



A Guide to Web Preservation

**Practical advice for web and records managers
based on best practices from the JISC-funded
PoWR project**

**Edited by Susan Farrell using content written by Kevin
Ashley, Richard Davis, Marieke Guy, Brian Kelly, Ed Pinsent
and Susan Farrell**

A note on the authors

This Guide has contributions from a range of people but was primarily written and reviewed by the JISC PoWR team:

University of London Computer Centre (ULCC), Senate House, South Block, Malet Street, London WC1E 7HU (<http://www.ulcc.ac.uk/>)

- Ed Pinsent
- Richard Davis
- Kevin Ashley

UKOLN, University of Bath, Bath BA2 7AY (<http://www.ukoln.ac.uk/>)

- Brian Kelly
- Marieke Guy

This Guide was edited by:

Susan Farrell Consulting Ltd (<http://www.farrellconsulting.co.uk/>)

- Susan Farrell

Acknowledgement

Cover image of a website structure provided courtesy of E J Fox.

Other information

This publication can be purchased from Lulu.com via the url:

<http://www.lulu.com/uk/buy/>

ISBN 0-9516856-7-8

Copyright © 2010 UKOLN / ULCC

Provided under a Creative Commons Licence

(Attribution - Non-commercial - Share Alike)



Table of Contents

Introduction	1
Chapter 1: What is preservation?	2
Chapter 2: What are web resources?	6
Chapter 3: Why do I have to preserve them?	10
Chapter 4: What is a web preservation programme?	14
Chapter 5: How do I decide what to preserve?	18
Chapter 6: How do I capture them?	25
Chapter 7: Who should be involved?	28
Chapter 8: What approaches should I take?	33
Chapter 9: What policies need to be developed?	37
Case studies	
Home page history	9
Institutional use of Twitter	13
Vanishing domain names	24
Student blogs	27
Capturing wiki contents	32
Appendices	
A: Further reading	39
B: The JISC-funded PoWR project	40
C: Tools for capturing	41
D: An introduction to records management	46
E: An introduction to web management	47
F: Legal matters	48
Glossary	50
Index	55

Introduction

The Preservation of Web Resources project (JISC PoWR) was funded by the JISC in order to identify emerging best practices for the preservation of web resources. The project was provided by UKOLN and ULCC and ran from April through to November 2008. A number of workshops were organised to help identify emerging best practices, and a blog was established to raise awareness of this work and to gain feedback on the approaches being taken by the JISC PoWR team.

The project handbook was published in November 2008. Since then we have seen a growing awareness of the importance of digital preservation in general and in the preservation of web resources (including web pages, web-based applications and websites) in particular. The current economic crisis and the expected cuts across public sector organisations mean that a decade of growth and optimism is now over – instead we can expect to see reduced levels of funding available within the sector which will have an impact on the networked services which are used to support teaching and learning and research activities.

The need to manage the implications of these cutbacks is likely to result in a renewed interest in digital preservation. We are therefore pleased to be able to publish this new guide, based on the original *PoWR: The Preservation of Web Resources Handbook*, which provides practical advice to practitioners and policy makers responsible for the provision of web services.

Brian Kelly, UKOLN

Project Director, JISC PoWR project

Chapter 1 – What is preservation?

Summary:

- Definition of ‘web preservation’.
- To ensure that everyone in the Institution agrees on what should be preserved and how, a web preservation programme should be developed.
- All resources must be managed in order to preserve them.
- There are issues to bear in mind which are specific to web resources, Web 2.0 resources and content management systems.

For the purposes of this Guide, we define web preservation as ‘the capture, management and preservation of websites and web resources’. Web preservation must be a *start-to-finish* activity, and it should encompass the *entire lifecycle* of the web resource.

Another definition to consider in this context is JISC’s definition of ‘digital preservation’ - ‘the set of processes and activities that ensures long-term, sustained storage of, access to and interpretation of, digital information’.

Institutional views of preservation requirements may vary so it is important for an Institution to agree on a web preservation programme which defines the web resources which will be preserved. When considering this bear in mind that:

- Resources must be managed in order to preserve them.
- Preservation will not apply to all web resources: a selective approach is recommended.
- Preserving every version of every resource is not always necessary.
- Permanent preservation (as defined by the OAIS model) is not the only viable option. Short-term protection of a resource from loss or damage is an acceptable form of preservation.
- Preservation actions do not have to result in a 'perfect' solution.

When considering web resources there are a number of specific preservation issues which apply. In addition, Web 2.0 and content management systems present unique issues.

Web resource preservation issues

1 - Frequency of change

Web resources change to a greater or lesser extent every day, and periodically change dramatically because of events such as re-branding, the implementation of a content management system, or changes to content providers.

2 - Quantity and range of resources

The quantity and range of resources potentially needing preservation are so large it is vital to: know what resources there are; where they are; and what to do about them.

A Guide to Web Preservation 2010

3 - Continuity

Because of the ease with which websites and pages can be edited, the possible impact on users expecting continuity in web resources can be overlooked. For example, a page may stay the same, but no longer be available from the same URL, or it may remain at the same URL but its content changes. So the issues are: persistence of resources at a given URL; and persistence of resources within a domain.

Ideally it should be possible to support versioning across a whole site, so that old versions of a page link to their associated contemporary versions, but this represents a large overhead.

4 - Integrity of web resources

Websites and pages need to be protected from careless or wrongful amendment, deletion, or removal, whether by malevolent hackers/crackers, or well-intentioned Institutional staff.

5 - Ownership

There may be issues of ownership resulting from web resources being managed by many different departments or members of staff, or by sub-sites sometimes being temporary or ad hoc (for example, a project site).

6 - Databases and deep websites

Databases present particular issues because preserving an underlying database may not preserve the user's experience on the web. Also database-driven websites are not always easy to capture by remote harvesting.

7 - Streaming and multimedia

The quantity and quality of data, and the range of formats, can cause issues when dealing with multimedia. In addition, these resources can be hosted elsewhere and therefore the same set of issues applies as for Web 2.0 applications (see below).

8 - Personalised websites

Some websites offer users customisable features. This raises the issue of whether every possible combination of every user's custom view should be preserved.

9 - Appraisal and selection

Appraising and selecting which web resources should be preserved raises many questions which are dealt with in Chapter 5.

10 - Providing access

Once preserved it has to be considered how access will be provided to the web resources and how to deal with issues of IPR and ownership.

11 - Resources for preservation

Both personnel and technical resource issues also have to be considered. Preservation work can be an overhead on day-to-day web and records management activities so assigning people to the preservation work needs to be balanced with routine web and records management.

In technical terms, it is necessary to estimate how much storage space will be required to store the old web resources and where this will be located.

Web 2.0 preservation issues

The two most important issues with Web 2.0 software and applications are ownership and retention.

1 - Ownership and responsibility

Often individuals create and manage their own Web 2.0 resources such as external (personal) accounts for Flickr, Slideshare or Wordpress.com. So it is possible for academics to conduct a significant amount of Institutional business outside any known Institution network. In these cases, the Institution either does not know this activity is taking place, or ownership of the resources is not recognised officially. In such a scenario, it is likely the resources are at risk.

2 - Retention of 'master copies'

Third party sites such as Slideshare or YouTube are excellent for dissemination, but they cannot be relied on to preserve materials permanently. So, if a resource is created on one of these third party sites and it requires retention or preservation, arrangements must be made within the Institution for the 'master copy'.

Content management system preservation issues

With digital preservation in mind, the features of particular value which content management systems (CMSs) may offer are:

- Version control - when changes are made to items in the CMS, the previous version is kept.
- Change logging - when changes are made to items in the CMS, the system records who made the change and when.
- Rollback/reversion - the facility to restore the website, or a part of it, to a previous state.
- Creating a snapshot of the website at a particular point in time.

Many CMSs offer one or more of these features but the extent to which they can easily be used to reinstate older versions of a website, or find what changes happened when, varies dramatically. Version control information is easy to create and store, but less easy to put to practical use. Discussions with web managers suggest that these features are rarely tested very vigorously.

The particular preservation issues of CMSs are:

- Page names and numbers.
- Rollback function is limited.
- Lifespan of system.
- Compatibility between systems.

Page names and numbers

Some CMSs may present problems to a remote harvesting engine, or crawler, as pages that are identified with numerical tags instead of page names, for example, may not be recognised, and hence may be missed by the remote harvester. This is especially true if the CMS generates pages dynamically. The severity of this behaviour may also depend on how the site was built in the first place.

A Guide to Web Preservation 2010

Rollback function is limited

A rollback may not be the same as restoring a full snapshot as it will tend to focus on a particular page or content element, but not its entire context. Web pages usually have many objects that they relate to - for example embedded images and stylesheets - so the rollback cannot be used to view the content of the whole page as seen by the user. The content is held in the database as layers of time-stamped pages and a script is required to retrieve it. It is therefore not clear to what extent the rollback functions and version control tools produce useful, tangible outputs that could be captured, managed or preserved.

Compatibility between systems

A CMS may not be supported indefinitely so the question arises about whether the new version will be compatible with the old version. Also, the Institution may decide to change the CMS and, as CMS internal management of content, data and metadata tends to be application-specific, this may mean that moving large quantities of interlinked website content between CMS packages is likely to be a manual and intensive process.

Backing up is not enough

A CMS is a database full of content, but simply backing up the database will not constitute preservation of the content. The backup action would capture a change history of the website for as long as it was kept in that CMS; it would not constitute a usable collection of page snapshots, or an archived website.

Metadata

The change history metadata would be extremely useful for records management and preservation purposes, but access to that metadata is not guaranteed: it would need to be exportable in a form that could be preserved.

Action

- Define web preservation in the context of your Institution.
- Consider to what extent the issues raised by Web 2.0 and content management systems affect your Institution.

Chapter 2 – What are web resources?

Summary:

- Web resources are those delivered by a web browser, and can be found on web servers, in managed systems (including content management systems, Institutional repositories and digital collections), and less well-managed systems (such as Web 2.0 applications and services).
- To help with determining whether they should be preserved, web resources should be categorised into records, publications or artefacts.

Web resources are those delivered by a web browser so they may appear not only on the Institutional website but also on other websites and web systems. Some examples are:

- Institutional and departmental records, with legal and business requirements governing their retention and good maintenance.
- Content affecting students, such as prospectuses and e-learning objects.
- Administrative, research, teaching and project outputs.
- Evidence of other activities (e.g. conferences).

Many of the resources of an Institutional website may be stored in a CMS or within well-established managed systems, such as:

- Those for managing assessments and examinations.
- Online libraries.
- Virtual Learning Environments (VLEs).
- Online teaching courses and course content.
- Institutional repositories.
- Digital collections used for study.
- E-learning objects and teaching materials.
- E-portfolios.

In these cases, the resources are simply being accessed or delivered by a web browser so, except for those in a CMS, the preservation should be of the system rather than the pages as they appear on the website. In the case of a CMS both the system and the pages produced need to be preserved.

But the increasing use of Web 2.0 services and applications means that many resources exist in less well-managed systems, many of which are hosted outside the Institution, including:

- Blogs Blogger (e.g. Wordpress, Edublogs, Warwick Blogs).
- Wikis (e.g. Mediawiki, Wetpaint, Tiddlywiki).
- Social bookmarking (e.g. Delicious, CiteULike, Connotea).

A Guide to Web Preservation 2010

- Media sharing services (e.g. Flickr, Slideshare, YouTube. (Scribd, DeviantArt)).
- Social networking systems (e.g. Facebook, Twitter, Ning, Elgg, Crowdvine, LinkedIn).
- Collaborative editing tools (e.g. Google Docs).
- Syndication and notification technologies (e.g. Netvibes, Technorati).
- Instant messaging (e.g. Facebook Chat, Google Chat, Skype, Jabber, Windows Messenger).

