



Approaches to the preservation of web sites

Brian Kelly

UKOLN, University of Bath, UK

Abstract

Web sites have a tendency to disappear which can result in the loss of valuable scholarly and cultural resources. Although there are a wide range of tools available for mirroring web resources there is still no clear agreement on the strategies which should be adopted for the preservation of web sites - a process fraught with technical and resourcing difficulties, as well as legal and copyright issues. This paper describes various approaches to the preservation of web site and outlines some of the technical challenges. Recommendations on best practices for web site managers and funding bodies are given.

Introduction

The importance of the preservation of web resources can be gauged from the often quoted statistic that the average life of a web page is 100 days ^[1] ^[2], although other sources give an estimate of 40 days ^[3]. The apparent short lifetime of web resources may seem to contradict the expectations of information professionals. Although storage space is getting cheaper and the number of web resources indexed by search engines continues to grow, web users experience 404 error pages far too often. Web sites are reorganised for a variety of reasons - organisations merge, change their name or have an internal reorganisation; the technologies used to provide the web services change - but even if resources are still available, the change of URL makes it appear that the Web page has disappeared.

We might expect digital information to be easier to preserve than physical resources. In practice it can be very easy to permanently delete digital resources: by issuing a single command a system administrator could delete an entire large-scale web site. In comparison destroying books in a library is a very difficult and expensive operation. And, of course, multiple copies of physical resources such as books, journals and videos normally exist.

This paper considers the issues of the preservation of web sites from a number of perspectives: the individual who creates web resources; bodies which fund projects which create web sites; organisations which host web sites; and, on a large scale, national and international initiatives which may have a broader remit for ensuring the long-term preservation of digital resources.

Why do web resources disappear?

Let us try to answer this question by looking at a case study from the UK Higher Education community.

eLib case study

The eLib programme ^[4] was funded by the Joint Information Systems Committee (JISC) ^[5] to develop aspects of a digital library for the UK higher education community. Over 70 projects were funded by eLib, the majority of which hosted a Web site. Web sites were typically hosted by the lead university within a project consortia. Some projects may have had a short-term focus or explored unproductive areas, or the approaches taken by projects may have been superseded by other developments. Other projects may have provided significant findings (the CEDARS project which addressed the area of digital preservation is worthy of note ^[6]) or may have evolved from a fixed-term project into an ongoing service (such as the SOSIG subject gateway ^[7] and the Netskills training organisation ^[8]).

Even for those projects which had negative findings to report it is still desirable for the project reports to continue to be available. As the saying goes "those who forget the lessons of history are doomed to repeat it". Unfortunately for a number of projects web sites which were provided at one stage are now no longer available. As reported in *Ariadne* in January 2001:

“Of the 71 project Web sites which were surveyed, 3 Web sites (PPT, ACORN and ERIMS) are no longer available (i.e. the domain no longer exists) and 2 Web entry points (On Demand and TAPIN) are no longer available (i.e. the domain exists but not the project entry point)” [9].

It should be pointed out that a recent revisit of the web sites revealed that the ACORN web site [10] is now available again. However it was also found that the entry points for the SEREN [11], HERON [12], ResIDe [13], Stacks [14] and SKIP [15] projects and the domain for the NetLinkS [16], ERIMS [17], DIAD [18] and JEDDS [19] projects are now no longer available. The entry point for information about the M25 project is no longer available, although the M25 project web site itself is still available [20].

Of the 71 projects which initially provided a web site only 58 still appear to be providing one.

Why does this happen?

Many of the eLib projects were managed by individuals, research groups and service staff with an understanding of the importance of long-term access to scholarly publications and of enhancing access to digital resources. So why have a number of the web sites disappeared or appeared to have disappeared? A number of reasons can be proposed:

- **Web site still exists but the location changed and an update not provided.**

A project may have changed the location of the entry point but have failed to inform the maintainer of central eLib programme web site.
- **Cost of maintaining pages.**

There is a cost to maintaining web resources. This can be particularly true if pages are dynamic, use advanced features or are managed by a content management system. In addition web pages suffer from “link rot”. There is a cost in ensuring that broken links are fixed. If broken links are left, this could be felt to reflect unfavourably on the institution.
- **Change of purpose of web site.**

