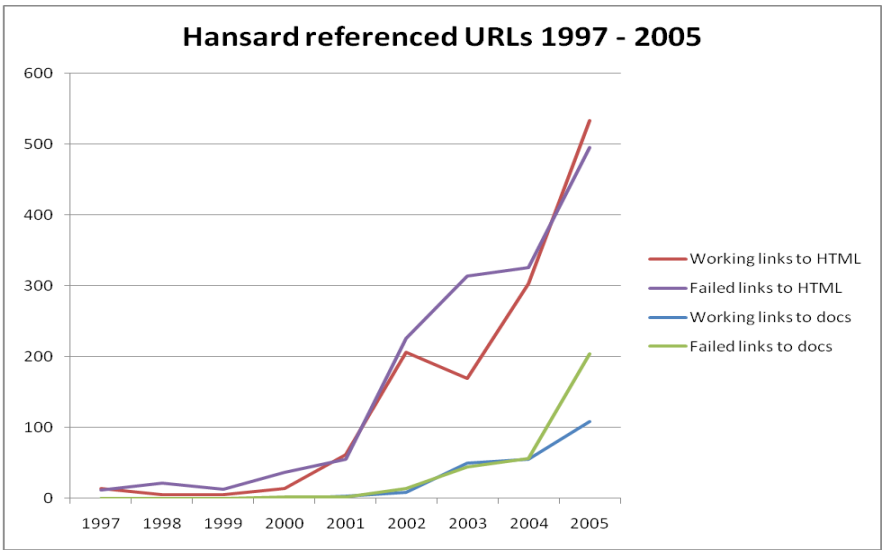

Figure 1. URL references made in Hansard (1997-2005) to UK Government websites (The National Archives [TNA], 2007b).

Comparing the harvests from whole years (2006 was only partially harvested), some clear trends emerge. Departments have increasingly cited a URL in their answer to Parliamentary Questions, and these URLs frequently fail to resolve. Some central government departments had a 100% record of persistent links, whereas others were particularly poor in maintaining links to documents, some having none working at all (Figure 2).

Year	Working links to docs	Working links to HTML	Failed links to docs	Failed links to HTML	Total working	Total failed	Percentage links to docs working	Percentage links to HTML working	Percentage links working
1997	0	14	0	11	14	11		56%	56%
1998	0	5	0	21	5	21		19%	19%
1999	0	5	0	12	5	12		29%	29%
2000	0	14	2	36	14	38	0%	28%	27%
2001	3	61	2	54	64	56	60%	53%	53%
2002	9	206	14	225	215	239	39%	48%	47%
2003	50	169	45	313	219	358	53%	35%	38%
2004	55	303	56	325	358	381	50%	48%	48%
2005	108	533	204	495	641	699	35%	52%	48%
2006	64	344	119	329	408	448	35%	51%	48%

Figure 2. Hansard URL references to UK Government websites (1997-2006). (TNA, 2007b).

The proportion of non-resolving URLs to government and non-government websites were the same (Figure 3). This illustrated that central government departments were no worse than other sectors; however one might expect them to be considerably better given the nature of the content.

Year	Working links to docum-ents	Working links to HTML	Failed links to docs.	Failed links to HTML	Total working	Total failed	Percentage links to docs working	Percentage links to HTML working	Percentage links working
1997		5	0	11	5	11		31%	31%
1998		10	1	13	10	14	0%	43%	42%
1999		4	0	4	4	4		50%	50%
2000		18	0	15	18	15		55%	55%
2001	3	15	1	25	18	26	75%	38%	41%
2002	3	63	4	139	66	143	43%	31%	32%
2003	9	96	17	238	105	255	35%	29%	29%
2004	6	104	18	221	110	239	25%	32%	32%
2005	13	154	31	210	167	241	30%	42%	41%
2006	9	92	16	120	101	136	36%	43%	43%

Figure 3. Hansard URL references to other websites (1997-2006) (TNA, 2007b).

The pattern for persisting URL references to cited documents is worse than the pattern for persisting links to pages of HTML[3]. This is evidence that references to information in document formats are more likely to be moved or lost. This pattern is illustrated by the charts in Figures 2 and 3 above.