Which web resources need to be preserved?

To determine which web resources to preserve it is helpful to consider which of three categories best describes the resource – record, publication or artefact. If it is a record or a publication, the resource should be considered in the context of existing policies and procedures for these types of document.

A record

‘Recorded information, in any form, created or received and maintained by an organisation or person in the transaction of business or conduct of affairs and kept as evidence of such activity.’

(http://www.recordsmanagement.ed.ac.uk/InfoStaff/RMstaff/RM_framework.htm)

A web resource can be considered a record if it:

- constitutes evidence of business activity that needs to be referred to again;
- is evidence of a transaction;
- needs to be kept for legal reasons.

Some examples:

- The website contains the only copy of an important record. Web resources should not be removed or deleted without establishing if they are the only copy.
- The website, or a set of web pages, in itself constitutes evidence of Institutional activity. The history of this evidence is visible through the various iterations and changes of the website.
- The website is in itself evidence of the publication programme, or has such evidence embedded within its systems. If it is necessary to show, as evidence, that the Institution published a particular document on a certain date, then the logs in the CMS constitute an evidential record. In some cases, this may be needed to protect against liability.
- A transaction of some sort that has taken place through the website (transaction does not just mean money has changed hands). If these transactions need to be kept for legal or evidential reasons, then they are records too. The transaction may generate some form of documentation (e.g. automated email responses), which may in turn need to be captured out of the process and stored in a place where it can be retrieved and accessed.

A Guide to Web Preservation 2010

A publication

‘A work is deemed to have been published if reproductions of the work or edition have been made available (whether by sale or otherwise) to the public’. (National Library of Australia <http://www.nla.gov.au/services/ldeposit.html>)

A web resource might be considered a publication if it is:

- a web page that is exposed to the public on the website;
- an attachment to a web page (e.g. a PDF or Word Document) that is exposed on the website;
- a copy of a digital resource, e.g. a report or dissertation that has already been published by other means.

Some examples:

- Websites containing the only copy of an important publication.
- Web pages constituting a version of information that is available elsewhere. By version, it is meant that it has been rendered in some way to bring it into the website. This rendering may include, for example, the addition of navigation elements that make it different from the original source.
- A web page constituting a mix of published information. For example, a page of original Institutional material combined with an RSS feed from outside the Institution.

An artefact

‘Anything else that isn't a record or a publication by the above definitions, but which is still worth preserving, can be understood as an artefact’.

A web resource might be considered an artefact if, for example, it:

- has intrinsic value to the Institution for historical or heritage purposes;
- is an example of a significant milestone in the Institution's technical progress, for example the first instance of using a particular type of software.

Preserved artefacts could include image collections (still and moving), databases, e-learning objects, digitised objects or research objects.

Actions

- Investigate the web resources in your Institution so that you can see the range of types and locations, particularly concentrating on those which are more hidden such as the Web 2.0 applications and services. This will be refined following Chapter 5.
- Assess the existing technical infrastructure and technical skills.
- Produce an outline list of the resources to be preserved.

Case Study – Home page history

Scenario: *Your institution is about to commemorate an important event and the Vice Chancellor wants to highlight that the Institution is actively engaging with new technologies, and so would like to provide an example of how the Institution's website has developed since it was launched.*

Issues

- How has your Institutional home page changed over time?
- Have you kept records of the changes and the decisions which were made (and how they were made)?
- If you needed to do this for your Institution, do you feel you would be able to deliver a solution? How far back could you go?

Approaches

- The Internet Archive (see Chapter 7) has taken snapshots of websites since 1996 and may have captured web pages from your Institution. The University of Bath used the snapshots of its home page captured by Internet Archive to illustrate how it had changed between 1997 and 2007: an animated visualisation of the changes, linking to the IA's snapshots, is available on UKOLN's website. However, there is no guarantee that the Internet Archive will have captured every iteration of your Institution's website, nor that the copies it has are complete and fully functional.
- Even if there are few, or no, surviving copies of previous versions of your website, there is no time like the present to start making sure snapshots are kept, either by taking your own copies, or ensuring the Internet Archive takes a copy. You can use an online form to nominate a site for crawling by the Internet Archive. It is also possible to nominate your site for capture by the UK Web Archive (see Chapter 7).
- Another approach is to build a compiled online history. The University of Virginia maintains a web page detailing 14 years of its website history. It includes fascinating statistical information based on analysis of the web server logs. Copies of the website are not available before 1996 and the image of the website in 1996 is taken from the Internet Archive. All subsequent snapshots are hosted on the main U.Va website, in subdirectories (/virginia1999, etc.). Some years are missing: whether because the changes were insignificant, or no copy survives, is not clear. Although there are broken links in the archived sites, or there are anachronistic links to current versions of pages, the archived snapshots provide a valuable view of the evolution of the Institution's web presence.

Chapter 3 – Why do I have to preserve them?

Summary:

- The drivers for carrying out web preservation are strategic, legal, financial, contractual and reputational.
- Web preservation also has a role to play in business continuity planning.
- The espida project provides a useful methodology for quantifying the value of web preservation which is helpful when developing a business case for a web preservation project.

There are many internal and external drivers for undertaking web preservation within an HFE Institution and some of these are summarised below.

Strategic, legal, financial and contractual obligations

Institutional websites contain evidence of Institutional activity which is not recorded elsewhere and may be lost if the website is not archived or regular snapshots taken. This loss could be construed as a threat to the business continuity of the organisation as reference to this content may be required for the checking of strategic, legal, financial, contractual or scholarly information. For example, if certain information is not recorded or protected the Institution is in danger of failing to comply with legal acts such as Freedom of Information and Data Protection. Also the Institution may be breaking contractual and auditing obligations, and putting itself at risk.

Reputational risk

The Institution's reputation can be put at risk by poor website continuity, broken links, and missing resources but more damaging is the risk that website users may make decisions based on misleading or out-of-date content.

Supporting the Institutional Mission

The case can be strengthened for carrying out web preservation by showing what it delivers to support the Institution's mission statement, and its associated strategies in the areas of research, teaching and learning, information, libraries, and records management. A policy for preserved web resources could feasibly assist in supporting Institution-wide aims and generic objectives, even when they do not explicitly mention digital resources by name, for example 'attracting a wide variety of students'.

Risk management and risk analysis

Possible risks associated with websites and web resources include: loss of data, records or resources; failure to be information compliant; the risk of litigation from students or the public; and the risk of breaching copyright.

Bringing web preservation into line with business continuity planning may help to change Institutional practice. The risks associated with possible IPR infringements are put in perspective by Charles Oppenheim's lightweight risk formula: $R = A \times B \times C \times D$, where:

A Guide to Web Preservation 2010

- A: probability that you are illegal;
- B: probability that you are found out;
- C: probability that action will be taken against you;
- D: extent of financial risk.

The aim is to keep all of these values as low as possible, but it is also the case that if any of these is zero, the overall risk is effectively nullified.

If a risk management strategy is required to enable web preservation, using the JISC InfoKit on Risk Management is a good place to start. The kit takes the view that Risk Management is an essential part of project management.

Saving money

Web resources cost money to create and store; failure to repurpose and reuse them would constitute a waste of money. Although web preservation may have an initial cost, once the process has begun the savings can be great. Having a good strategy in place (which means selection, retention, and deletion where appropriate) will save both money and energy in the long run. The website may also contain digital assets and electronic resources - assets which may be of continued value which may increase in value through sharing and repurposing.

Responsibility to staff and students

The Institution has a responsibility to the people who use the resources. Students and staff may make serious choices about their academic careers or their jobs based on website information, and the Institution has a responsibility to make sure a record is kept of the publication programme.

Responsibility to users

The Institution has a responsibility to the people who may need to use the resources in the future. Many of the resources which the Institution publishes are unique, and deleting them may mean that invaluable scholarly, cultural and scientific resources (heritage records) will not be available to future generations.

Gaining a competitive edge

Starting a web preservation programme will make the Institution look 'forward thinking' and therefore give them an edge over their competitors. The Institution could be one of the first to start an official web preservation programme which will be great marketing fodder. Embedding web preservation strategies will also help the Institution think about the continuity of resources, broken links and other aspects of web management.

Quantifying the value of web preservation

The espida project in Glasgow (<http://www.gla.ac.uk/espida/>) offers a useful methodology which could be used to quantify the value of web preservation. It takes a pragmatic view of the way that HFE Institutions operate in the real world recognising that preservation activities will continue to vie with other services for funds.

A Guide to Web Preservation 2010

espida can help:

- demonstrate the value of websites and web resources;
- communicate the intangible benefits of web preservation and web resource preservation to senior management, and articulate those benefits;
- make a case for a web preservation programme, based on a formalised and transparent communication process between the proposer and the funder;
- identify costs and benefits of web preservation, using scorecards and cost templates;
- produce a decision-making process that is transparent and based on all relevant information.

The results of the process will enable the development of a business case which not only answers the question ‘how much does web preservation cost?’, but also ‘why do we need web preservation?’ and ‘why should we spend money on web preservation, rather than on the primary business of the organisation?’.

Action

- Think about the drivers and how they relate to your Institution so that you can develop a business case to get senior management buy-in for your web preservation programme.
- Carry out a risk assessment.

Case study – Institutional use of Twitter

Scenario: *An Institution has set up an Institutional Twitter account for disseminating news on activities and events. The unspoken expectation is that Twitter will be used across the Institution as an individual productivity and social tool. However, one department has become an early adopter of the technology and is using it in teaching and learning, and research contexts. The Head of this department is now suggesting that a formal policy for capture and preservation of Twitter messages be enacted.*

Approaches

Capturing Twitter tweets is straightforward as they are easily downloaded from the RSS feeds that the service generates. These could be converted into text documents or web pages, or even imported into a database or blogging software. A number of solutions are available for tweet preservation such as Twapper Keeper (<http://twapperkeeper.com/index.php>), FriendFeed (<http://friendfeed.com/>), the WordPress Lifestream plugin (<http://wordpress.org/extend/plugins/lifestream/>), What the Hashtag (http://wthashtag.com/Main_Page) and Tweetdoc (<http://www.tweetdoc.org/>). More significant is the need to define an Institutional decision about the value of these resources. Why keep Twitter tweets?

- Corporate record: Is a Twitter a digital resource that information professionals should be interested in capturing or preserving? At what point does a Twitter turn into a record that requires the attention of the records manager?
- Scholarly record: can it be demonstrated that tweets are part of the scholarly record that is not captured in other forms?
- Legal reasons: Is Twitter being used to deliver learning? Is there a legal requirement to record what has been sent, as part of the assessment record?

Possible outcomes from the decision:

- It is agreed that Twitter posts are transient and do not need to be preserved. Records of corporate activity are something Institutions consider within their archiving policies, and tweets could be considered part of that corporate record. The decision needs to be reviewed as some information / communication media may take over the role of others.
- As a recognised official Institution publication, the posts need to be subject to Quality Assurance and editorial processes, which include keeping a record of the posts.
- An informal log of posts is kept, in order to have a record of topics that have been covered, audit the number of posts, be able to identify any significant impact from these services, etc.

Chapter 4 - What is a web preservation programme?

A web preservation programme must involve planning and activities for *all the stages in the lifecycle*. It involves two main phases: analysis and planning; and execution. The table below shows these phases divided into sequential stages, and the tasks included in each stage. The column on the right points to the relevant chapter(s) in this Guide which will provide help.

It may prove more successful for an Institution to manage the individual stages as a series of clearly defined projects, with owners. Indeed, some of the suggested actions fall into the category of generic project management actions, and do not have a corresponding chapter in the Guide.