A web site initially used for scholarly and research purposes may have evolved into a marketing web site and non-marketing resources removed.
- **Removal of content when staff and students leave.**

Content on a web site associated with an individual may be removed when the individual leaves.
- **Concerns over accessibility and legal issues.**

There may be concerns that old web resources may be inaccessible or are potentially in breach of legislation (e.g. the *Data Protection Act*) and are removed as a safeguard.
- **Completion of project.**

Once a project has finished, it may be felt desirable to “tidy up” which could include removal of web pages.
- **Project web site could be seen to be still running.**

A web site for a completed project may give no indication that it has been completed.
- **Project web site has too high a profile.**

A web site for a completed project may be more highly rated in search engines than more significant resources on the Web site.
- **There may be nobody to maintain the project web site.**

Once a project has been completed there may be nobody available to maintain the resource, argue for its preservation, etc.
- **The web site architecture changes.**

The structure of the web site changes, leading to broken links, or the architecture changes through the deployment of a content management system, server-side scripting, etc.

Avoiding disappearing resources

What approaches can be taken to minimise the numbers of disappearing web resources? A number of approaches can be taken including contractual agreements, technical solutions, and education.

Contractual issues

When a funding body launches a new programme it is desirable that contracts with projects should include a statement about the long term availability of project web sites. It may also be felt necessary for projects to store key deliverables in a central repository. It may also be appropriate for funding bodies to actively monitor project web sites, so that they receive notification from an automated system if a project entry point ceases to exist. If this happens, there may be time to negotiate further access to the resource and to ensure that key resources are deposited in a secure location.

Technological issues

There will be technological factors which will affect the long-term persistency of web resources.

URI naming conventions

As described in "*Guidelines for URI Naming Policies*"^[21] it is desirable to adopt URI naming conventions which will remain persistent if web sites are reorganised or new technologies of file formats are deployed. Ideally URIs will be independent of the technology which is used to provide access to the resources. For example (<http://www.foo.org/key-documents/policies.asp>) is, in all probability, processed by Microsoft's ASP (Active Server Pages) technology. If the organisation decides to move to an alternative technology (PHP, Java Server Pages, etc.) it is likely that the URI will change. This can be avoided in a number of ways, including making use of directory name defaults: (<http://www.foo.org/key-documents/policies/>).

URIs will also ideally be independent of the file format of the resource. For example if a reference to a paper is cited as: (<http://www.foo.org/key-documents/policies.pdf>) is in Adobe PDF format. If the resource is migrated to an alternative format the link will either break or will not provide access to alternative resources. A better approach to citation is to always cite a neutral file format such as HTML: (<http://www.foo.org/key-documents/policies/>) which can then contain links to appropriate formats.

It is also desirable to map dynamic URIs to static URIs. For example, messages stored in the Mailbase Web archive had the format (<http://www.mailbase.ac.uk/lists/web-support/1998-02/0001.html>). As well as being memorable due to its simple, easily understood structure such resources can be easily mirrored. However the current JISCMAIL service has a web archive in which messages have the format (<http://www.jiscmail.ac.uk/cgi-bin/wa.exe?A2=ind0208&L=web-support&D=1&T=0&O=D&F=&S=&P=11547>). Such URIs are difficult for use by individuals and are also difficult to mirror. In this case it should be possible to cover the JISCMAIL format to one similar to Mailbase's using an Apache rewrite directive.

Mirroring

As well as adoption of appropriate URI naming conventions it is also desirable that web site administrators take appropriate steps to ensure that their web site can be mirrored with the minimum of difficulty. Ideally web site administrators will attempt to mirror their Web site to another location and then run validation and checking tools to identify areas in which the mirroring failed.

Since mirroring tools make use of automated robot software they would normally be expected to obey the Standard for Robot Exclusion (SRE)^[22]. The SRE is a simple mechanism which enables web site administrators to prevent robot software from accessing the web site or portions of the web site by use of a robots.txt file in the root of the web site. Web site administrators normally either do not provide a robots.txt file or allow most robots to access the web site, except for certain areas such as a directory of images (in which it has traditionally been felt that such directories do not need to be indexed). However there is a danger that restricting *indexing* robots from such directories will hinder mirroring tools. It would be advisable for organisations to rethink their policies on use of the SRE, if they have one.

