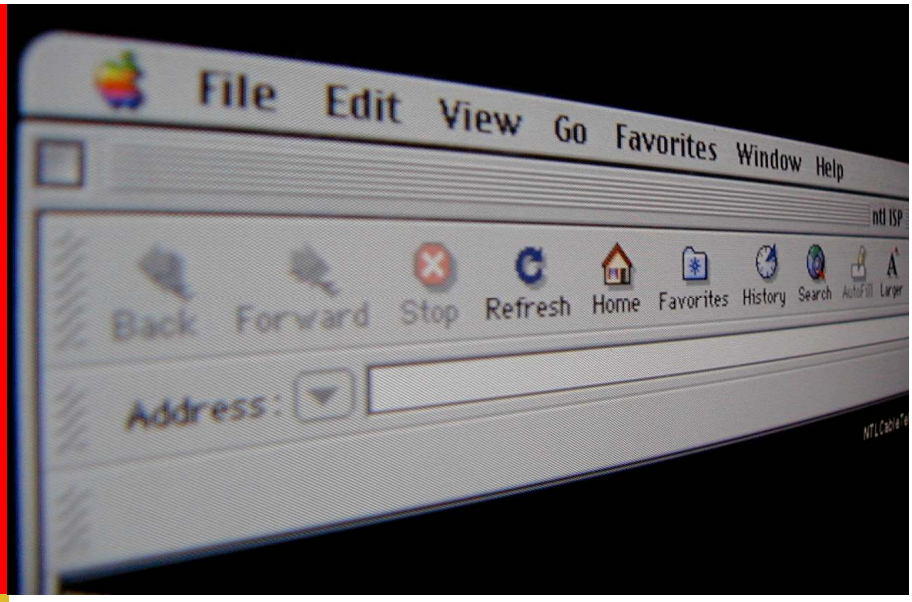

PoWR

The Preservation of Web Resources Handbook

Digital preservation for the UK HE/FE web management community

ulcc
University of London Computer Centre

UKOLN

JISC

Table of contents

The JISC-PoWR team

University of London Computer Centre (ULCC), 20 Guilford St, London WC1N 1DZ
www.ulcc.ac.uk

- Ed Pinsent
- Richard Davis
- Kevin Ashley

UKOLN, University of Bath, Bath BA2 7AY
www.ukoln.ac.uk

- Brian Kelly
- Marieke Guy

OpenContent Lawyer
http://opencontentlawyer.com

- Jordan Hatcher

Acknowledgments:

- Stephen Morrin (proofreading)
- Tim Britton (cover photograph of compact disk)
  http://flickr.com/photos/timbobee/

Introduction

Aims and objectives

This Handbook is one of the outputs from the JISC-funded PoWR (Preservation Of Web Resources) project.

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, run on 27 June (London), 23 July (Aberdeen) and 12 September (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, webmasters, and other information professionals with an interest in web preservation, or experience of the difficulties and issues.

The PoWR blog

We built the blog (http://jiscpowr.jiscinvolve.org/) at the very start of the project in April 2008. Several key chapters of the Handbook 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 gave us 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 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 aimed at UK higher and further educational institutions, to enable the preservation of websites and web-based resources.

We want the Handbook to be accessible and practical, and the content has been structured, as far as possible, as a narrative, starting with familiar ideas and issues, and moving towards more complex issues.

The Handbook is structured in two parts. The first part deals with web resources and makes practical suggestions for their management, capture, selection, appraisal and preservation. It includes observations on web content management systems, and a list of available tools for performing web capture. It concludes with a discussion of Web 2.0 issues, and a range of related case studies. The second part is more focussed on web resources within an Institution. It offers advice about institutional drivers and policies for web archiving, along with suggestions for effecting a change within an organisation; one

such approach is the adoption of Information Lifecycle Management. There are separate Appendices covering Legal guidance (written by Jordan Hatcher) and records management. The Handbook also contains a bibliography and a glossary of terms.

The Handbook is aimed at an audience of information managers, asset managers, webmasters, IT specialists, system administrators, records managers, and archivists.

The Web landscape

Sometime in the mid-1990s, institutions everywhere will have set up a web server in their Internet domain. At first it was probably a few pages of perfunctory contact details and an institutional overview. In some cases, departments and individuals may have been able to create their own sites in sub-directories in the main domain. Since then, everything about the web has grown phenomenally. Expectations of both design and content grew, both for external publicity on the Website, and internal information management on the Intranet. The Web has become the platform and interface of choice for virtually every kind of information system: anything that cannot be found on or through the web is in danger of never being discovered at all.

The kind of web resources that the PoWR Handbook is addressing are still the many diverse, and much more sophisticated descendants of those early web objects. This includes many objects commonly managed in a web CMS, whether available externally or just on the Intranet. Objects may be common native web objects (HTML, CSS, JPG), or other commonly disseminated formats (PDF, DOC, MP3, PPT). They may be database-driven blogs, wikis, or data resources. They may have URLs within the Institution's main domain, or in a subdomain, or within a third-party domain that may be paid-for or not.