See also the Programme Timeline on page 16 which may assist with planning.

Analysis and planning phase	
Tasks	Chapter
Stage 1A – Institutional analysis	
Define web preservation and web resources	1, 2
Discuss preservation with relevant staff	7
Set up a working group or project, with an owner	7
Define the legal requirements and Institutional priorities	App F
Review all related strategies, policies and guidelines	9
Stage 1B – Resource analysis	
Assess the existing infrastructure and technical skills	2
Assess the resourcing implications of the programme	3, 7, 8
Consider the desired speed of implementation of the programme, and define its aims, scope and duration	8
Stage 2 – Getting buy-in	
Choose which approach model you will use (in-house, contracted out, collaborative, consortial)	7, 8
Develop a business case and secure buy-in	3

A Guide to Web Preservation 2010

Assess the cost effectiveness of your approach	7,8
Perform a risk assessment	3
Stage 3 – Decide on methodology	
Define policies for the appraisal and selection of web resources, identifying the key functions which need owners or managers	5
Consult stakeholders	5
Decide on selection and capture approaches	6 AppC
Choose preservation tools	6
Determine an access policy	6
Complete web preservation programme contents	All chapters

Execution phase	
Tasks	Chapter
Carry out the appraisal of web resources and identify those to be included.	5
Capture the selected resources	6, App C
Manage the storage and access of the resources, with appropriate metadata	
Develop, and ensure adoption of, policies to embed preservation into the life of the Institution	9
Manage the digital repository where web resources are stored, and take such actions over time as are needed to ensure their continued access and permanency	

Tasks	Timeline (months)								
	1	2	3	4	5	6	7	8	9
Programme timeline (shaded areas indicate timelines for tasks)									
Analysis and planning: Stage 1A – Institutional analysis									
Define web preservation and web resources									
Discuss preservation with relevant staff									
Set up a working group, with an owner									
Define the legal requirements and Institutional priorities									
Review all related strategies, policies and guidelines									
Analysis and planning: Stage 1B – Resource analysis									
Assess the existing infrastructure and technical skills									
Assess the resourcing implications of the programme									
Consider the desired speed of implementation of the programme									
Analysis and planning: Stage 2 – Getting buy-in									
Choose which approach model you will use									
Develop a business case for senior management and secure buy-in									
Assess the cost effectiveness of your approach									
Perform a risk assessment									
Analysis and planning: Stage 3 – Decide on methodology									
Define policies for the appraisal and selection of web resources									
Consult stakeholders									
Decide on selection and capture approaches									
Choose preservation tools									
Determine an access policy									
Complete web preservation programme contents									
Execution									
Carry out the appraisal of web resources and identify those to be included.									
Capture the selected resources									
Manage the storage and access of the resources, with appropriate metadata									
Develop, and ensure adoption of, policies to embed preservation									
Manage the digital repository									

Resources required

The resources required are: staff; hardware and infrastructure; and software.

Staff

Chapter 7 covers which staff should be involved in the programme but it useful to bear in mind that the skills required are:

- technicians who understand websites;
- selection and curator skills (archivists or librarians);
- digital preservation skills;
- permissions management.

Hardware and infrastructure

The following may be required:

- dedicated servers for data storage;
- sufficient web server capacity for delivery of resources;
- high speed and high bandwidth Internet connection for capture;
- sufficient connection bandwidth for users.

Software

The following may be required:

- collection tools (See Appendix C);
- harvesting methodologies;
- database archiving;
- specialised software for access and delivery;
- specialised software for metadata management.

Action:

- Consider the desired speed of implementation of the programme, and define its aims, scope and duration.
- Draw up a task plan with a timeline.

Chapter 5 – How do I decide what to preserve?

Summary:

- The appraisal process looks at the location, use and ownership of the web resources, and particularly aims to ensure that web resources containing unique information are preserved.
- Resources which are managed elsewhere (e.g. in asset collections or Institutional repositories), or are of little or no value, can be omitted from the web preservation programme.
- The MoSCoW method enables prioritisation of resources.
- The selection approach can be unselective (domain harvesting) or selective (criteria-based or event-based).
- Which aspects of resources (information and experience), and which elements (content, appearance and behaviours) to preserve must be considered.

Careful appraisal and selection of web resources will make the task of web preservation more manageable as this will make it clear what should, and should not, be included in the scope of the web preservation programme. Also, within that list, each web resource can be assigned a priority for action.

Some issues to bear in mind when deciding which of the resources to preserve are:

- The Institutional structure and its aims.
- The policies and drivers for preservation.
- The legal record-keeping and audit requirements for the Institution.
- The potential reuse value of resources.
- Is the resource needed by staff to perform a specific task?
- Has the resource been accessed in the last six months?
- Does the resource represent a significant financial investment in terms of staff cost and time spent creating it?

It is particularly important to facilitate the survival of web resources which contain unique information such as those:

- which only exist in web-based form - for example, teaching materials designed as web pages;
- which do not exist anywhere else but on the website;
- whose ownership or responsibility is unclear, or lacking altogether;
- that constitute records, according to the definitions in Chapter 2;
- that have potential archival value, according to definitions supplied by the archivist.

Appraisal and selection

Appraisal identifies those web resources which constitute records, publications and artefacts (as defined in Chapter 2) for preservation. The web resources selected should provide information about, and evidence of, what the Institution has done and why, what it and its staff and students have achieved, and its impact locally and in the wider world.

In simple terms, appraisal of HFI resources for preservation should focus on:

- substantive functions (i.e. teaching, research, academic award administration) and;
- substantive elements (e.g. strategy development, policy development) of facilitative functions (e.g. governance, estate management, public relations).

(From JISC Infonet (2007) *Guidance on Archival Appraisal*)

The questions that need to be answered for appraisal relate to: location; use; and ownership. However there are some resources which can immediately be excluded from the selection list.

Location

As indicated in Chapter 2, web resources can be found on web server(s), in managed systems or in less well-managed systems which may be hosted externally. The locations of these servers and systems must be discovered so consulting the web manager and the IT manager is the first step. This should result in finding out the:

- number and names of domains and sub-domains being used, including staff and student intranets and portals, funded project websites and any others. It may prove difficult to track them all down as departments and other areas of the Institution may register their own domains;
- number and location of web servers. Again this may be difficult to do because of the possible autonomy of departments but the IT Security staff should be able to help as they will manage the firewall;
- managed systems and their locations;
- less managed systems (where known). Discussions with stakeholders are most likely to track these down;
- backup schedules for all relevant systems and servers;
- resources with external dependencies.

Use

Once the locations of the web resources are known, each should be appraised to find out its purpose and how it is being created. It is particularly important to find out if original resources are being created (i.e. those that are not available in any other format).

Ownership

Finding out the owners of the web resources will identify the stakeholders of the web preservation programme. Each owner will be able to provide valuable help in finalising the programme as they can say which resources they would like to be kept and why, and for how long they need to be kept. This approach will also help get people on board and start to embed a culture of good practice and web management within the Institution.

A Guide to Web Preservation 2010

However it is important to be objective when considering the information provided by the stakeholders as it is likely that they may want to keep everything indefinitely. Conversely, it is all too easy to sweep resources away simply because the stakeholder is not around to defend their interests.

What resources can be excluded?

There are some resources which can be immediately excluded from the preservation programme including those that are already being managed elsewhere, and those which have little or no value.

Web-based resources that are already being managed elsewhere

Asset Collections. For some asset collections, or e-resource collections, the web is often just an access tool for the underlying information resource so preservation actions are best concentrated directly on that resource. This class might include: digitised images; research databases; electronic journals; ebooks; digitised periodicals; examples of past examination papers; and theses.

Institutional repositories (examples include DSpace, eprints or Fedora). Institutional repositories are web-based tools, but the materials stored in an IR are already being managed as there are elements such as metadata profiling, secure and managed storage, backup procedures, audit trails of use, and recognised ownership. A well-managed IR therefore already constitutes a recognised digital preservation method in itself.

Duplicate copies. In some cases, the website is a pointer to resources that are stored and managed somewhere else. Or the resource has been uploaded from a drive which is owned and maintained by another department. If it is ascertained that the 'somewhere else' is already being preserved, then it may not be necessary to keep the website copies.

Web-based resources that have little or no value

Institutional web-based applications which deliver a common service. In this case the web application is an incidental component used in the management of such services. Quite often the important record component in such instances is actually stored or managed elsewhere, for example in a database of underlying data.

Services which do not generate any informational material of lasting value to the Institution. Some examples of this are room booking systems, systems which allow automated submission of student work for assessment, or circulation of examination results.

Resources which clearly fall outside the scope of an agreed records retention policy, or an archival selection policy. Examples of this might include Twitter and Instant Messaging, unless evidence can be found of a strong Institutional driver to retain and manage such outputs. (See case study on page 13.)

Prioritisation

Prioritisation is fundamental to successful preservation - keeping everything is rarely possible. So, when considering what to preserve and what not to preserve, the MoSCoW method can be used as this classifies the requirements as one of the following:

- M: resources which must be preserved.
- S: resources which should be preserved, if at all possible.
- C: resources which could be preserved, if it does not affect anything else.

A Guide to Web Preservation 2010

- W: resources which won't be preserved.

Even after carrying out the MoSCoW method, it is possible that the list of web resources in the M and S categories will be long and potentially unmanageable. So the MoSCoW method can then be carried out again with just those in the M category to find out the order in which the resources should be tackled.

Selection approaches

Three main approaches (one unselective and two selective) have arisen from the work carried out by National Libraries and these can be adapted to the requirements of an HFE Institution.

1. Unselective approach - bulk/domain harvesting

This could involve harvesting the entire website, and/or all its associated domains. Some argue that it is cheaper and quicker to be unselective than to go through the time-consuming selection route; that it is demonstrably less 'subjective' and will produce a more accurate picture of the web resource collections; and that since it is technically feasible, why not?

However, aspects of those arguments are more applicable to a digital archive or repository trying to scope its collection within certain affordable and pragmatic boundaries. Secondly, there is no point in capturing 'everything' if it has already been established that there are significant quantities of web resources in the Institution that do not even need capture, let alone preservation. In running a frequent domain-wide harvest of Institutional networks, there is a risk of creating large amounts of unsorted and potentially useless data, and committing additional resources to its storage.

2. Selective approach - criteria-based selection

This could entail selecting web resources according to a pre-defined set of criteria, for example:

- All resources owned by one Department.
- One genre of web resource (e.g. all blogs).
- Resources that share a common subject, or related subjects.
- All resources that affect students or staff only.
- All funded projects with web-based deliverables.
- All resources thought to be at risk of loss.
- All records or all publications.
- Resources that would most benefit an external user community (e.g. alumni).
- Resources covering a pre-determined theme.

3. Selective approach - event-based

Consider if there would be value in taking 'before and after' snapshots of certain web pages, if agents of change are known to be at work. The sort of time-based events which might trigger a decision to capture are:

- End or beginning of term, or beginning of a new academic year.
- Appointment, or departure, of a senior official.

A Guide to Web Preservation 2010

- Completion of a major piece of research.
- Publication of the new prospectus.
- Purchase of new authoring software which affects web content.
- Corporate or Institutional re-branding.
- Formation of a new department.

Having decided on the approach, the next step is to decide which aspects and elements of web resources must be captured.

Aspects

It is possible to make a distinction between preserving an *experience* (experience of accessing web content including all its attendant behaviours and aspects) and preserving the *information* (all meaningful content including words, figures, images, audio) which the experience makes available. Both are valid preservation approaches and both achieve different ends.

Deciding which aspects of web resources to capture can be informed to a large extent by the Institutional drivers, and the agreed policies for retention and preservation.