Mirroring a web site can be difficult since a web site may be composed of elements from a range of locations. For example, a project web site is likely to contain images of the logos of the host organisation, funding body, etc.

A web site is likely to contain not only static resources but also dynamic features, ranging from simple search facilities to bulletin boards, online voting facilities, personalised features, news and other dynamic content, etc. Static resources can be mirrored relatively easily using a harvesting approach. However this approach will not work with server-side software since a harvester can only access the output of software which runs on the server and cannot mirror the software itself.

Since it will be difficult to mirror the full functionality of web sites which provide such dynamic features, it may be felt useful to investigate how such web site will function after mirroring.

As well as the content of a web site itself, it is also necessary to define the extent of a web site. Web sites are often defined by their domain, for example, (<http://www.bath.ac.uk/>) is the main University of Bath web site. However a project web site may well be hosted on an institutional web site. In such cases, it is necessary to define the extent of the project web site. In the absence of mature standards for defining web collections this can

only be done using directory names, for example, (<http://www.foo.com/projects/preservation/>). In this case we will expect all resources beneath the /preservation directory to be part of the project web site. With this approach, the project web site can normally be easily mirrored.

If the entry point for the project is located above the /preservation directory, such as (<http://www.foo.com/projects/preservation.html>), automated mirroring tools will be unable to distinguish between directories beneath /projects and so will mirror not only resources in the /preservation directory but other directories as well.

Educational issues

Perhaps most importantly is the need for education. There is a need for funding agencies, institutions, authors and web managers to be aware of the dangers of losing valuable scholarly resources and to ensure that approaches which minimise these dangers are adopted, including those mentioned above.

Approaches to preservation

We have provided an example of disappearing scholarly resources, explained why this can happen and provided some suggestions on approaches to minimising such difficulties. However there will still be a need for web resources to be preserved. What approaches can be taken to this task?

Harvesting

An approach which has been touched upon in this article is harvesting web sites. This tends to be the approach favoured by the IT community who regard the issue as one which can be solved by computational power, software and large amounts of disk space. This approach has the advantage that it can be carried out automatically. It is not necessary to rely on individuals to deposit resources. This is likely to be the only feasible approach in cases in which a project has completed and no staff are available to take part in manually depositing selected resources.

Depositing

An alternative approach is for web site owners or other parties to deposit selected resources into a repository. This is an approach which tends to be favoured by the library and archiving communities, as it provides a mechanism for section of quality resources and for providing appropriate metadata. With this approach it may be possible to ensure that a functional web service is deposited, and not just the output of the service as is the case with the harvesting approach.

National approaches

A number of countries are developing strategies for the preservation of national web resources. A brief summary of a number of initiatives is given below.

Australia

PANDORA^[23] provides an archive of selected, significant Australian web sites and web-based online publications. PANDORA aims to retain 'look and feel' of publications:

"... we should strive to capture not just the content but the look and feel of digital publications. Many contain software plug-ins or dynamic features that make this one of the most challenging aspects of our work."

It has an emphasis on selectivity rather than comprehensiveness: "We currently do not attempt to capture everything published online" and it is creating records for the National Bibliography^[24].

PANDORA uses a harvesting approach, based on the HTTrack tool^[25]. Although this has proved a useful tool it has a number of shortcomings, with difficulties in mirroring resources containing multimedia, JavaScript and frames. In addition it does not record the MIME type of harvested resources. Following an evaluation of a number of tools it was decided to continue to make use of HTTrack in conjunction with the Teleport Executive software^[26], which was particularly useful for mirroring database-driven web sites and handling sites containing Java applets and image maps.

Finland

The first round of harvesting of Finnish web space was completed in June 2002 [27]. The archive consists of 11.7 million files, from 42 million URLs. Duplicate entries have been removed and the archive has been compressed to a size of 401 GB. The most popular file formats are HTML (48%), GIF (25%), JPEG (20%), PDF (3%) and other (4%). The approach in Finland has been to harvest the .fi domain, and also Finnish servers which host domains such as .com and .org. In additional linguistic analysis techniques were used to identify web sites written in Finnish.