[3] HyperText Markup Language (describes structure of text-based information in a document)

The large percentage of non-working links was of considerable concern and as such there was an urgent need to develop a solution.

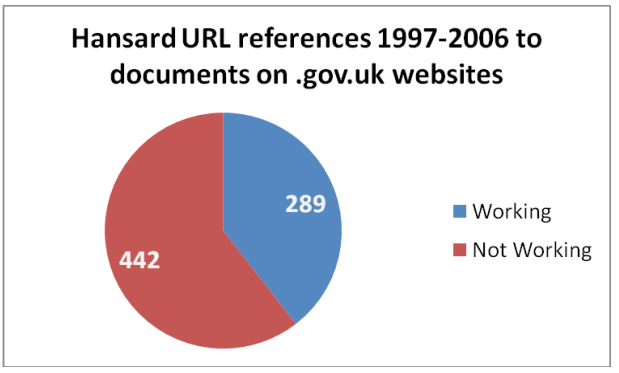

Figure 4. Hansard URL references (1997-2006) to documents on '.gov.uk' websites (TNA, 2007b).

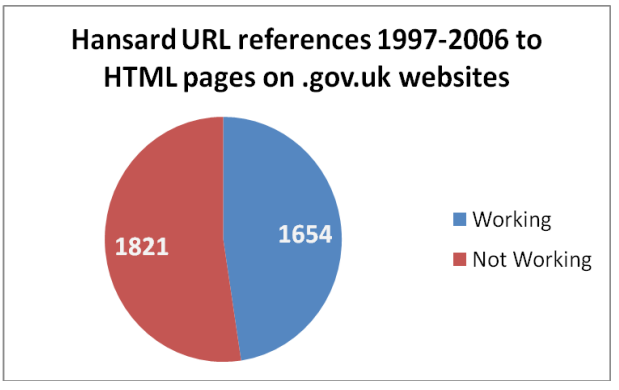

Figure 5. Hansard URL references (1997-2006) to HTML pages on '.gov.uk' websites (TNA, 2007b).

Links Analysis Using Google Webmaster Tools: A Snapshot

The Hansard research showed that a significant proportion of content had been moved or removed. Was this typical or specific to the kind of content cited in responses to Parliamentary Questions?

The second piece of research undertaken used a snapshot approach from data collected via Google Webmaster tools[4]. This involved an analysis of broken links on the indexed Web, focused on assessing the extent to which information had been removed from the Web. Government departments were approached and invited to use Google's Webmaster tools software to help the working group gather data for our analysis.

Google Webmaster tools allow website owners to access a range of metrics and reports collected about their site by the Google Web Crawler.

By collecting this information for a number of key government websites a picture was built up, providing an evidence base for understanding the nature, type and extent of linking to UK Government websites. This included whether external links tended to

[4] Google Webmaster tools http://www.google.com/Webmasters/tools

be to document formats (.pdf[5], .xls[6], .doc[7]) or to HTML pages, as well as the extent of "missing" content (indicated by e.g. HTTP[8] 404 errors) on the indexed Web. This was achieved by analysing snapshots of the error logs of unreachable pages.

There were limitations to this approach. Using the information from Google Webmaster tools did not provide a complete understanding of the extent to which links to documents were broken. In addition Google does not index the whole Web (although they try to index as much of it as they can). URL references in the deep (largely un-indexed) Web, and those from intranets and in users' browser bookmarks, are not included. It is thought that the deep Web is many times the size of the indexed Web[9]. A further limitation was that the indexed Web tends to get repaired (people monitor and fix broken links); those that are not are removed from the Google index. Thus this method could only provide a snapshot of the problem rather than a longitudinal study. Nevertheless links analysis using Google's Webmaster tools offered the most complete snapshot picture of the Web available.

The pattern of persistence of URLs varied between departments and largely echoed that of the research on Hansard. Particularly noticeable was the poor maintenance of links in two departmental websites that had recently been redeveloped; in neither case was the persistence of links a consideration.