A few examples are given below.

Evidential and record-keeping: As well as the content, this would involve preserving some form of change history, with as much contextual information as possible. This may not apply to all the web resources, just to those which are needed for legal purposes, to protect the Institution, where decision-making is involved, etc. For such resources the following should be captured and preserved:

- an audit trail of changes and a change history;
- contextual information about people (authors, users etc), and dates and times (creation, change and publication dates etc);
- the content, appearance and behaviour of the resource.

Repurposing and reuse: For web resources which are being reused and potentially repurposed in a different context (or even on a different server), it would make sense to preserve:

- the content, appearance and behaviour of the resource;
- contextual metadata about its creation, its original location, its authorship, its access rights, etc.

Social history: For web resources which are not needed for evidential purposes, but are being preserved to retain something about the history of the Institution, the capture requirements may not be as exacting. For example, if it was decided to preserve a sample of student home pages, appearance of the resource could be preserved to demonstrate how home pages looked five years ago.

Elements

The elements of web resources that need to be considered are content, appearance and behaviour.

A Guide to Web Preservation 2010

Content

This is just the words. No links, no behaviour, no framesets, no stylesheets, no images - just plain text.

Appearance

This is the look and feel of the page including navigational devices, images, page layout etc.

Behaviour

This covers web resources which have dynamic or animated features. One example of a website with a mixture of behaviours would be a blog, which might have behaviours such as a live feed, comments which can change, site administration features, and bookmarking and tagging features.

If the preservation of content, appearance and behaviour is required, the job of preservation becomes more complex. It may not be feasible or desirable to capture all of these elements therefore it is important to specify which are most significant for preservation.

DPC decision tree

A potentially useful tool is the Decision Tree produced by the Digital Preservation Coalition. It is intended to help build a selection policy for digital resources, although it should be pointed out that it was intended for use in a digital archive or repository. The Decision Tree may have some value for appraising web resources if it is suitably adapted.

Action

- Define a policy for the appraisal and selection of resources, consulting stakeholders as appropriate.
- Refine the list of resources for preservation drafted after Chapter 2.

Case study – Vanishing domain names

Scenario: *A project team in the Institution has purchased a top level domain with a .org suffix, outside the main Institution domain, in order to expose and store its project outputs. The project is now developing into a successful service, there are numerous dependencies, and users have come to trust the domain. But the project manager failed to renew the domain name subscription, and it has now been purchased by a third party. This third party is requesting a significant fee to release the domain name back to the Institution.*

Issues

If your resources are located on the main Institutional website (usually in the .ac.uk second level domain, managed by JANET), then your domain is unlikely to disappear unless there are major changes affecting your institution.

If, however, you are using an alternative domain name (such as .org, .org.uk, .co.uk or .com) then care is needed in managing domain registrations. Internal administrative management procedures will need to be in place to ensure that the domain name is renewed prior to the expiry.

You may ask why anyone would wish to make use of a non-.ac.uk domain in light of such possible dangers. JANET does not sell off its domains to the highest bidder. Instead it has strict eligibility guidelines that may not be met by short-term, collaborative or cross-sectoral projects and services. Equally within Institutions, the allocation of fourth level sub-domains (e.g. specialproject.london.ac.uk) is often tightly controlled or subject to considerable bureaucracy.

Approaches

- Carry out an audit of the Institution's use of non- .ac.uk domains.
- Ensure that such domains have adequate administrative processes in place to ensure that the domain name is not lost if, for example, project funding ceases and staff involved in the project leave the Institution.
- Carry out a risk assessment of the dangers of losing such domains, and the costs your Institution may be willing to pay to claim back the domain.

Chapter 6 – How do I capture them?

Summary:

- The capture of web resources can be carried out within the authoring system or server, at the browser or with a crawler. Each process has its advantages and disadvantages.
- Tools are available for each type of capture.

Tools

A number of tools are available for capturing web resources:

- Tools for capturing web resources.
- Workflow systems and curatorial tools.
- Snapshot tools.

For more see Appendix C. Netpreserve.org and Harvard University Library also have lists of tools.

Point of capture

There are three points in the journey of a web page, from server to user, where its capture is likely to be most feasible.

1 Capture within the authoring system or server

This involves retrieving web pages directly at their point of origin, usually the content management system, or the server on which web pages are held. It is possible to get all the content (HTML, CSS, GIFs, JPEGs, etc) however, increasingly web pages are formatted 'on-the-fly' to suit the specific needs of the user that is requesting them (e.g. type of browser, small screen or large screen, desktop device or mobile phone). So this method raises the question of which of these possible versions should be captured.

2 Capture at the browser

This could also be described as capture post-rendering, or at the point of the HTTP transaction. It implies something of a snapshotting approach, and such a snapshot is going to result in frozen content.

3 Capture with a crawler

Using a crawler is going to resolve some of the problems of other methods but not all. Crawlers are unlikely to succeed totally as they miss other external sources such as document servers, databases and datafeeds, internal databases, subscription databases, and file management platforms. Web content management systems, access methods, protocols, and security and logins, may also present barriers.

Many crawlers, including Heritrix, are also prone to the 'collateral harvesting' problem. This means they can gather lots of content which is not needed, by blindly following links. There are ways of setting exclusion filters to prevent this, but its behaviour can still be unexpected.

A Guide to Web Preservation 2010

Types of capture	Advantages	Disadvantages
Capture within the authoring system or server	<p>Easy to perform if the server is owned by the Institution.</p> <p>Works in the short to medium term, for internal purposes.</p>	<p>Captures raw information, not presentation.</p> <p>May be too dependent on authoring infrastructure or CMS.</p> <p>Not good for external access.</p>
Capture at the browser	<p>Relatively simple for well-contained sites.</p> <p>Commercial tools for doing it exist.</p>	<p>What the user sees is captured (but it is not necessarily known why).</p> <p>Treats web content like a publication: frozen.</p> <p>Loses behaviour and other attributes.</p>
Capture with a crawler	<p>Most widely-used method.</p> <p>Defers some access issues.</p> <p>Provides link re-writing.</p> <p>Provides embedded external content: from archive or live.</p>	<p>Lots of work, tools and experience are necessary.</p> <p>Presents many problems for capture: often not everything is captured, or too much is captured.</p>

Action

- Decide on your capture approach.
- Assess the tools and decide which is/are the most appropriate for your Institution.
- Define policies for the capture of, and access to, preserved web resources.

Case study – Student blogs

Description: *Your Institution runs a blog service for students and staff. One enthusiastic alumna wants to migrate the extensive blog she has kept for three years, but your Institution systematically deletes files and accounts held by students on Institution servers shortly after they graduate. How should the Institution respond if students wish to maintain or migrate the contents of their blog (plus embedded resources, comments, etc.)?*

Issues

- Should the option be open to students to have their resources persist on Institutional servers after they leave - perhaps as part of an Alumni programme?
- Should this be an opt-in or opt-out process, and should fees be involved?
- Does an Institution have permission to archive the content of blogs? This might include permission not only from the blog's author but also from creators of third party content.
- Is it possible to excise potentially offending material, or is the risk (probably negligible) that an Institution might be sued for copyright breaches acceptable?
- Are Institutional staff and students well-informed about the issues of online copyright? Is it possible to include a default Creative Commons licence in the terms of use of the system?
- Is it more sustainable for the Institution to host and manage a blogging service or to use third party providers such as Blogger.com or Wordpress.com?

Approaches

- Decide that the issue is predominantly one of policy, not of in-house hosting versus third party hosting. If an educational Institution is encouraging the use of blogs to support reflection, discourse and deep learning, it has a responsibility to make that online environment as safe as it tries to make its physical campus.
- Institutions could recommend the use of mature hosted blogging services for students who will normally only be at the Institution for a short period. Third party hosting might be a reasonable alternative to the costs of service development and maintenance, but the Institution must examine the terms and conditions and functionality very carefully to ensure they meet standards it can recommend to those in its charge.
- Seek permission from the owners of the blog content before making copies, investigate wider application of Creative Commons licences and work towards resolving third party issues.

Chapter 7 – Who should be involved?

Summary:

- The Institution must take ownership of the web preservation programme at the highest level so that staff and students are motivated to take part.
- There are some national and international initiatives which might help.

Success with the preservation of web resources will potentially involve the participation and collaboration of a wide range of experts: information managers, asset managers, web managers, IT specialists, system administrators, records managers, and archivists. Each of the participants will have an interest derived from their role which means some may be more driven in terms of the process than others.

For example, the interest of records managers will be in legal compliance, long-term RM goals, retention, disposal, and classification of those web resources classed as records. This is a central activity in their role, whereas web managers may be less driven as their role is more about managing current content and new technological initiatives. Each needs to consider the other's view of preservation and work together to make the web preservation programme a success.

Records managers need to:

- recognise that the web is a potential place where records can occur, and identify where, how, when and by what agency this is happening;
- rethink some of their traditional models as centralised control over web resources is not possible;
- overcome any fear of IT, and forge relations with people like the web manager, system administrator, and IT manager.

Web managers:

- need to recognise that the web is a potential place where records can occur, and that they have some responsibility to ensure they are protected;
- should think twice before deleting everything or disabling an account;
- should exploit software for better ways to capture and manage the content;
- consider preservation-friendly software for the next purchase of web tools.

As well as those with the specific roles identified above the preservation programme must also involve the stakeholders identified during the appraisal process (see Chapter 5).

The Institution as owner

It is vital that the Institution takes ownership at the highest level as this will motivate all Institutional staff and students to take part in the web preservation programme. Also, effective web preservation needs to be policy-driven as it is about changing behaviour, and consistently working to policies. A clear policy is needed that states the importance and value of the Institution's web resources, and makes it clear why some of them are being preserved. There should be a sense of corporate ownership of the Institution's

website, the web publication programme, web resources that have value and need to be preserved, and all other issues associated with making the resources ready for preservation such as capture, storage and management.

There are many drivers for an Institution to take ownership of the web preservation programme and these are discussed in Chapter 3. However, the strategic, legal, financial and contractual obligations coupled with the risk issues should provide enough impetus for the Institution to take action.

Technology is not the answer

Web resources have existed in UK HFE Institutions for many years, and so have the tools that would help an Institution capture, manage and store those resources. The fact that these tools are not being widely used is an indicator that web preservation is not primarily a technological problem: the solution does not lie in buying new software or more software. There is no single tool that addresses all possible web preservation issues (behaviour, dynamic content, scripts, versioning, etc.), so any programme of work associated with web preservation needs to be properly resourced, with team-based and collaborative approaches drawn from across more than one discipline.

Can other people do it for you?

Internet Archive

The Internet Archive (<http://www.archive.org>), also known as the 'Wayback Machine', is unique in that it has been gathering pages from websites for so long that it holds web material that cannot be retrieved or found anywhere else, and would have been lost.

The Internet Archive has ways which allow anyone to submit a website for inclusion in the Archive including a subscription service called Archive-It (<http://www.archive-it.org/>). The advantage of this service is that distinct web archives (collections) are created containing only the content selected for harvesting and at the chosen frequency. These can be catalogued and managed directly by the subscriber. The assumption is that archived copies will be made public, via the Internet Archive, although arrangements can be made to keep them private.

Additionally, people are encouraged to use the Internet Archive as a sort of 'People's Repository'. By registering, it is possible to upload images, texts, moving images, and audio material, thus making use of IA's considerable storage capacity. Again, in return for free storage, it is expected that your resource is made public.

