Past and Present: Using the UK Government Web Archive to Bridge the Continuity Gap

Amanda Spencer
The National Archives (UK)
Kew
Richmond TW9 4DU UK
Amanda.Spencer@nationalarchives.gsi.gov.uk

Brian O'Reilly
The National Archives (UK)
Kew
Richmond TW9 4DU UK
Brian.Oreilly@nationalarchives.gsi.gov.uk

Gabriel Vasile
European Archive Foundation
Montreuil
93100 France
gabriel@europarchive.org

ABSTRACT

The UK 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.

Categories and Subject Descriptors

To be determined

Keywords

To be determined

1. INTRODUCTION

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 'disappearing', 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[1] in response to parliamentary questions[2]. 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.

2. RESEARCH AND BROKEN LINKS

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 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, did not resolve, suggesting that many government departments do not consider the issue of long-term access to government information. 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 [2].

This issue is compounded by the fact that much of this information is increasingly only available electronically, and 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.

3. THE WEB CONTINUITY SOLUTION

The Web Continuity solution being taken forward by The National Archives adopts the following elements:

- Archiving of all content no longer considered topical or relevant, by The National Archives, extending their previous snapshot approach to one of systematic and comprehensive website archiving;
- Development of a tool that maintains a link to content whether live or archived by tying up an original URL with the archive version in the Web Archive;

[1] Hansard (the Official Report) is the edited verbatim report of proceedings in both Houses of the UK Parliament.

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

- Maintaining 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 redirected to the archived version, in the website archive. By customizing existing open source components and installing them on departments' web servers, this behavior can be implemented across the entire government Web estate at relatively little cost.

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 [2].

4. IMPLEMENTATION

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

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 (the Government Websites 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 Rationalization program, and evidence of compliance with government web standards guidelines (such as the accessibility standard). The database, launched in Spring 2009, is 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.

4.1 Capture

The archiving of government websites is operated by the European Archive Foundation[3] using the most popular method of capture for large-scale programs: remote harvesting using a Heritrix web crawler [4]. The Web Continuity Project has meant a significant increase in the number of websites captured, moving from a selective archiving program to a comprehensive program 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 program 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 been working with government content creators to raise awareness of the importance of best practice in website design and in terms of managing websites through change. To this end, The National Archives has worked closely with the Central Office of Information, responsible for mandating Web Standards across government, to produce several new standards and guidance around Web Archiving and Managing URLs [1]. The National Archives has also explored the possibility of using the XML sitemap protocol[4]; to ensure that capture of the Government web estate is comprehensive. It is now used as an additional seed for the crawler completing the discovery process.

4.2 Redirection of 404s to the archive

The most innovative element of the project involves the configuration of government web sites to perform automated checking in the web archive when users request a resource not found on the department's own site.

4.2.1 Implementation at the WebSite end

In its simplest implementation, the redirection process works as follows (and as illustrated in Fig. [1] below):

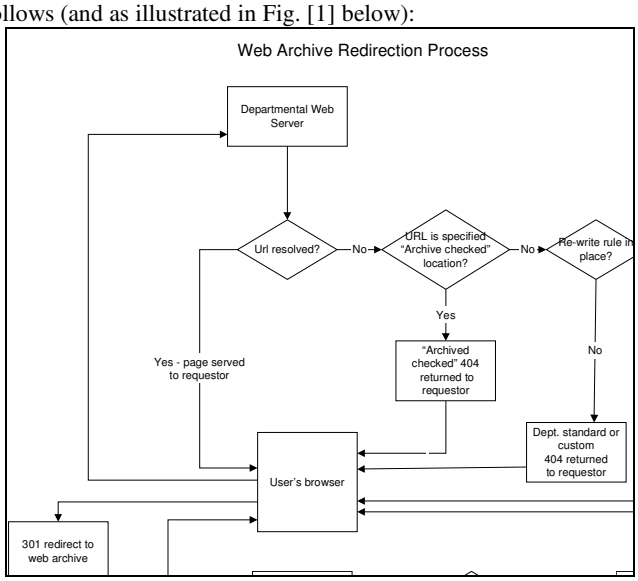

Fig [1] Basic web archive redirection process

1. The user requests a URL, e.g. http://www.mydepartment.gov.uk/page1.html. If the request to the URL can be resolved, the resource is served back to the user in the normal way;

[3] http://www.europarchive.org/ [Last visited 10th of Aug 2009]

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

2. If the request cannot be resolved, the departmental web server issues a redirection instruction (HTTP status code 301 – 'moved permanently') to the user's browser. This instructs it to check the Web Archive to see if the resource exists there. A 301 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.