It was evident from this research that large volumes of website content, including content published in document formats, was disappearing. Even if the content had been archived, any links to that content were broken.

Our analysis also found that there was a wide variance in practice about whether information is published in HTML or document formats. Most links, including most URL references cited in Hansard are to HTML pages; document formats are the minority of the content that is linked to or referenced. By observation it is apparent some departments use an HTML page as a place to "hang off" documents. Maintaining links to key information, ensuring its persistence and long-term findability means not differentiating between HTML and other document formats such as PDF. For example DEFRA, a good performer with URL persistence and a department that has content widely cited in Hansard, consistently publish policy information in HTML.

[5] .pdf (the format used with the Portable Document Format file format)
[6] .xls (the format for data within the Microsoft Office Excel spreadsheet application)
[7] .doc (the format for data within the Microsoft Office Word word-processing application)
[8] HyperText Transfer Protocol (a communications protocol for the transfer of information on the Internet and the World Wide Web)
[9] The size of the deep Web is unknown. In a White Paper designed for marketing purposes and written in 2001, Bergman estimates the size of the deep Web at over 500 times the size of the indexed Web (http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104). This is probably an overestimate as Lewandowski has pointed out (http://www.ib.hu-berlin.de/~mayr/arbeiten/Lewandowski-Mayr_BC06.pdf), but his research concurs that the deep Web is an order of magnitude larger than the indexed Web.

Analysis of the Causes

A comparison of the proportion of broken links with the history of particular government websites revealed a number of reasons for poor maintenance of URL persistence:

- rebuilding of websites on different technical infrastructures;
- changes in the machinery of government, for example when a department has its responsibilities and name changed;
- management attention focused elsewhere.

The above represent major events, indeed, disruptions, in the history of a government department website, and unless URL persistence is planned for, links will be broken, in some cases universally. According to Tim Berners-Lee, the inventor of the Web, "cool URIs don't change"(Berners-Lee, 1998). We agree. Unfortunately the structures of government can shift almost as rapidly as the Web itself is changing. Departments are split, merged and renamed. New domain names are created and websites reconfigured, making it very difficult to maintain persistent links to documents.

An important conclusion from this research was that current practice was not sufficient to maintain links to information in the official record, and that only a government-wide solution could suffice. The issue is compounded by the trend towards the electronic-only publishing of official publications. This increasing reliance on solely electronic versions makes the integrity of Web links all the more crucial to the business of government.

In summary, there is evidence that links are not persistent and there may also be extensive loss of information from the public domain. The prevalence of broken Web links reflects badly on the reputation of government because it is perceived that government is managing its information poorly. Where attempted usage is frustrated by such lack of persistence, it can only serve to reduce public confidence in the services the state provides online. Parliamentary scrutiny of government is impaired by the inability to refer to key government documents published on the Web with any reliability.

Given this situation, 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 by making a request from the original location. The latter idea reflects an acknowledgement that the Web has changed user behaviour in searching for and finding information. It is about to change further with the opening up of online information for data mashing and the growth of the Semantic Web. Therefore any solution to the problems identified needed to take account of the changing nature of and potential uses of the Web.

The Nature of Linking to the Government Web Estate

The research also revealed a situation with much broader implications for government and its use of the Web. The pattern of links to UK Government websites follows a power law distribution pattern so common on the Web. This gives rise to the distinctive "L" shape graph when shown using a logarithmic scale as in Figure 6.

Across departmental websites there are large numbers of pages (many hundreds) with between 1 and 10 external links. Of the sites surveyed there were 21,689 pages, with between 1 and 10 external links – but these account for just 5% of the external links identified. Compare this to 12 pages in the government domain that among them hold more than 10,000 links, and which account for 785,495 (or 62%) of the external links identified.