A few caveats about the suitability of this solution to UK HFE Institutions:

- IA lacks an explicit preservation principle or policy, and has no real mandate to capture websites outside of a societal desire to see it happening and to share the results with the public. This may cause severe problems to HFE Institutions; as it may not cover everything the Institution needs to do within its remit.
- There is potential for legal difficulties and litigation. IPR issues may not be adequately dealt with by Creative Commons and the IA's 'notice and take down' approach.
- IA may not have a sustainable funding model and its continuance is largely dependent on the generosity of its creator, Brewster Kahle.

There are additional caveats about the technical failings of Internet Archive:

A Guide to Web Preservation 2010

- IA will not capture all web-based assets and cannot capture any site or service that depends on a database, or a login.
- IA cannot guarantee capture to a reliable depth, or reliable quality. Note that this applies to dynamic content.
- There can be large gaps between capture dates.
- The image assets in IA are always smaller than archive quality copies.
- IA may not be preserving the resources they capture to OAIS standards.
- There is little in the way of contextual information in their catalogues.

A number of Institutional assets are missed out by the IA approach including library catalogues, image collections, e-print collections with a database, and interactive teaching materials.

UKWA

The UK Web Archive (UKWA) is a British Library initiative which started gathering and curating websites in 2004. The archive is free to view, accessed directly from the web itself and has collected thousands of websites. UKWA's approach is selective, and determined by written selection policies. The Archive contains UK websites that publish research, reflect the diversity of lives, interests and activities throughout the UK, and demonstrate web innovation.

It may be possible to nominate a Institutional website for capture with UKWA but it may not be selected or archived. The current position on this process is that owners of UK website are especially encouraged to nominate their sites but that UKWA reserves the right to decide whether to include the nominated sites.

Institutions may also want to bear the following in mind:

- The capture will be a snapshot of the website at a certain date and time.
- Certain resources will be beyond the reach of the Heritrix crawler (e.g. databases, secure and password-protected pages, and hidden links).
- Similarly, if the website depends heavily on server-side architecture, then remote capture may fail.

If the website is selected by UKWA, it will involve a few practical things:

- Signing a permissions agreement to allow remote harvesting and copying.
- Agreeing to have the archived copy made publicly available.
- Allowing the remote harvester to ignore robot exclusions.

A UKWA solution is better than nothing but there are limitations, and it may not constitute a quality solution to preservation of all the Institution's web resources. (<http://www.webarchive.org.uk/ukwa/>)

The International Internet Preservation Consortium (IIPC)

The IIPC will not help to harvest an Institution's website, but they are an internationally-recognised body of excellence for website preservation. The mission of the IIPC is to acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations.

A Guide to Web Preservation 2010

The goals of the consortium are (<http://www.netpreserve.org/about/index.php>):

- To enable the collection, preservation and long-term access of a rich body of Internet content from around the world.
- To foster the development and use of common tools, techniques and standards for the creation of international archives.
- To be a strong international advocate for initiatives and legislation that encourage the collection, preservation and access to Internet content.
- To encourage and support libraries, archives, museums and cultural heritage institutions everywhere to address Internet content collecting and preservation.

Other external initiatives of interest

These initiatives, mostly library-based and sponsored at a National level, aim to complete selective web collections, often based on the aim of archiving the entire 'national' domain. They are provided here as they may be able to offer some useful lessons learned. However, they will not be able to assist with the archiving of an Institution's website. Some of the National Library collections are not open to the public.

MINERVA, Library of Congress (<http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html>). A selective web archive based on themes of national importance, i.e. national political elections, wars and terrorism.

PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) (<http://pandora.nla.gov.au/>). National Library of Australia with nine other Australian libraries and cultural collecting organisations.

UK Government Web Archive (<http://www.nationalarchives.gov.uk/webarchive/>). A selective collection of UK Government websites, archived at regular intervals from August 2003, developed by The National Archives using Internet Archive.

Austrian On-line Archive (AOLA), Austrian National Library and Technical University of Vienna (<http://www.ifs.tuwien.ac.at/~aola/>).

DAMP (Digital Archive for Web Publications) (<http://www.nsk.hr/DigitalLib.aspx?id=80>), University of Zagreb and the National and University Library (NUL) in Zagreb, Croatia.

Kulturarw3 - KB Web Archive, Royal Library - the National Library of Sweden.

The State and University Library and the Royal Library, Denmark (<http://netarchive.dk/>).

WebArchiv, National Library of the Czech Republic and Masaryk University in Brno (<http://en.webarchiv.cz/>).

Action

- Consider who needs to be involved to ensure the web preservation programme is a success not just in the short-term but also in the long-term.
- Discuss the programme with all the relevant people, perhaps setting up a working group to facilitate the discussions.

Case study – Capturing wiki contents

Scenario: *Wikis are now at the online heart of innumerable projects - for teaching, research, publishing and business. When the project is complete what should be done with the content to ensure it is retained? Does wiki software allow for this?*

Many wikis have a backup option which will enable the capture of wiki content, such as Wetpaint, Wikidot, Mediawiki, and Confluence. However all these options produce imperfect results. In contrast to this, if a spidering engine like HTTrack or Wget (see Appendix C) is used to harvest the site remotely, a working local copy of the wiki, looking much as it does on the web, will be the result. This might be an attractive option if a record of what it looked like on a certain date is required.

However, it may not be necessary to gather every web page since the wiki contains many automatically generated pages: versioning, indexing, admin etc. So a selection decision is needed. For example, the edit history and discussion pages may be excluded as the user community only wants to look at the finished content.

The change history is important to the current owner-operators of the wiki, however is this really needed for long term (or even permanent) preservation. Indeed, could their access requirement be satisfied merely by allowing the wiki (presuming it is reasonably secure, backed-up etc.) to go on operating the way it is, as a self-documenting collaborative editing tool?

Approaches

All this suggests some basic questions to ask when setting up a wiki for a project:

- What aspects of the wiki do we want to preserve and for how long?
- Is there a business need to capture the wiki change history, and for how long?
- Will it need preserving at intervals, or at a completion date?
- Is it more important to preserve text content, complete functionality, or its look?
- Should we back it up? If so, what should we back up?
- Does the wiki provide backup features? If so, what does it back up (e.g. attachments, discussions, revisions)?
- Once 'backed up', how easily can it be restored?
- Will the links still work in our preservation or backup copy?
- If the backup includes raw wiki markup, do you have the capabilities to re-render this as HTML?

Chapter 8 – What approaches should I take?

Summary:

- Four main operational models can be considered.
- The quick win approaches include domain harvesting, carrying out pilot projects and considering using the EDRMS.
- Strategic approaches include information lifecycle management, adapting records management approaches, and the continuity approach.

Operational models

There are various operational models for developing and implementing a web preservation programme. Selecting a model will depend on balancing such factors as costs, risks, priorities, and available resources and infrastructure.

In-house: the programme is resourced, managed and implemented within the Institution. This approach offers the most flexibility and control, assuming the necessary skills and resources are available. In the USA, the Harvard Web Archive Collection Service (WAX) is an example of a successful in-house project that relied heavily on the active participation of the creators of the resources. 48 Harvard websites were put into the collection, representing Departments, Committees, Schools, Libraries, Museums, and educational programmes.

Contracted-out: all or some of the work is performed by a contractor. It is unlikely a single contractor will have all the skills and resources to perform the entire package, and Institutions may need to look at contracting out services for e.g. hosting and storage or website hosting, while retaining the curational and selection elements in-house.

Collaborative: one or more Institutions work together, pooling skills and resources towards a common goal. JISC often fund such projects which bring together initiatives from two or more HFE Institutions, particularly for collaborative tools and e-learning projects. However, as yet, no funded programme for website preservation exists.

Consortium: the programme is implemented by a consortium of organisations, using some shared resource or infrastructure. The IIPC is an example of the consortium approach. In this country, the UK Web Archive used to fill this role, although its role as a consortium appears not to be as active. The activities of the Digital Preservation Coalition's Web Archiving and Preservation Task Force will also prove valuable in this collaborative area. It aims to 'provide a mutually supportive environment for continued policy development for members and a mechanism through which non-members can engage with web archiving policy'.

Approaches

Once the operational model is decided, there are two main classes of approach:

1) **Quick win:** This can be used to protect resources identified as being most at risk and the approaches include domain harvesting, remote harvesting, pilot projects, and using an EDRMS. They may be attractive because they are quick, and some of them can be performed without involving other people or requiring changes in working. However, they may become expensive to sustain if they do not evolve into strategy.

2) **Strategic:** This includes longer-term solutions which take more time to implement, involve some degree of change, and affect more people in the Institution. These approaches are adapted from Lifecycle Management and records management, and also may involve working with external organisations that will do the work (or some of it) for the Institution. The pay-off may be delayed in some cases, but the more these solutions become embedded in the workflow, the more web archiving and preservation becomes a matter of course.

Quick win

Domain harvesting

This refers to two possible approaches:

- The Institution conducts its own domain harvest, sweeping the entire domain (or domains), using appropriate web crawling tools.
- The Institution works in partnership with an external agency to do domain harvesting on its behalf.

Domain harvesting is only ever a partial solution to the preservation of web content as there are limitations to the systems which currently exist. Too much content may be gathered, including that which does not need to be preserved. Conversely, content which ought to be collected may be missed including hidden links, secure and encrypted pages, external domains, database-driven content, and databases. Simply harvesting the material and storing a copy of it may also not address all the issues associated with preservation.

Pilot projects

Instead of trying to solve the web resource problem for an entire Institution, a pilot project could be carried out from which a visible result will be seen quite quickly. This may make it more persuasive for other departments to participate and will add credibility to the programme. Pilot projects can also generate useful reports about lessons learned that can be used to make the next project even more successful. In addition fewer stakeholders may be involved, if the project is scoped tightly enough, thus saving time on consulting users and owners of the content. Pilot projects could be targeted as outlined in Chapter 5.

Migration

Migration of resources from one operating system to another or from one storage/management system to another is a form of preservation. This may raise questions about emulation and performance. Can the resource be successfully extracted from its old system, and behave in an acceptable way in the new system?

Can web resources be put into an EDRMS?

It is not yet known how feasible it is to use an EDRMS (Electronic Document and Records Management System) for the management of web resources. These systems seem to work best with static documents; authors of reports, for example, understand that a good time to declare their report as a record is when the final approved version has been accepted. Yet one of the distinctive features of Web 2.0 content is that the information is very fluid, and often there is no obvious point at which to draw this line and fix content.

It is technically feasible, for example, to capture Instant Messaging outputs as text or HTML files which could be saved into an EDRMS. The question remains whether there

is a defined policy that supports doing this; one that recognises use of IM as a legitimate record-keeping tool, and as a practice that is acceptable to the Institution. The attraction of storing certain web-based output in an EDRMS is that then such resources could be managed in line with agreed retention schedules; and that related records are filed together, like with like.

Strategic Approaches

These include Information Lifecycle Management, adapting records management approaches, or following the web continuity methodology. Any approach taken should ensure that web resources are protected from careless or wrongful destruction, deletion, or removal of the resource.

Information Lifecycle Management

Information Lifecycle Management (ILM) involves recognised professional standards and practices, leading to better management of information, and is one possible approach. If a lifecycle model can be applied to web resources, they will be created, managed, stored and disposed of in a more efficient and consistent way. It can also assist with the process of identifying what should and should not be retained, and why; and that in turn will help with making preservation decisions. ILM makes no assumptions about software or IT systems, nor does it assume that all information will be managed through a single software tool; rather, it is a conceptual framework to help ensure consistency within an organisation. It can be especially helpful when introducing new systems, or reviewing existing ones.

Information moves through a series of phases over time. JISC's approach to ILM proposes four distinct phases: creation; active use; semi-active use; and final outcome.