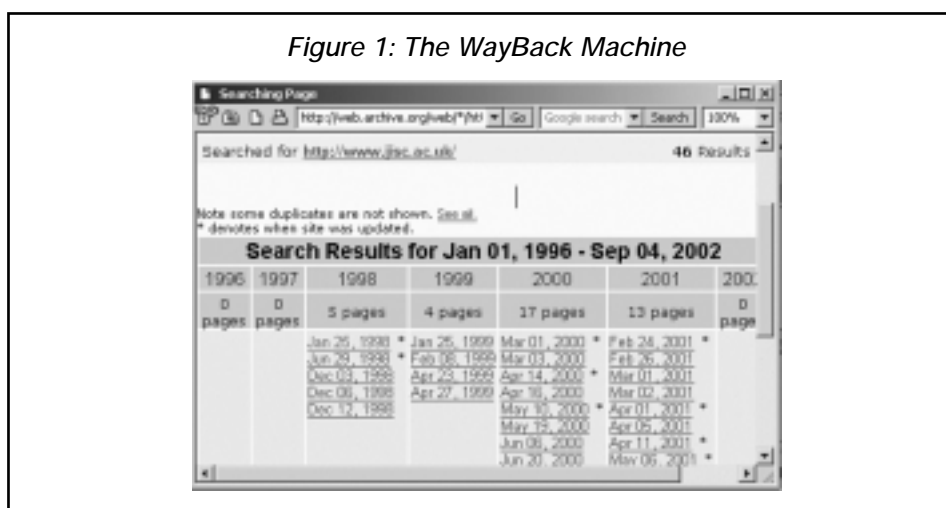
Sweden

The Royal Library (KB), Sweden's National Library is running a project known as Kulturarw3 [28] which aims to collect, preserve and make available Swedish documents from the internet. Kulturarw's approach, which is similar to that of the Internet Archive Foundation, is to preserve everything with the aid of computer technology [29].

Internet Archive

Perhaps the most interesting approach to the preservation of web sites is being taken by the Internet Archive [30]. The Internet Archive is a non-profit-making body based in San Francisco, USA. It was launched in 1996 and it currently stores over 150 terabytes of data.

The Internet Archive's Wayback Machine provides public access to more than 10 billion pages stored in the Internet Archive's web archive. Use of the WayBack Machine for accessing archived copies of the JISC web site is illustrated in Figure 1.



The Internet Archive takes a pragmatic approach to concerns over copyright infringement and other legal concerns. The service supports the Standard for Robot Exclusion. Web site owners who wish to remove their web site from the archive can do so by updating their web site's robots.txt file. In addition they will respond to requests for removal of resources provided they receive an official request.

Challenges of the harvesting approach

A harvesting approach will be needed to ensure that resources which have no responsible owner are preserved. A harvesting approach can also be used to preserve selected resources. It is therefore useful to consider some of the challenges of an automated harvesting approach to the preservation of web pages.

Scoping

In an automated trawl it will be necessary to define the extent of the trawl. For funding bodies such as JISC, Resource, etc., this is likely to be the web sites of funded projects. From a broader perspective a national body, such as the British Library or the Public Records Office (PRO) may have an interest in preserving the entire UK web site.

Let us briefly consider the challenges of preserving UK web space. The first question which needs to be asked is "what do we mean by UK web space?" This could have several meanings:

- **Web sites with a .UK domain name**

This is too simple a definition as it will ignore UK web sites which use domain names such as .org, .com, etc.

- **Web sites hosted in the UK**

This will be more difficult to identify - it will include overseas web sites which are hosted in the UK and UK web sites which are hosted in the UK.

- **Web sites owned by UK organisations**

This will be more even difficult to identify as ownership details can be difficult to process in an automated way.

- **Web sites containing significant British intellectual content**

This may well be impossible to identify.

Web site configuration

The configuration of a web site may affect the ease by which it can be mirrored. Some examples are given below.

The Standard for Robot Exclusion

The Standard for Robot Exclusion (SRE) provides a mechanism for web site owners to request that robot software should not access or index areas of the web site. There has been a tendency for web site owners to prohibit robots from accessing directories which contain resources which may not be usefully indexed (such as those containing images) in order to minimise the load on their web site. However this could mean, for example, that well-behaved mirroring software which supports the SRE will not mirror images.