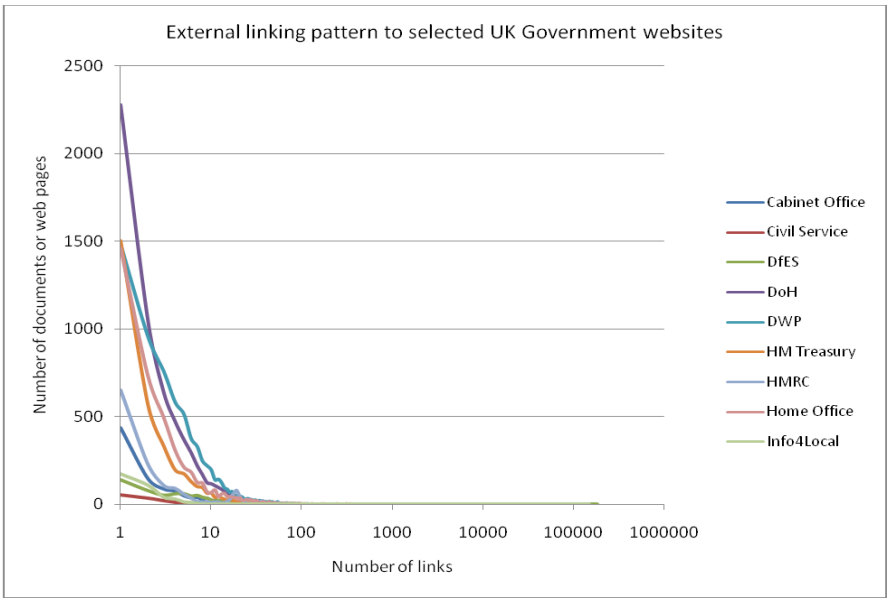

Figure 6. External linking pattern to selected UK Government websites. (TNA, 2007a).

It is evident from the table below that these "link-heavy" pages are home pages or other key general pages.

Page	Number of external links
http://www.dfes.gov.uk/	183229
http://www.info4local.gov.uk/	151428
http://www.dh.gov.uk/	118784
http://www.homeoffice.gov.uk/	75388
http://www.dwp.gov.uk/	62143
http://www.dfes.gov.uk/copyright/	49494
http://www.hm-treasury.gov.uk/	32055
http://www.homeoffice.gov.uk/security/current-threat-level/	30481
http://www.dh.gov.uk/en/Home	29650
http://www.dh.gov.uk/en/ContactUs/index.htm	21198
http://www.cabinetoffice.gov.uk/	16275
http://www.cabinetoffice.gov.uk/chartermark/	15370

Figure 7. Table of pages with over 10,000 external links (TNA, *2007a*).

Three observations follow:

- The Government can ensure the persistence of the vast majority of external links to its websites by maintaining the links to a small number of Web pages; arguably the relatively small number of links that point to the majority of "linked-to" pages are disproportionately important in terms of providing "link metadata" about those pages.
- Whilst less than half of the links are to government websites, the links to the majority of pages are key to aiding the discovery of much of the content via search engines such as Google. It may also be the case that the failure of a significant number of government websites to ensure URL persistence for much of their content forces sites linking to government locations to take the safe option and link to the home page, rather than a page lower down that may be moved.
- If people cannot trust in the persistence of government-created URLs they will be less inclined to link to government domain content, thus directly impairing the discovery of content on its sites.

Another striking result is the number of links to UK Government websites from other UK Government websites. As Figure 8 shows, some 55% external links identified to the selected sites are from other UK Government Websites.

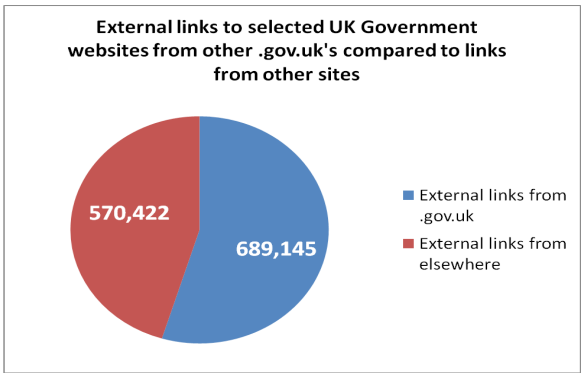

Figure 8. Proportion of external links to selected UK Government websites from other '.gov.uk' sites compared to external links from other sites (TNA, 2007a).