3. The browser then requests the page from the Web Archive automatically. For the URL in (1) above, the request would be http://webarchive.nationalarchives.gov.uk/+/http://www/mydepartment.gov.uk/page1.html. The '+' in this URL indicates that the web archive should serve the latest available archived version of the page, if several versions exist captured at different times

4. The web archive will then serve this page if it exists.

5. If it does not exist, then the web archive (which also has the URL rewriting software installed) rewrites the URL and sends a further redirection request back to the user's browser.
This rewrite uses the pattern http://www/mydepartment.gov.uk/ukgwacnf.html. This indicates to the department that this is a redirection following an unsuccessful check in the web archive. The department can therefore intercept such requests and serve a 404-error page that indicates that the page the user requested does not exist either on its own site or in the web archive.

This can be achieved on Apache server installations by levering the URL rewriting capabilities of the native mod_rewrite module (available in different versions for Apache 1.3, 2.0 and 2.2)[5]. For Internet Information Services (IIS) 5.0 and 6.0, a number of third party products can be used to provide similar functionality, and the National Archives has sourced the component provided by Ionics[6]. Both this and the Apache mod_rewrite have been through a testing process and form TNA's suggested options to departments (both are free and open source). In both cases, the Perl Compatible Regular Expression (PCRE) syntax is used to configure the URL rewriting and redirection behaviour. Departments are at liberty to use alternative proprietary or bespoke solutions as they see fit, however these options should be suitable for most government Web server installations[7].

At the time of writing, the above approach has been implemented on the main websites of the former Department for Business, Enterprise and Regulatory Reform (BERR – www.berr.gov.uk) and Ministry of Justice (MOJ – www.justice.gov.uk). The BERR implementation uses the Ionics Rewriter, whilst the MOJ one use Apache mod_rewrite.

A variation on the above approach is to use an intermediate or "bridging" page to enable the user to choose whether or not to look for the page in the web archive, instead of redirecting automatically. This has been implemented by the Cabinet Office (www.cabinetoffice.gov.uk) using their own bespoke solution, rather than a third party rewriter such as Ionics. This use of a bridging page is arguably less disorientating for the user (see also the discussion below), and preserves an element of choice, although on the other hand it loses the potential to maintain search engine rankings that a straightforward 301 redirection might provide.

A further variation of this approach has been implemented by the National Archives, also using a bespoke solution, and running on a load balanced IIS 6.0 installation. This introduces a further innovation by pre-checking the web archive to see whether or not the page the user requested exists there. The process works as follows:

1. The user requests a URL, e.g. http://www.nationalarchives.gov.uk/events/default.htm.

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, a redirection instruction with the HTTP status code 303 (see other) is sent to the user's browser. For example, if the original request was for http://www.nationalarchives.gov.uk/news/stories/10.htm, the 303 redirect's location will be:

http://www.nationalarchives.gov.uk/PageNotFound/PageNotFound.aspx?url=http://www.nationalarchives.gov.uk/news/stories/10.htm

3. Before displaying to the user, the PageNotFound.aspx page checks "behind the scenes" whether or not the original URL, as passed in the 'URL' query string parameter, exists in the web archive. This is achieved via a HTTP HEAD request, rather than the standard GET, since we do not actually want to view the page at this point (and this also reduces network traffic and the load on the web archive servers).

4. If the requested page is found to exist in the archive, PageNotFound.aspx displays in "bridging page mode". This includes a hyperlink to the web archive via which the user can retrieve the latest version of the page (in this case the link would be:
http://webarchive.nationalarchives.gov.uk/+/http://www.nationalarchives.gov.uk/news/stories/10.htm).

In this mode, PageNotFound.aspx returns a HTTP 200 code to the user.

5. If the page does not exist in the web archive, the archive, PageNotFound.aspx is displayed in "Not Found mode" and returns a HTTP 404 code. The wording informs the user that the originally requested page was not found on either the National Archives website or in the web archive.

The introduction of redirection behavior onto a government website can normally be carried out without disrupting existing functionality on the site, with a simple implementation of redirection using Apache mod_rewrite or Ionics requiring very

---

[5] Module mod_rewrite URL Rewriting Engine http://httpd.apache.org/docs/1.3/mod/mod_rewrite.html , http://httpd.apache.org/docs/2.0/mod/mod_rewrite.html and http://httpd.apache.org/docs/2.2/mod/mod_rewrite.html

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

[7] 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) (unpublished).

little PCRE programming using basic rules covered in TNA guidance. However, given the wider capabilities of these components, they can also be used for other website management tasks. For example, departments may wish to redirect requests for resources in specific directories to another area of their site, or to another government site. In other cases, it may be desired to rewrite URLs to allow more user (and search engine) friendly versions to be presented publicly, but for the web server to use original more complex syntax to communicate with a database or content management system (with URL rewriting, as opposed to redirection, the original URL remains visible in the browser's address bar). The PCRE syntax can accommodate these and similar scenarios, alongside redirection to the web archive for resources that are not covered by other rules. The rich capabilities of the PCRE syntax, however, means that programming certain more complex scenarios will require suitable skills in this technology.

As redirection behavior is introduced across government, 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.