Information should be managed throughout each phase, and there are pertinent issues which apply. ILM can also be aligned very closely to the records management programme. An ILM approach always takes a start-to-finish, cradle-to-grave view. A model can be adapted according to your Institutional needs. The model should have a chronological structure, clearly defined phases, user identification, and consistency.

There is a lot of literature available including the JISCInfoNet published guidance, *Managing The Information Lifecycle*, which is geared towards the HFE sector. (See <http://www.jiscinfonet.ac.uk/infokits/information-lifecycle>.)

Adapting records management approaches

If a records model can be applied to web resources, the same benefits associated with ILM apply: web resources will be created, managed, stored and disposed of in a more efficient and consistent way. The RM programme will already be established, and through the agreed retention schedules it can assist with the process of identifying what should and should not be retained, and why. All of that in turn will help with making preservation decisions. Under records management, these things will take place within a legislative and regulatory framework that enables and obliges the creation and disposal of records.

Continuity approach

The 'Web Continuity' project involves a 'comprehensive archiving of the government web estate by The National Archives', and aims to address both 'persistence' and 'preservation' in a way that is seamless and robust. Web continuity offers concepts and ways of working that may be adaptable to a web preservation programme in an

A Guide to Web Preservation 2010

Institution, particularly as a main area of focus is the integrity of website links. The project's use of digital object identifiers (DOIs) can marry a live URL to a persistent identifier. To achieve persistency of links, a redirection component derived from open source software is used. This component will 'deliver the information requested by the user whether it is on the live website, or retrieved from the web archive and presented appropriately'. Of course, this redirection component only works if the domains are still being maintained, but it will do much to ensure that links persist over time.

Part if the project involves building a centralised registry database of Government websites, and is a means of auditing the website crawls that are undertaken. Such a registry approach is worth considering on a smaller scale for an Institution as would the project's methodology for rolling out XML site maps across government. These site maps can help preservation because they help to expose hidden content that is not linked to by navigation, or dynamic pages created by a CMS or database.

The intended presentation method will make it much clearer to users that they are accessing an archived page instead of a live one and helps to address any potential liability issues arising from members of the public acting upon outdated information.

Action

- Decide on the most appropriate approach for your Institution. Options include: doing everything in-house; contracting out all or some of the work'; collaborating with other Institutions; working within a consortium.

Chapter 9 - What policies need to be developed?

Summary:

- All existing policies which may have an impact on web preservation should be located and assessed.
- A policy review should be carried out.

It is unlikely that any Institution will have a single policy or mission statement that governs everything that should be happening in relation to websites and web resources. Any relevant Institutional statements are probably scattered across several places and departments; further, any guidance relating to the creation, storage and preservation of web-based materials may only be implied rather than made explicit.

However, the first step is to investigate the following sources (where available) as they may prove helpful.

- Institutional mission statement.
- Legal or legislative mandate.
- Regulatory requirements.
- Change management policy and procedures.
- Terms and conditions of website use, website privacy statement, accessibility policy, disclaimer, and copyright notice.
- Acceptable use policy / regulations concerning use of Institutional computing.
- Code of conduct for work areas and use of software.
- Web publishing policies and guidelines.
- IT security policy.
- System administration code of practice.
- Blogging terms and conditions.
- Records management policy.
- Archivist's collection and preservation policies.
- Digital library guidelines.
- Institutional Repository deposit agreements.
- E-learning object repository policies.
- Any institutional or departmental policies governing information management, asset management, or knowledge management.

It may also be useful to locate the Minutes of any Committees or Advisory Groups in the Institution which formulate web development strategies or advise on policy and current development activities.

Assessing policies

Once all relevant policies are located, the following should be considered:

- Do any policies refer explicitly to web resources?
- Do the policies refer to the three proposed web resource classes? See Chapter 2.
- Do the policies suggest any action with regard to keeping web resources?
- Is there any scope for influencing the behaviour of those who create and use web resources?
- Is there any scope for assigning responsibilities for creation, capture and management of web resources to individuals?
- Would these policies allow the carrying out of preservation actions?
- Would these policies prevent the carrying out of preservation actions?

For example, a records manager's retention schedules may not explicitly mention web resources by name as the content of the record tends to be identified rather than describing the form it is in. Therefore these, and other policies, will need to be translated, or adapted, to address the preservation of web resources. This includes all web resources and must stand the test of time without the need for endless revision.

Policy review

Reviewing policies and procedures is vital and should be embedded as a continual review action within the preservation process itself. As part of its long-term and evolving strategy, the Institution should:

- Strive to define technology-neutral policies. The policies should not be dependent on a choice of software, nor the format of the resource.
- Apply the policies to emerging systems.
- Make sure that its web resources and their management are explicitly covered by appropriate policies.
- Separate decisions about what policy says would be ideal, from what is achievable using current resources and technology.

Decisions made at these stages should be taken at Institutional level, so that ways can be found of embedding the decisions in practice, or matching them up to existing policies.

Action

- Review all related strategies, policies and guidelines, and identify any gaps.
- Start the process of ensuring web preservation is covered by all appropriate strategies, policies and guidelines. This is important if the Institution is to be the owner of the web preservation programme.

APPENDIX A

Further reading

Chapter 1

van Harmelen, M. Briefing paper on Web 2.0 technologies for content sharing: *Web 2.0 – An introduction* (<http://franklin-consulting.co.uk/LinkedDocuments/Introduction%20to%20Web%202.doc>).

MacGlone, E. (2008). *YouTube and the National Library of Scotland* in *WIDWISAWN*, Vol. 6, No 1. (http://widwisawn.cdlr.strath.ac.uk/issues/vol6/issue6_1_4.html).

Chapter 3

Beagrie, Neil *Digital Preservation Policies and their implementation* (http://www.jisc.ac.uk/fundingopportunities/funding_calls/2008/01/dppolicy.aspx).

JISC/University of Glasgow (2007) *espida: Making it happen by getting real*. (<http://www.gla.ac.uk/espida/>).

Korm, N. and Oppenheim, C. (2007) *IPR Risk Assessments, Rights Clearances And Rights Management: Practical guidelines for content creators within FE and HE*. HEFCE.

Chapter 5

Digital Preservation Coalition (2006). *Decision Tree* (<http://www.dpconline.org/decision-tree.html>).

Hallgrimsson, Thorsteinn, National Library of Iceland in *Proceedings of the Fifth iPRES Conference 29-30 September 2008*, pp 305-306.

Brown, Adrian (2006) *Archiving Websites: A Practical Guide for Information Management Professionals*, Facet Publishing 2006.

JISC Infonet (2007) *Guidance on Archival Appraisal* (<http://www.learninfonet.ac.uk/partnerships/records-retention-he/archival-appraisal>).

Chapter 6

Netpreserve.org. Downloads (<http://www.netpreserve.org/software/downloads.php>).

Harvard University Library. Web Archiving Resources (<http://hul.harvard.edu/ois/systems/wax/resources.html>).

Paynter, Gordon, et al (2008) *A Year of Selective Web Archiving with the Web Curator at the National Library of New Zealand*. In *D-Lib*, May/June 2008 (<http://www.dlib.org/dlib/may08/paynter/05paynter.html>).

Chapter 8

Guy, M. (2008) *When Do We Do Fixity*. On JISC-PoWR blog: (<http://jiscpowr.jiscinvolve.org/2008/07/14/when-do-we-fixity/>).

JISC Infonet (YEAR) *Managing The Information Lifecycle*. (<http://www.jiscinfonet.ac.uk/infokits/information-lifecycle>).

JISC (2008) *Change management Infokit* (<http://www.jiscinfonet.ac.uk/infokits/change-management>).

APPENDIX B

The JISC-funded PoWR (Preservation of Web Resources) project

This Guide is one of the outputs from the JISC-funded PoWR project carried out jointly by ULCC (University of London Computer Centre) and UKOLN (University of Bath).

One of the goals of PoWR is to make current trends in digital preservation meaningful and relevant to information professionals with the day-to-day responsibility for looking after web resources. Anyone coming for the first time to the field of digital preservation can find it a daunting area, with very distinct terminology and concepts. Some of these are drawn from time-honoured approaches to managing things like government records or Institutional archives, while others have been developed exclusively in the digital domain.

PoWR workshops

The Project ran three workshops: in London, Aberdeen and Manchester. The workshops, organised by UKOLN, were a mixture of presentations and break-out groups, where a great deal of useful discussion took place and many ideas were generated. Much valuable and interesting input was gleaned from the mixture of professionals who participated, including people from a records management background, web managers, and other information professionals with an interest in web preservation, or experience of the difficulties and issues.

The PoWR blog

A blog was built (<http://jiscpowr.jiscinvolve.org/>) at the very start of the project in April 2008. Several key chapters of the Handbook (see below) originated on this blog, many of them starting life as a series of ‘what if’ scenarios or actual case studies, focusing on various challenging aspects of web content and the actual use made of systems in an HFE context. The resulting discussions and comments provided a great deal of content to assess and assimilate.

The Handbook

The Handbook, written by ULCC staff, is a distillation and synthesis of the material gathered via the workshops and blog; it also draws heavily on the expertise of the PoWR team in the areas of website management, records management, digital preservation, etc. The Handbook aims to provide suggestions for best practice and advice specifically for UK higher and further educational institutions, to enable the preservation of websites and web-based resources.

The Guide

This Guide was developed using content from the Handbook and turning it into a practical guide for all those who would benefit from knowing how to go about preserving web resources.

APPENDIX C

Tools for capturing

Tools for capturing web resources

Web harvesting engines are essentially web search engine crawlers with special processing to extract specific fields of content from web pages. The shortcomings of crawlers are described in Chapter 6.

Heritrix

Heritrix is a free, open-source, extensible, archiving quality web crawler. It was developed, and is used, by the Internet Archive and is freely available for download and use in web preservation projects under the terms of the GNU GPL. It is implemented in Java, and can therefore run on any system that supports Java (Windows, Apple, Linux/Unix).

More information: <http://crawler.archive.org>

Download: <http://sourceforge.net/projects/archive-crawler>

HTTrack

HTTrack is a free offline browser utility, available to use and modify under the terms of the GNU GPL. Distributions are available for Windows, Apple, and Linux/Unix. It enables the download of a website from the Internet to a local directory, capturing HTML, images, and other files from the server, and recursively building all directories locally. It can arrange the original site's relative link structure so that the entire site can be viewed locally as if online. It can also update an existing mirrored site, and resume interrupted downloads.

Like many crawlers, HTTrack may in some cases experience problems capturing some parts of websites, particularly when using Flash, Java, Javascript, and complex CGI.

More information: <http://www.httrack.com/>

Download: <http://www.httrack.com/page/2/en/index.html>

A Guide to Web Preservation 2010

Wget	<p>GNU Wget is a free software package for retrieving files using HTTP, HTTPS and FTP. It is a non-interactive command line tool, so it can easily be used with other scripts, or run automatically at scheduled intervals. It is freely available under the GNU GPL and versions are available for Windows, Apple and Linux/Unix.</p> <p>GNU Wget's features include:</p> <p>Converting absolute links in downloaded documents to relative links, so that downloaded documents may link to each other locally.</p> <p>Using filename wild cards, and recursively mirroring directories.</p> <p>Resuming aborted downloads.</p> <p>Multilingual message files.</p> <p>Support for cookies, proxies and persistent HTTP connections.</p> <p>Using local file timestamps to determine whether documents need to be re-downloaded when mirroring.</p> <p>More information: http://www.gnu.org/software/wget/</p> <p>Download: http://www.gnu.org/software/software.html</p>
DeepArc	<p>DeepArc was developed by the Bibliothèque Nationale de France to archive objects from database-driven deep websites (particularly documentary gateways). It uses a database to store object metadata, while storing the objects themselves in a file system. Users are offered a form-based search interface where they may input keywords to query the database. DeepArc has to be installed by the web publisher who maps the structure of the application database to the DeepArc target data model. DeepArc will then retrieve the metadata and objects from the target site.</p> <p>More information: http://bibnum.bnf.fr/downloads/deeparc/</p> <p>Download: http://sourceforge.net/projects/deeparc/</p>

Workflow systems or curatorial tools

These are used for controlling a web harvest, conducting quality assurance checking, initiating and scheduling archiving processes, managing the metadata (including access restrictions), and producing management reports. They may interface with access tools, for repositories engaged with publishing their harvested copies.