This is clear evidence that the set of government websites forms an ecosystem of links. This whole system aids the discovery of content on all of the government websites involved. With so much interlinking, each of the websites in this ecosystem is supplying data for the search engine algorithms responsible for indexing, ranking and relevance. In essence, by creating websites and linking to other '.gov.uk' sites, departments and agencies have been meta-tagging each other's content (in the form of link text) – and by doing so have provided key metadata that are used by all the major search engines.

In summary, the research identified that:

- A significant proportion of Central Government content was unable to be accessed through their original cited URLs.
- This was worse for documents than for HTML pages, although many URLs link to pages holding document links.
- There is a widespread problem. The content lost does not appear to be specific to the kinds of information referenced in responses to Parliamentary Questions.

- As an unproven hypothesis, those that link to central government websites may not have the confidence to deep-link to content on government websites, given the degree of lack of persistence and prefer instead to link to top-level navigation pages.
- Many of the links to government websites come from other government websites, thus forming an interlinked network of sites.

Possible Solutions

Faced with the research results, the working group moved away from its initially narrower focus on preservation of content, to recognising that links also need to be maintained. Thus a solution was needed not only for the topicality – reference store tension – but also to include maintenance of linking; in other words, persistent URLs.

The articulation of this can be summarised as:

- allow central government departments to focus on topicality;
- find a way of archiving all content no longer considered topical or relevant;
- continue to be able to link to content whether live or archived;
- maintain the interlinking between the content, whether live or archived.

To guide the solution space, a set of principles were drawn up, with a view to minimising change to current practice and to cost; they were:

- archive sufficient to maintain public-facing government resources, including all information, not just a snapshot of look and feel;
- use URLs - they can be employed effectively as identifiers with appropriate guidance and do not change the current system;
- ensure links issued by government departments continue to work:
    - through changes of political administration (the supposition is that a new administration may not wish to be associated with the policy statements of its predecessor and thus significant volumes of content might be removed)
    - through machinery of government change
    - when websites close
    - when content archived from live sites
- avoid large new technical systems unless proven necessary;
- facilitate process to ensure access and long-term preservation of government Web-only information;
- ensure that any policy is realistic and that guidance and advice prove practicable enough for departments to implement.

An additional principle addressed the improvement of searching through effective metadata, but was remitted to another group exploring how to improve the discovery of content across the government Web estate.

The solutions we considered included, but were not limited to, the approaches described in the three sections that follow.

Depositing All Content into British Library

Long term, this is the home for preservation of all published information. However to obtain delivery of all Web only publications would involve legislative change on legal deposit of Web publications. It would then be unlikely to help users as they would still have to locate the material.

Government Departments to Take Responsibility for Link Maintenance

Website owners and managers should take responsibility for link maintenance. The practical focus of government is always in presenting topical information relevant to audiences and not in older superseded content. There is only funding for the former and not the latter. Indeed, the Public Records Office, now The National Archives, was established to preserve such material. The departments' responsibility was therefore considered to be limited to ensuring link maintenance for topical material.

Use Digital Object Identifiers

Digital Object Identifiers (DOIs) permit the ownership, rights and location associated with a unique identifier of a piece of information to be changed but the information still accessed. This was a popular idea as one particular form had been successfully implemented in academic journals.

Such a system would require a high degree of centralised management and control on an ongoing basis, with each publisher giving information a unique reference before publication and maintain its associated information. This could be quite straightforward with those official documents published in a centralised way and laid before Parliament. In government most Web publishing is carried out by numerous devolved editors who have access to their departmental content management system (CMS), and whose task is to publish, not to maintain the system.

Government is intrinsically distributed, and the government's Web publishing activities are too diverse to attempt to manage centrally on a document-by-document basis. Moreover, we did not need a generic solution, just one that would work for a relatively closed (albeit large) set of departments and their agencies. There was no comparable example of DOIs being successfully deployed in such a highly decentralised and distributed publishing environment.

Given the fragmentation of publishing responsibilities within government, exacerbated by the use of social media, such as blogs, the systematic use of DOIs across government was considered both too complex and too expensive. What was needed was a fundamentally Web-centric solution, rather than a publishing process solution. By working through each of these options, we arrived at the solution that is now being implemented under the name Web Continuity.