4.2.2 Implementation at the Archive end

In order to ensure navigability on the archive, Web Archive have to apply a set of transformations on the archive content when serving it [3]. The main one being to change absolute links (pointing to the original site) into links going to the archive. This can be done by in 3 different manners:

1. Changing links in the code of archived pages (at the risk of loosing authenticity)
2. Appending a JavaScript that automatically re-writes links when the browser displays the page (client-side rewriting)
3. Rewrite links when the page is extracted from the archive before it is served to the user (server-side rewriting)

The main issue with option 2/ is the impossibility to apply rules on time (before the browser actually renders the page) or in some cases, where rewriting requires more complex operations, to rewrite at all. The main advantage of this method is that it does not put the rewriting load on the archive, which is critical for large-scale archives (like the Internet Archive).

The European Archive has opted several years ago to the third method (server-side rewriting) to achieve better quality for its partners (among which are The National Archives of the UK).

European Archive rewriting engine is composed of regular expression and runs as an Apache module. Taking into account the URL and the mime type, a decision is made about applying a specific set of rules for URL discovery and rewriting.

Currently EA uses several sets of regular expression, one for html rewriting, another for JavaScript, some other for CSS and XML for instance. Another class of regular expression deals with injecting EA specific code into pages, for instance code replacing video players with our own player.

These rewriting occur for both embeds (images, CSS, video and other secondary elements constituting a page) and main navigational links (to other pages for instance).

Having this server-side rewriting engine already in place has significantly facilitated the implementation of the redirection process for which the requirement is slightly different: although all embeds needs to be rewritten to display pages properly, other navigational links should, on the contrary, not be rewritten, to ensure that users continue their navigation on the original website and not on the archive (which could be very confusing)[8]. The navigation, from the user point of view, goes from the live site to archived pages only when the live pages don't exist any longer. The archive is only a way to avoid 404 in this case, it is not used for continuous navigation on older versions of the site.

4.2.3 Redirection infrastructure

The redirection is being deployed on large sites that get much more traffic than what web archives usually do. In order to support Government wide redirection implementation, the European Archive have carried out various enhancements to cope with the estimated 10 fold increase in demand (based on analysis of webservers logs of several of these sites). EA's serving capacity which was previously of app. 100 requests per second, has been increased by one order of magnitude.

Located in two different datacenters, one in Paris and one in Amsterdam the serving farm is composed of different classes of machines:

1) Fast processors machines for url rewriting

2) Fast disk machines for large indexes (10 k rpm server grade disks)

3) Sizeable memory machines for special purpose indexes and for caches (16GB per machines)

4) Machines with large storage capacities for storing all our collections (Paris machines store 30 TB in RAID5, 1 PB in Amsterdam, used in RAID 1 for these collections).

5) Monitoring machines constantly checking services and ready to switch a service from a failed machine to a running one.

An external monitor periodically verifies the state of our datacenters and decides whether to switch services from one to another if the parameters fall bellow an established threshold.

To access pages in this way, EA has developed a new syntax for archive URL.

---

[8] See for instance the difference in navigational links between (archive version):

http://webarchive.nationalarchives.gov.uk/20080107223656/http://www.nationalarchives.gov.uk/default.htm

and (redirect version):

http://webarchive.nationalarchives.gov.uk/+/http://www.nationalarchives.gov.uk/default.htm

4.3 Discussion of the solution

The Web Continuity project and its use of the redirection component signals a new avenue for traditional web archiving programs 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 organizations 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, departments have the option to use a bridging page as described above. Also, pages served from the archive are headed with a banner bearing the National Archives logo, and containing wording to the effect that the Web page is an archived snapshot, taken on a particular date, as well as the option to navigate to other archived versions of the page, as 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 banner is inserted via server-side code in the HTML of the captured page, including new tags to hold the banner and to act as a container for the original page <body> content. In most cases, this can be done without adversely impacting on the page's appearance. However, this is not always the case, and work is ongoing to address issues with certain pages and sites. Also, it is not currently possible to insert the banner on PDF documents using this approach.

Another important aspect to differentiate between the archived and the live version is how pages appear in search engine results. In our case, archived version have the description [Archived Content:] in the title tag, in order to help the Web audience make sense of material return through Search.

5. 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 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 Government Websites database and an automated crawling process. 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.

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 labeling of archived material. Redirection will also bring enormous benefits to the user of the Web, with its potential to bring the Web Archive to a more diverse audience.

.

6. REFERENCES

[1] C.O.I. Digital Policy Manager, "COI - Web Standards and Guidelines," Mar. 2009.

[2] A. Spencer, J. Sheridan, D. Thomas, et D. pullinger, "UK Government Web Continuity: Persisting Access through Aligning Infrastructures," The International Journal of Digital Curation, vol. 4, Jun. 2009.

[3] J. Masanès, "Web Archiving: issues and methods," Web Archiving, J. Masanès, éd., Springer Verlag, 2006.

[4] G. Mohr, M. Kimpton, M. Stack, et I. Ranitovic, Introduction to Heritrix, an archival quality web crawler, Bath (UK): 2004.