File formats and MIME types

Standard HTML resources with static URIs are particularly suited for mirroring. For other file formats which are accessible on a web site there may be difficulties in ensuring that the mirrored resource can be accessed. When a web resource is delivered, a MIME type for the resource is sent in the HTTP header. The MIME type is determined by the server configuration file, normally by associating a file extension with a particular MIME type. The MIME type is used by the web browser to determine which application should render the resource. If a mirrored site does not make use of a MIME type access to mirrored resources may be difficult since only the file extension will be known.

There may also be difficulties in accessing different versions of file formats, since the MIME type does not include version details.

External resources for HTML files

HTML resources themselves often contain embedded resources such as images. Increasingly other types of files may be embedded, such as JavaScript and CSS files. The link mechanism for such resources is well established (use of the <LINK> element). However there may be addition be other resource types for which the link mechanism is not widely deployed.

Generated resources

Through use of client-side scripting languages such as JavaScript allows HTML resources to be generated dynamically. Unless the mirroring service provides a JavaScript environment and the code is executed the resource, and links from the resource, will not be mirrored.

Redirects

Many web sites make use of redirects, so that when a user goes to one resource he is taken to another. Typically redirects are used when an URI is published or following a reorganisation of a Web site. Redirects can be implemented in a number of ways including use of the <META> element or JavaScript within HTML documents, or through the web server configuration file, which provide a number of ways in which resources can be redirected. Given this complexity there may be a danger that mirroring software may not satisfactorily mirror redirected resources.

Dynamic web sites

Dynamic web sites pose particular challenges for mirroring. The term 'dynamic' can be applied to a number of distinct features of web sites: use of server-side management tools (e.g. PHP or ASP scripting); use of animated features (e.g. animated GIFs or dynamic HTML); or personalised and customised resources. Server-side management tools can create resources which can easily be mirrored and animated GIFs and dynamic HTML will normally be mirrored satisfactorily. Personalised and customised resources may prove more difficult. The customisation may be influenced by a number of factors, such as the client machine (PC or PDA) or browser

(Internet Explorer, Netscape or Lynx) or factors such as the time of day, location of the end user, navigational path, etc. Personalisation is normally provided by user selection or by factors which are known about the user. Personalisation will often require use of technologies such as cookies or registration.

Loops

There is a danger that mirroring software may be caught in a loop. This can happen with CGI resources or through malformed relative links.

Web collections

Collections of web resources can be identified by automated tools through use of the domain name or directory name.

Future developments

There are a number of areas in which developments could affect strategies for mirroring.

The Standard for Robot Exclusion has not been developed significantly since it was released. It currently allows robots to be managed only by their name. Resources can be managed at a directory level. An enhanced standard could usefully provide much richer control, including concept of collections of related resources, managing access by type of robot, control, etc.

Resources which form part of the 'invisible web' are difficult to access by robot software. A number of approaches to disclosing metadata for such resources have been taken, including the Open Archives Initiatives. It would be useful to investigate the potential of this type of approach to disclosing 'invisible web' resources to mirroring tools.

It would be useful to provide a test bed containing examples of resources which are known to be difficult to mirror which could be used for the evaluation of mirroring tools.

Conclusions

Web sites are disappearing and there is a grave danger that invaluable scholarly, cultural and scientific resources will be unavailable to future generations. In many instances this is due to organisational and human factors, compounded by technical challenges. There is a need to provide education for various players in order to bring about a cultural change so that deleting web resources is not seen as a regular maintenance procedure. There may need to be changes in contractual agreements and it is likely that there will be increased costs.

However, even if best practices become more widely adopted, there is likely to be a continued need for web sites to be preserved. We are likely to see a mixed economy, with subject specialists preserving resources within their own subject area, organisations and funding bodies preserving resources which they have a clear responsibility for and national and international bodies taking a broader approach. There is a need for owners of web sites and bodies which fund the development of web sites to ensure that consideration is given to the long-term availability of the web sites.