Web Continuity

In essence the Web Continuity solution adopts elements of all the options:

- allow central government departments to focus on topicality but ensure that they deploy links with effective URLs that, for example, persist through technical changes;
- ask The National Archives to archive all content no longer considered topical or relevant, extending their previous snapshot approach to one of

systematic and comprehensive website archiving;
- develop a tool that would maintain a link to content whether live or archived by tying up an original URL with the archive version in the Web Archive;
- maintain the interlinking between content either live or archived (achieved by the tool development in the previous point).

The solution is based on the recognition that the Domain Name System (DNS) provides an effective method of resolving requests to a document using an identifier and that this can be based on the original URL of the resource. All that is needed therefore is a means of capturing the content (provided by Web-archiving technology), coupled with a means of redirecting a user to that content: When a document is requested which is no longer on a department's website, the user is automatically redirected to the archived version, in the website archive. By customising existing open source components and installing them on departments' web servers, this behaviour can be implemented across the entire government Web estate at relatively little cost.

The solution devised for Web Continuity is entirely Web-centric and hinges on the use of identifiers. On the Web such identifiers are called URIs (Uniform Resource Identifiers). These are strings of characters used to identify a resource uniquely, such as a document. The W3C's Architecture of the World Wide Web, Volume One (Jacobs & Walsh, 2004) says that everything of importance deserves a URI. Moreover, the first architectural principle of the web is that "global naming leads to global network effects." Consequently it is very important to give resources persistent URIs as by doing so it is possible to create network effects of immense value.

The correct use of URIs should exhibit a set of architectural principles referred to as Representational State Transfer or REST. Fielding and Taylor (2002) set out these principles after the emergence and popularisation of the Web. Whilst the RESTful principles were inherent in design on the web from its earliest days, Fielding's work established them in formal terms.

The principles of REST are that everything is a resource and every resource is given a URI. The URI provides a means of uniquely addressing each resource. Every resource shares an interface for the transfer of state between the client (the user of the resource) and the resource itself. This interface consists of a set of well-defined operations which are part of the HTTP protocol. The resource itself is stateless (there is no memory of previous operations as these are immaterial to the resource, there is only "now") and the resource can be cached and layered. A "resource" can be anything in the universe about which a statement can be made, not only entities, such as official documents, which can be returned in electronic form using the Web.

There are two types of URI, URLs (Uniform Resource Locators) and URNs (Uniform Resource Names). The difference between the two is that where a URN provides a unique name or label for a resource, a URL returns a representation of the resource from the Web, for example, a Web page loaded in a browser. Where anything can be given a URI, only resources that can be returned across the Web can have a URL. In other words, a URL is a specialised type of URI.

For the purposes of Web Continuity, our concern is repeatedly capturing representations of resources on the Web (such as official publications) in the Web Archive at different points in time, which occurs when those resources are harvested. A new representation of the resource is created in the Web Archive and is given a URI which dereferences to the resource.[10]

At the point of capture of a resource in the Web Archive, we are creating a new representation of the resource and assigning to that representation a new URI. This URI is formed from the combination of the day the new representation of the resource was created and the location (the URL) of the original resource that was used to create the archival resource at the time of its capture and creation in the Web Archive. The new URI for the resource in the Web Archive is both a URN, a unique and permanent label for the resource, and a URL – enter the URI into a browser and the resource is returned.

The key idea of Web Continuity is that of mandating a behaviour for URLs across the government Web estate, such that when a representation of a resource ceases to be available from its original location on a department's website, the most recent representation of a copy of that resource (the archival resource) is automatically returned to the end-user from the Web Archive, using a 301 redirection. This behaviour has the effect of maintaining any links which have been made to the document by associating them to the archival resource.

An important objective of Web Continuity is that the valuable network of links between resources is maintained and moved from the current to the archival resource at the point the current resource is removed from the department's website. Both the inbound network from the wider Web to the archival resource and the outbound network from the archival resource to the wider Web are sustained.