Web Curator Tool

Web Curator Tool (WCT) is for managing the selective web harvesting process, is designed for use in libraries and other collecting organisations, and supports collection by non-technical users while still allowing complete control of the web harvesting process. The WCT is now available under the terms of the Apache Public License.

WCT is a tool that interfaces with the Heritrix crawler, allowing a certain amount of configuration of the target's profile, the addition of extra seed URLs, and enabling filters to be applied to gather more (or less) material from the target. It also generates several log files which are more accessible than HTTrack's, and can help determine why gathers are going wrong and how to fix them.

It was developed by the National Library of New Zealand and the British Library, and initiated by the International Internet Preservation Consortium. Since December 2007, Web Curator Tool is being used by the UK Web Archive. (See ARIADNE issue 50, January 2007, www.ariadne.ac.uk/issue50/beresford/).

More information and download: <http://webcurator.sourceforge.net/>

PANDORA Digital Archiving System (PANDAS)

The PANDORA Digital Archiving System, known as PANDAS, was developed by the National Library of Australia as an integrated, web-based, web archiving management system. The need for such a system arose from the scale of the Library's archiving activity and the necessity to enable other PANDORA participants to contribute to the Archive from various geographic locations.

Like WCT, PANDAS is a workflow system; with the crawling being done by HTTrack. It was created to enable very selective harvesting and is not intended for large-scale automated harvests. Its main functions include managing workflow, creating publisher and title entities, access permissions, gather schedules, and metadata. One caveat is that the tool has a very strong bias towards library models (it treats websites and web pages as titles that have authors and subjects).

Also, the software is built from web objects and lacks robustness; its interface with HTTrack is far from clear, particularly when it comes to applying filters to the gather.

A Guide to Web Preservation 2010

	<p>More information: http://pandora.nla.gov.au/pandas.html</p>
NetarchiveSuite	<p>NetarchiveSuite is a curator tool which allows librarians to define and control harvests of web material. The system scales from small selective harvests to harvests of entire national domains. The system is fully distributable on any number of machines and includes a secure storage module handling multiple copies of the harvested material as well as a quality assurance tool automating the quality assurance process.</p> <p>It was developed by the Royal Library and the State and University Library in the virtual organisation netarchive.dk</p> <p>More information and download: http://netarchive.dk/suite</p>

Snapshot tools	
Adobe Acrobat web capture tool	<p>Adobe Acrobat WebCapture generates tagged accessible PDF files from web pages. Acrobat adds the Adobe PDF toolbar and Convert Current Web Page To An Adobe PDF File button to Internet Explorer 5.01 and later, which allows the user to convert the currently displayed web page to a tagged Adobe PDF file.</p> <p>This tool allows web pages, or entire sites, to be captured to a PDF file. Tools like this have their place, but (like all web capture and preservation technologies) they also have their drawbacks. A PDF's print-oriented format isn't a good match to some sites, much as some sites don't look good when printed. Acrobat Web Capture effectively uses the browser's print engine combined with PDF writer pseudo-printer to do its work, so there will be a close correlation.</p> <p>More information:</p> <p>http://www.wap.org/journal/acrobat4capture.html</p> <p>http://www.planetpdf.com/enterprise/article.asp?ContentID=6057</p>

A Guide to Web Preservation 2010

OpenOffice web wizard	Open Office has many advanced features, including the ability to use some of its conversion features in batch mode, therefore it could be used to mass convert web pages into PDF.
A.nnotate	A.nnotate will allow the user to do web page capture. By entering a URL or using a bookmarklet it will take a snapshot of a web page and store a copy of the HTML in a private space on the a.nnotate.com site. This gives a page at a particular point in time. Currently it does a shallow copy (i.e. just the HTML) so if the images are required it would need to download those. The A.nnotate server is also available for local installation (with an API) if it is to be integrated with a CMS. PDFs can also be uploaded to A.nnotate and these get converted to images and rendered in the browser using pure HTML / AJAX (without any dependency on Flash or Adobe reader.
Snagit 9	Snagit is an example of an advanced, commercial screen-capture tool that includes features to capture images and linked files from a web page, and save the source code and URL of web pages. <i>More information:</i> http://www.techsmith.com/screen-capture.asp

APPENDIX D

An introduction to records management

All Institutions have corporate and business records which need to be managed within the framework of a records management programme. This may include web resources.

The website as a record

- The Institutional website, as a site where information is frequently added, updated, and removed, could be viewed as a record of Institutional activity.
- It is also a place where unique records can be stored or generated.
- Transactions may take place which generate record evidence and audit trails.

Things to consider:

- Does the website contain unique digital records?
- Are staff, students, or members of the public making business or career decisions, based on the information they find on the website?
- Are there unique, time-based, evidential records being created via the browser?
- Is the website itself considered an Institutional record that is worthy of capture?

How records management applies to web resources

- Records management may assist with web resources which require managed retention and disposal, particularly for legal, audit, or business reasons.
- It should not be seen as the only way to manage web resources.

Including a website in a records management programme

- The web manager and records manager could cooperate and start to include the site, or parts of it, in the Institutional records management programme.
- This may require interpretation of Institutional policies and procedures, and published records retention schedules. See Chapter 9.
- Bringing a website in line with an established retention and disposal programme means the website will work to defined business rules.
- Retention schedules will enable the destruction of expired web materials.
- Records management will enable the protection and maintenance of web records that need to be kept for business reasons.

Information compliance

- The website should be managed within a legal and regulatory framework, in line with FOI, DPA, IPR and other information-compliance requirements; and the business requirements of the Institution itself.
- Data Protection and Freedom of Information can be two very strong drivers for records management. Further details are in Appendix F.

APPENDIX E

An introduction to web management

Web management is less easy to define than records management as what is covered by the web manager varies from Institution to Institution, plus the location of the web manager within the Institutional structure and their level of seniority may also make some difference in what is included. So what is listed below are the areas generally covered by web management.

Strategy and leadership

- Development, promotion and management of web strategies.
- Leadership and advice on all aspects of the web and emerging web technologies.
- Liaison with all areas of the Institution to promote the web and ensure buy-in.

Content creation and management

- Ensuring consistency, accuracy and currency of web content.
- Setting standards for content creation and management, usability and accessibility.
- If content management is devolved, training and overseeing the provision of web content providers around the Institution.
- Ensuring the web content meets the needs of stakeholders by carrying out user research and usability testing.
- Maintaining the integrity of the information architecture.
- Ensuring the appropriate use of branding.
- If there is a content management system, providing training and support.
- Ensuring content is accessible and meets legal requirements.
- Collecting web statistics.
- Registering domain names.

Web development

- Maintaining the content management system and other web applications.
- Developing web applications as required by the stakeholders.
- Ensuring web standards compliance.

Web infrastructure

- Liaising with IT staff to ensure that the web infrastructure is appropriate to the needs of all the web activities of the Institution.

APPENDIX F

Legal matters

Legal issues for an Institution to consider when capturing and preserving web resources include: copying; republishing; preservation; ownership; Intellectual Property Rights; third party authorship; and information compliance

These activities, and others associated with capture and preservation, can carry some legal risks – many of the same risks faced by the creator of the resources in the first place.

Copying

Institutions may need to be clear about the ownership of their web content. The two main questions that can arise are:

- Can we make copies of the content?
- Can we republish the content, making our copies available to others through a managed retrieval system, most likely doing it online?

Permissions

It may be desirable or prudent for an Institution to consider whether it needs to seek explicit permission from stakeholders. Websites with multiple contributors (such as forums, wikis or blogs) are complex, because the individual contributors may not all agree to having their input copied.

Creative Commons

Creative Commons may be one way of licensing copyrighted material in a way that will allow the Institution to perform copying for website purposes. Creative Commons licences allow distribution and copying of work without seeking further permission from the rights-holder. This is done through easy-to-use licences which are available from the CC website. The licence elements include attribution, commercial use, derivations and sharing.

DRM

Digital Rights Management may restrict what can and can't be done with a digital work, including a web page or an element of a website. It may cover material locked to a particular system or software, or possession of licence key. For website resources, this could apply to other authored elements such as an underlying database, scripts, and flash animation that appear in the website.

Preservation rights

Can copies be stored in a preservation system and perform such actions as are needed to ensure their continued accessibility? This may include making security copies, surrogates, migrating data between formats, and combining content from different sources.

Freedom of Information

- The FOI Act affects all public bodies, including HFE Institutions.

A Guide to Web Preservation 2010

- It makes a general presumption of rights of access.
- Exemptions can be claimed - some subject to a public interest test.
- Some exemptions expire after 30 years.
- If you are affected, someone in your Institution should be aware; most likely the records manager.

In order for the Institution to comply with FOI legislation, a proactive approach is needed when performing the preservation of certain web resources, and FOI could be used as a driver for the web preservation programme. Such requirements could be taking place in a records management context.

Data Protection

The Data Protection Act 1998 establishes a framework of rights and duties which are designed to safeguard personal data. The Act is underpinned by a set of eight data protection principles which, taken together, define the standards that must be met when processing personal data. It gives rights to people about whom the Institution holds information and gives the Institution responsibilities regarding that information.

DPA may include legal requirements **not** to preserve a web resource, such as the Fifth Data Protection principle: 'Personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes'.

In terms of how this impacts on web preservation activities, the Data Protection Act may prevent an Institution from:

- Holding information collected for other purposes.
- Providing access to information if it identifies living people.
- Linking information about individuals from multiple sources.
- Holding information which may be incorrect .

An Institution may have its own local data protection policy, perhaps supported by written guidance so the records manager or Data Protection Officer should be consulted.

Links

- JISC Legal (<http://www.jisclegal.ac.uk/>) offers a number of legal guides for areas covered here, such as Data Protection, FOI, and IP. They are a free information service specialising in legal advice on the intersection of further and higher education and ICTs, including the web.
- Web2Rights project (<http://www.web2rights.org.uk/>). This project addresses many of the legal issues, particularly those around intellectual property, in the use of these services and has an 'IP Toolkit' service available.
- OSS Watch (<http://www.oss-watch.ac.uk/>) provides advice on the use of open source software (which can be used on some websites) and how to comply with their terms.
- DPC Handbook: see (<http://www.dpconline.org/advice/digital-preservation-preservation-issues.html>) for general advice on legal issue in digital preservation.

Glossary

Appraisal

The aim of archival appraisal is to identify and select records which, collectively, build a comprehensive but compact picture of the Institution over time as a corporate entity, a teaching and learning organisation, a research and innovation organisation, a contributor to economic and cultural development, etc. See Chapter 5.

Archiving (1)

The permanent preservation of those records which have permanent value to the Institution, or records which may be deemed to have historical, heritage, and research value to others. The Institution will aim to preserve those records designated as having permanent legal, administrative or research value at the earliest possible stage in the records lifecycle.

Archiving (2)

Backup of digital resources. It is best to consider the scope of digital preservation as much broader than digital archiving, although the terms are often used interchangeably. In computing generally, 'archiving' is the process of backup and offline storage of data so the term 'digital preservation' helps avoid confusion when referring to the broader issues of managing digital materials and information in and about them.