References

- 1 "Wayback Goes Way Back on Web", Wired News, 29 Oct 2001, (<http://www.wired.com/news/culture/0,1284,47894,00.html>)
- 2 *Seeing The Future In The Web's Past*", BBC News, 12 November, 2001, (http://news.bbc.co.uk/1/hi/in_depth/sci_tech/2000/dot_life/1651557.stm)
- 3 *"Digital Sexualities: A Guide To Internet Resources"*, Theory.org, (<http://www.theory.org.uk/ctr-que6.htm>)
- 4 *"eLib: The Electronic Libraries Programme"*, UKOLN, (<http://www.ukoln.ac.uk/services/elib/>)
- 5 *"Joint Information Systems Committee (JISC)"*, JISC, (<http://www.jisc.ac.uk/>)
- 6 *"CEDARS Project"*, University of Leeds, (<http://www.leeds.ac.uk/cedars/>)
- 7 *"SOSIG: Welcome"*, ILRT, University of Bristol, (<http://www.sosig.ac.uk/>)
- 8 *"Netskills: Quality Internet Training"*, University of Newcastle, (<http://www.netskills.ac.uk/>)
- 9 *"WebWatching eLib Project Web Sites"*, Brian Kelly, Ariadne, issue 26, January 2001, (<http://www.ariadne.ac.uk/issue26/web-watch/>)

- 10 "ACORN: Access to Course Reading via Networks",
(<http://acorn.lboro.ac.uk/>)
- 11 "SEREN: Sharing of Educational Resources in an Electronic Network in Wales", UKOLN,
(<http://www.ukoln.ac.uk/services/elib/projects/seren/>)
- 12 "HERON: Higher Education Resources ON-demand", UKOLN,
(<http://www.ukoln.ac.uk/services/elib/projects/heron/>)
- 13 "ResIDe: Electronic reserve for UK Universities", UKOLN,
(<http://www.ukoln.ac.uk/services/elib/projects/reside/>)
- 14 "The Electronic Stacks Project", UKOLN,
(<http://www.ukoln.ac.uk/services/elib/projects/stacks/>)
- 15 "SKIP: SKills for new Information Professionals", UKOLN,
(<http://www.ukoln.ac.uk/services/elib/projects/skip/>)
- 16 "NetLinks: Collaborative Professional Development for Networked Learner Support", UKOLN,
(<http://www.ukoln.ac.uk/services/elib/projects/netlinks/>)
- 17 "ERIMS: Electronic Readings in Management Studies", UKOLN,
(<http://www.ukoln.ac.uk/services/elib/projects/erims/>)
- 18 "DIAD: Digitisation in Art and Design", UKOLN,
(<http://www.ukoln.ac.uk/services/elib/projects/diad/>)
- 19 "JEDDS: Joint Electronic Document Delivery Software Project", UKOLN,
(<http://www.ukoln.ac.uk/services/elib/projects/jedds/>)
- 20 "M25 Link", UKOLN,
(<http://www.ukoln.ac.uk/services/elib/projects/m25/>)
- 21 "Guidelines For URI Naming Policies", Brian Kelly, Ariadne, issue 31, March 2002,
(<http://www.ariadne.ac.uk/issue31/web-focus/>)
- 22 "A Standard for Robot Exclusion", Martijn Koster,
(<http://www.robotstxt.org/wc/norobots.html>)
- 23 "Pandora Archive: Preserving and Accessing Networked Documentary Resources Of Australia",
(<http://pandora.nla.gov.au/index.html>)
- 24 "Towards a Preserved National Collection of Selected Australian Digital Publications", Colin Webb, National Library of Australia, Preservation 2000: An International Conference on the Preservation and Long Term Accessibility of Digital Materials,
(<http://www.rlg.org/events/pres-2000/webb.html>)
- 25 "HTTrack Website Copier - Offline Browser",
(<http://www.httrack.com/>)
- 26 "Teleport Exec", Tennyson Maxwell Information Systems Inc.,
(<http://www.tenmax.com/teleport/exec/home.htm>>
- 27 "Harvesting of the Finnish Web space completed", Juha Hakala, 19 August 2002, email message to web-archive mailing list
- 28 "Kulturarw",
(<http://www.kb.se/kw3/ENG/>)
- 29 "The Kulturarw3 Project - The Royal Swedish Web Archiw3e - An example of "complete" collection of web pages", 66th IFLA Council and General Conference, Jerusalem, Israel, 13-18 August 2000,
(<http://www.ifla.org/IV/ifla66/papers/154-157e.htm>>
- 29 "Internet Archive",
(<http://webdev.archive.org/>)

Contact

Brian Kelly
 UK Web Focus
 UKOLN
 University of Bath
 Bath BA2 7AY
 UK

B.Kelly@ukoln.ac.uk
 www.ukoln.ac.uk