Implementation

In November 2008 The National Archives began the comprehensive archiving of the government Web estate. This involves the harvesting of content from around 1,500 websites three times a year, or by request. Fortunately this number will decline over the next three years, as through the Transformational Government programme the number of websites is being reduced to deliver a smaller number of higher-quality ones focused on particular audiences.

Government departments will need to introduce XML (Extensible Markup Language) sitemaps as a supplementary means of directing website crawls. This supports more comprehensive capture of Website content, by providing information on the location of "hidden" (unlinked-to) Web pages or "virtual" pages generated by dynamic (CMS or database-driven) applications, to avoid missing content.

Associated with comprehensive collection by The National Archives is the use of a software component on each government department's Web server. This affects the

[10] The term *dereferencing* describes the act of obtaining a representation of a description of an entity via its URI. A dereferenceable Uniform Resource Identifier or dereferenceable URI is a resource identification mechanism that uses the HTTP protocol to obtain a representation of the resource it identifies.

desired redirection behaviour, so will serve the resource from the Web server in response to a request if that resource still exists and, if not, initiates a checking process with the Web Archive to see if the resource exists there. The components, based on open source software, configured, tested and supplied to departments by The National Archives, have been designed to work with Microsoft Internet Information Server versions 5 and 6, and Apache versions 1.3 and 2.0[11]. The IIS component is produced by Ionics[12]. The Apache component is the module `mod_rewrite`[13].

If the user requests a URL: http://www.mydepartment.gov.uk/page1.html, then:

1. If the request to the URL can be resolved, the resource is served back to the user in the normal way;
2. If the request cannot be resolved, the Web Archive is checked to see if the resource exists there. If it does, the user is served with the latest version of the resource held there, for example:
   http://Webarchive.nationalarchives.gov.uk/*/http://www/mydepartment.gov.uk/page1.html;
3. If the resource does not exist in the Web Archive, the user is served a "custom 404" from the original department Website, which states that the page was not found on the original site, or in the Web Archive.

The component works by rewriting URLs not found on the departmental website, and then sending the redirection instruction (HTTP status code 301 – 'moved permanently') to the user's browser. This is the most search engine-friendly redirect and aids the long-term discovery of the resource beyond the point when it has been removed from its original home. The browser then requests the page from the Web Archive. These components 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.

The component is also installed on the Web Archive. Here its role is to rewrite the URL for the original department website, if it is not found in the Web 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.

As these components are introduced, one useful feature is that any pages that The National Archives have previously archived will then become available and accessible again. In other words, some links that currently do not work from, Hansard, for example will, in time begin to work again.

Consistency of user experience is maintained through always serving the latest available version of the document from the Web Archive when the page no longer exists on the live website, as this is consonant with the live Web experience. In certain situations, the user will want to find the information at a link *as it was at the time the link was recorded* (for example, if cited in Hansard). The introduction of a "stripe" at

[11] Research conducted in December 2007, surveying 1101 central government websites identified by the Central Office of Information (COI) 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 v.1.3 and 76 were using v.2.0)

[12] Ionics Isapi Rewrite Filter http://www.codeplex.com/IIRF

[13] Module `mod_rewrite` URL Rewriting Engine
http://httpd.apache.org/docs/1.3/mod/mod_rewrite.html

the top of the archived page provides users with the option to navigate to earlier versions of the page. This stripe also enables content in the Web Archive to be clearly labelled as archival material, reducing the likelihood of users becoming disorientated by automatic redirection from live sites to the Web Archive. The National Archives has also taken measures to ensure that search engine results will have the description "[Archived Content:]" in the title tag, in order to help the Web audience make sense of material return through Search.

Conclusions

The government set out to address the problem that links from Hansard were not maintained over time and that therefore there was need for some long-term storage and stewardship of information, including maintaining access. As the work progressed, a tension was identified in government departments between their desire to maintain relevance and topicality to their users, and the need to ensure long-term access to material (which might also mislead or make finding the latest relevant content harder).