Asset or Asset collection See **Digital asset**

Blogs

Method of publishing online journals for a wide variety of information dissemination purposes. Often readers can leave comments on individual entries.

Cloud computing

An approach to management of computer systems and data where users access technology-enabled services from the Internet ('the cloud') without needing to have knowledge about or control over the technical infrastructure that supports them.

Collaborative editing tools

Used for collaborating when editing documents and spreadsheets, allowing users in different locations to edit the same document at the same time.

Content Management System (CMS)

A content management system (CMS) is a server-side web application which uses scripts and a database to store, manage and present content on the web. Typically it offers ways to manage common website features such as templates, menu bars and search functions, in ways that allow content creators to focus on the content, without having to bother about the code in HTML headers, JavaScript and CSS files.

Continuity

Among the objectives of web continuity are to ensure that all links work in perpetuity; no cited information is lost through deletion; and information is preserved long-term, even if the web is no longer the dominant publishing medium it is today. Source: The TNA Web Continuity Project 2008.

Digital asset

Any form of salient information that plays a role in the Institution's efficiency and effectiveness. If managed properly, assets can maximise efficiency, productivity and profitability. They could be stored (sometimes permanently) in an archive, a digital library, or an Institutional Repository. Or they could be kept for the short to medium term for business reasons, then disposed of according to a records management schedule. They may be both shared and shareable. They could have reusable content that can support both short-term and long-term use. On the other hand, some of them may contain confidential or sensitive information that means sharing has to be managed and secure. Digital objects can be thought of as assets because they help defend the value of other things (as evidence for patent claims, for instance), because they are needed for regulatory compliance, because they have intellectual value, or because they meet some other organisational need. Source: The AIDA self-assessment toolkit (ULCC 2008).

Digital curation

The term digital curation has recently gained prominence. It places greater emphasis on the activities required to maintain the integrity of digital collections over time, and keep them usable. It promotes a proactive approach to managing digital resources and the use of technological solutions, like web services, to address the problems that technology itself has created. It also paves the way for the emergence of 'digital curators', continually monitoring collections and intervening when necessary - a role analogous to their non-digital counterparts. Source: Digital Curation Centre.

Digital preservation See **Preservation**

Disaster recovery

‘The process, policies and procedures of restoring operations critical to the resumption of business, including regaining access to data (records, hardware, software, etc.), communications (incoming, outgoing, toll-free, fax, etc.), workspace, and other business processes after a natural or human-induced disaster.’ Source: Wikipedia:Disaster recovery.

Document Management System (DMS)

A system used to manage, track and store electronic documents (whether born digital, or digitised from paper originals).

Domain harvesting

A domain harvest involves attempting to harvest all the web material within an Internet domain; for example, all the websites whose URLs end in '.ac.uk'. Source: National Library of New Zealand. See Chapter 5.

Electronic Document and Records Management System (EDRMS)

A safe, secure and governed information and recordkeeping system that applies business classification, disposal, metadata management and security to enable the capture and management of information and records. It facilitates the efficient management and discovery of digital information and records.

Emulation

A means of overcoming technological obsolescence of hardware and software by developing techniques for imitating obsolete systems on future generations of computers.

Fixity (digital preservation)

Fixity, in preservation terms, means that the digital object has not been changed between two points in time or events. Technologies such as checksums, message digests and digital signatures are used to verify a digital object's fixity. Fixity information, the information created by these fixity checks, provides evidence for the integrity and authenticity of the digital objects and are essential to enabling trust. Source: http://www.library.yale.edu/iac/DPC/AN_DPC_FixityChecksFinal11.pdf

Fixity (record declaration)

The initial point at which the content of the record is fixed, a process commonly known as declaration. All records will have a life before they are declared as a record and their contents fixed. They will be drafted, edited and redrafted as draft documents many times before their contents are agreed, finalised and ready for any formal sign-off procedure. It is at this point that the process of declaration should occur and a record be created. Source: <http://www.jiscinfonet.ac.uk/infokits/records-management/creation/fixity-and-declaration>

Information lifecycle See **Lifecycle management**

Ingest

The OAIS entity that contains the services and functions that accept Submission Information Packages from Producers, prepares Archival Information Packages for storage, and ensures that Archival Information Packages and their supporting Descriptive Information become established within the OAIS. This is a very specific term from the OAIS reference model. Source: <http://public.ccsds.org/publications/archive/650x0b1.pdf>

Instant Messaging

Method of creating informal messages for conducting informal business. However, a thread may start as an informal chat and develop into something more formal later, it may be decided in advance that IM is going to be used for formal work, thus obliging a recordkeeping step.

Institutional Repository (IR)

An Institutional Repository is an online locus for collecting, preserving, and disseminating - in digital form - the intellectual output of an Institution, particularly a research Institution. This includes materials such as research journal articles, before (preprints) and after (postprints) undergoing peer review, and digital versions of theses and dissertations, but it might also include other digital assets generated by normal academic life, such as administrative documents, course notes, or learning objects. Source: http://en.wikipedia.org/wiki/Institutional_repository

Internet Archive

Based in San Francisco, the Internet Archive is an open, online archive of digital material. It uses Heritrix, a remote harvesting system for mirroring websites; and the Wayback Machine, an access tool. See Chapter 7.

Lifecycle management

The information that an Institution creates and uses can either represent an asset or a liability, largely depending on how it is managed. The concept of information lifecycle management is about making sure the right questions are asked at the right time

A Guide to Web Preservation 2010

regarding the management requirements of internally produced information. It does this by breaking down the 'lifecycle' that all information moves through into four distinct phases and identifying what are the most pertinent issues that influence how information should be managed during each phase. See Chapter 8.

Media sharing services

Method of sharing multimedia resources such as images, videos, podcasts and presentations. These may be used for teaching and research.

Metadata

Metadata is a popular way of referring to that data that supports the discovery, understanding and management of other data and information. Capturing and maintaining the correct metadata is increasingly being viewed as perhaps the key to the reuse and preservation of digital objects. A large number of metadata schemas and standards have been developed including some initiatives specifically concerned with the development of metadata schemas for long-term preservation. Source:
<http://www.dcc.ac.uk/resources/curation-reference-manual>

Migration

A means of overcoming technological obsolescence by transferring digital resources from one hardware/software generation to the next.

Open Archival Information System (or OAIS)

An Open Archival Information System (or OAIS) is an archive consisting of an organisation of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community. OAIS is the ISO reference model for Open Archival Information System. The OAIS reference model is defined by a recommendation of the Consultative Committee for Space Data Systems. The information being maintained has been deemed to need 'long-term preservation', even if the OAIS itself is not permanent. Source:
http://en.wikipedia.org/wiki/Open_Archival_Information_System

Preservation

Digital preservation is defined as a 'series of managed activities necessary to ensure continued access to digital materials for as long as necessary'. Web preservation is 'the capture, management and preservation of websites and web resources'. Web preservation must be a *start-to-finish* activity, and it should encompass the *entire lifecycle* of the web resource. Source: Digital Preservation Coalition, 2002

Record

Records can be defined as 'recorded information, in any form, created or received and maintained by an organisation or person in the transaction of business or conduct of affairs and kept as evidence of such activity'. Source:
http://www.recordsmanagement.ed.ac.uk/InfoStaff/RMstaff/RM_framework.htm. See Chapter 2.

Records management

Records management is a discipline which utilises an administrative system to direct and control the creation, version control, distribution, filing, retention, storage and disposal of records, in a way that is administratively and legally sound, whilst at the same time serving the operational needs of the Institution and preserving an adequate historical

record. Source:

http://www.recordsmanagement.ed.ac.uk/InfoStaff/RMstaff/RM_framework.htm

Remote harvesting

A method of web capture which takes snapshots of websites by following all the links in each web page.

Repository See **Institutional Repository**

Retention schedule

A process which applies various 'appraisal criteria' such as legal, operational, administrative and historical requirements, to determine how long a particular [record] series needs to be retained. A schedule for retention and disposal of records is often drawn up as a result of applied best practice i.e. based on records surveys, analyses, agreements with business units, etc.

Social bookmarking

Method of recording lists of bookmarks (links to online resources). Bookmarks are usually tagged, and can be viewed, shared and discovered by others using the same application. Useful for teachers creating reading lists, or students and researchers creating bibliographies.

Social networking systems

Allow people to communicate and share information online, either openly or in by invitation-only groups. Virtual, online study, project or research groups can easily set up an environment which combines many other Web 2.0 features. Conversations begun face-to-face, for example at a conference, can continue online, and vice-versa.

Syndication and notification technologies

Method of collating and aggregating news items from diverse sources. Uses XML newsfeeds (RSS and Atom) in diverse ways to alert subscribing users to events, such as new blog posts, podcasts and other new or updated online resources.

Web 2.0

Web 2.0 services and applications are highly interactive and personalised, and have collaboration and social networking as key features. There are seven types: blogs, wikis, social bookmarking, media sharing services, social networking systems, collaborative editing tools, and syndication and notification technologies (all defined in this glossary). In addition, instant messaging, while not new as a standalone application, can be added to the list because of its renewed prominence as a web-based tool. Some Web 2.0 systems may combine features from more than one of these categories, or straddle slightly arbitrary boundaries (Twitter, for example, has aspects of blogging, instant messaging and social networking).

Wikis

Method of collaboratively creating online hypertexts, electronic research or reference resources, mini-websites, and class projects. Some wikis can feasibly be used as a form of EDRMS.

Index

- A.nnotate 45
- Adobe Acrobat web capture tool 44
- Appearance 23
- Appraisal 18-21
- Artefact 8
- Aspects 22
- Asset collection 22

- Backup 4-5
- Behaviour 23
- Blogs 27
- Bulk harvesting 21
- Business continuity 10

- Capture 25-26
- Capture tools 41-45
- Change history 22
- Content Management Systems 4-5
- Continuity approach 35-36
- Crawler 25-26
- Creative Commons 48
- Curatorial tools 43

- Data Protection Act 49
- Decision Tree 23
- DeepArc 42
- Digital Rights Management 48
- Domain harvesting 21, 34
- Domain names 24
- Drivers, internal and external 10-12

- e-learning 6
- Electronic Document and Records Management System 34-35
- Elements 22-23
- E-portfolio 6
- espida 12

- Freedom of Information 48-49

- Heritrix 41
- Home page history 9
- HTTrack 41

- Information Lifecycle Management 35
- Institutional Repositories 6, 20
- International Internet Preservation Consortium (IIPC) 30-31
- Internet Archive 9, 29-30
- IPR 10-11

- JISC-funded PoWR project 1, 40

- Legal matters 48-49

- Metadata 22
- MoSCoW method 20-21
- Multimedia 3

- Netarchive Suite 44

- OpenOffice Web Wizard 45
- Operational models 33

- PANDAS 43
- Pilot project(s) 34
- Policies 37-38
- Policy review 38
- Preservation policy rights 48
- Preservation, definition of 2
- Programme timeline 16
- Publications 8

- Records management 35, 46
- Records managers 28
- Records 7
- Reputation 10
- Retention schedule 35, 46
- Risk management 10-11

- Selection 21-22
- Selection by criteria 21
- Selection, event-based 21-22
- SnagIt 45
- Snapshot tools 44-45
- Stakeholders 19-20

- Task plan 14-15
- Third party sites 4
- Twitter 13

- UK Web Archive (UKWA) 30
- User experience 22

- Virtual Learning Environment 6

- Web 2.0, 4
- Web Curator Tool 43
- Web management 47
- Web managers 28
- Web preservation programme 14-17
- Web resources, definition of 6-7
- Wget 42
- Wikis 32
- Workflow tools 43