Further investigation revealed that linking was key, not only in maintaining access to information, but also to the discovery of information. Because government websites form a significant linked ecosystem on their own, those links contribute greatly to resource discovery through the major Web search engines. Information would become harder to find through both linking and searching if links are broken. It therefore became imperative that a solution was found to accommodate this too.

The solution delivered levers the basic building blocks of the Internet (DNS) and the Web (HTTP and URIs) in a pragmatic way, to ensure that we preserve both the contents of the government's Web estate over time and the value of the network – the rich links from and to information on the government Web estate. As information becomes increasingly interlinked, the concept of preserving the value of the network is likely to become more important.

The principle of redirection to an archive in the event that content cannot be served has applications far beyond the narrow confines of the UK Government Web estate, and could bring benefits to a variety of Web communities – for example, in the academic, library and commercial sectors. Redirection to the Government Web Archive also introduces a temporal dimension to the Web, raising important user considerations, which needed to be addressed through the careful labelling of archived material. Redirection brings enormous benefits to the user of the Web, by ensuring continuity in the user experience, and has the potential to bring website archives to a more diverse audience.

Moreover the principles of Web Continuity could equally be applied to other distributed heterogeneous information systems, not just the Government Web estate. Finding pragmatic approaches to preserve the value of networks between resources as well as the content they contain is an important concept with which the archival and digital preservation community needs to grapple. The emergent linked data initiative and prospects for enterprise linked data mean that these issues are of direct concern to The National Archives core business of preserving government records.

The selection of open source components has provided an inexpensive and readily available redirection solution for government. The commissioning of a bespoke solution would have been more costly and time-consuming to implement, and, because

of varying infrastructure across government, would have had limited benefits in its application. The open source alternative is, however, not without its issues. The generic nature of the tools means that problems could be encountered on attempted use within different areas of government, even if they are using one of the Web server infrastructures for which the component has been developed, because of the variety of Web environment configurations in operation. The use of open source software also has the potential to be problematic for government organisations so accustomed to high levels of support from third-party IT service providers. Government will need to embrace this new method of working, and collaborate with peers to share experiences of implementation and to learn more about the Web and its potential use in government. With the express purpose of bringing together individuals and groups in government with responsibility for digital issues, the Government has established a peer-to-peer discussion forum - Digital People. In using this platform, government is already utilising innovations in technology brought about by the Web 2.0 revolution and changing the way in which it works.

Aside from the technical benefits that the Web Continuity Project brings to government and the end-user, the solution has made a positive contribution to other initiatives being taken forward by The National Archives: the Web Continuity Project relates to the permanent preservation of digital records transferred typically from government Electronic Records Management Systems (developed as the Seamless Flow Programme), and the current Digital Continuity Project concerned with providing a shared service for government, which tackles the issues of the potential technical obsolescence of digital formats which may soon impede access to semi-current information required by government in the future.

The innovative solution the Government is introducing is cost-effective, needing many small changes to practice, rather than introducing one large central system, and keeps the interests of each of the parties focused and in line with their organisational responsibilities. Most importantly, it provides users with effective access to information and will deliver and maintain easy routes to the online public record of government over the coming decades.

References

Berners-Lee, T. (1998). *Cool URIs don't change*. Retrieved May 25, 2009, from World Wide Web Consortium Web site http://www.w3.org/Provider/Style/URI

Fielding, R. T., & Taylor, R. N. (2002). Principled design of the modern Web architecture. *ACM Transactions on Internet Technology (TOIT) 2 (2)*: pp. 115–150. New York: Association for Computing Machinery. ISSN 1533-5399, doi:10.1145/514183.514185

The National Archives. (2007a). *Government websites link analysis* (unpublished survey of external weblinks on UK Central Government websites conducted to support the work of the 2007 Working Group on Digital Assets and Link Management). John Sheridan.

The National Archives. (2007b). *Hansard link study* (unpublished survey of weblinks cited in Hansard conducted to support the work of the 2007 Working Group on Digital Assets and Link Management). John Sheridan.

Jacobs, I., & Walsh, N. (2004). *Architecture of the World Wide Web, Volume One*, W3C Recommendation, December 15, 2004. Retrieved May 25, 2009, from: http://www.w3.org/TR/webarch/