The Handbook does not, however, directly address the preservation of:

- Management information systems which use a web-interface (e.g. Agresso finance system, room booking systems)
- Library, record, archival and administration systems that manage a well-defined class of resource, like an Institutional Repository (IR) or Document Management System (DMS)
- Virtual Learning Environments
- Online assessment systems (including e-Portfolios)

The reason for this is that we see these systems as hermetic and essentially self-managed by their professional user base (librarians, finance departments, teachers and learning technologists). A preservation policy, or an Information Asset policy, must encompass all web resources, including these; but the data in these systems will generally be less at risk than less strictly controlled content on the web. In many cases, these classes of system can be considered highly specialised types of Content Management System, increasingly vested with Web 2.0 features, and therefore much of the advice about CMS and Web 2.0 will be relevant.

Web management and records management

The JISC-PoWR workshops have revealed that web managers are likely to see their main responsibility as being to their users – keeping online systems useful, usable and up-to-date. That alone requires a lot of running just to stand still. In addition to changing technology and standards, and ever-greater demands from creators and consumers of information and publications, there is also an ever-changing regulatory and legislative environment, which may require a complete overhaul of the design of the system. Therefore, experience suggests that, perhaps more so than in the library or accounts department, preservation management issues, slip easily off a Web Manager's radar, if

ever they were there in the first place. Yet as a result many valuable institutional resources, and records of them, may be at risk of not even being considered for preservation, let alone preserved.

The PoWR project workshops and blog discussions have highlighted some of the cultural and intellectual differences between the aims of records managers and webmasters. This characterisation can imply that they might even have mutually exclusive aims and priorities. Web managers are portrayed as being interested solely in delivering content and information to a community of users and consumers, and want to keep abreast with technological developments - perhaps at the expense of preservation. Conversely, a records manager might like to capture or manage some web-based outputs, but doesn't know how to do it, is afraid of digital records and rarely communicates with institutional IT staff.

This distinction however tends to oversimplify the case. Records managers don't have all the answers, they aren't necessarily interested in preservation (archivists do that), and even the best records management programme in the world won't address all web preservation issues. Conversely, leave the management of everything solely to the webmaster and you may risk losing valuable resources. The message is that, if we are to achieve optimal longevity and security for all our web resources, records managers must change, as must webmasters.

Permanent preservation

Some reviewers have commented that JISC-PoWR has not dealt in detail with the **permanent preservation** process, as set out in the OAIS standard. Outside of the fact that it would probably require a second Handbook to do so, there were several other reasons for this decision:

1. The specific aims and objectives of the JISC ITT were to *raise awareness* of preservation issues amongst the web manager community, establish reasonable *strategic principles*, and to lay out practical steps necessary to ensure that web material remains *accessible*.

2. Even the international preservation community has yet to engage fully with the issues specific to the permanent preservation of web objects. This would have to include a detailed study of the file formats and file types used on the web, as well as their dynamic behaviours, before attempting to formulate preservation strategies for them. Outside of the ARC/WARC formats, which are really just containers for crawled copies of web pages and their associated metadata, we are not aware of viable, fully developed solutions in this area.

3. The OAIS model is better understood than it was five years ago, but is still in the process of gaining acceptance within the preservation community (such as national libraries and archives). OAIS is not necessarily gaining ground within HFE Institutions within the UK. The Handbook's emphasis has therefore been on selection, management, and capture. In the event that an OAIS system is implemented, Web-based material will form itself into better Submission Information Packages and Archival Information Packages (SIPs and AIPs, in OAIS terms) when selected and managed in this way, and hence be much more easily prepared for long-term digital preservation.

4. Websites and web resources, and the tools used to create them are changing rapidly: Web 2.0 presents both new challenges, as well as some old challenges in new clothes; the Digital Preservation community continues to revisit its understanding of what preservation means when thinking about Web resources, and highlight the importance of different aspects, like *continued access*, *continuity* and *persistence*.

None of this is intended to imply that preservation is out of scope, but the Handbook takes the view that resources must be captured, managed, and selected before they can be preserved, and the Handbook is designed to assist with establishing these first steps.

Part I: PRESERVATION OF YOUR WEB RESOURCES

Chapter 1: What do we mean by preservation?

Institutional views of preservation requirements, and what is meant by preservation, can vary. It is important for those involved to ensure that, broadly, they share the same views and agree on what resources will be included for capture, management, storage or preservation. The Handbook will demonstrate how best to select and appraise your web resources, help to determine which approaches are most suitable for each resource, and which collection methods to use. Summary outlines of certain definitions (which shade into suggested approaches and solutions are listed below. These will be explored in fuller detail in subsequent chapters of the handbook.

Managed resources: We must *manage* resources in order to preserve them. An unmanaged resource is difficult, if not impossible, to preserve. Information lifecycle management, if adapted, can help manage web resources. A records management approach may help to enact preservation for business records or legal reasons, even if you don't intend to keep the resource beyond its expiration.

Protection: Protecting a resource from loss or damage, over the short-term, is an acceptable form of preservation, even if you don't intend to keep it for longer than, say, five years.

Permanent preservation: This means preservation as defined by the OAIS model, which is published and internationally accepted as a feasible model for digital preservation. For web resources, we would assume, in this case, that an institutional decision has been made to keep the resource permanently.

What web resources need to be preserved?

This question will be considered throughout the Handbook, but as a starting point we propose that particular attention be paid to *publications* and *records*. Any other web content worth preserving might be considered an *artefact* with some intrinsic interest. Deciding this will help to determine what kind of approaches to adopt, when considering web resources for preservation purposes. Full definitions of these characterisations can be found in Chapter 4: What are your web archiving requirements?

We also want to be clear about what aspects of the resources should be preserved: web content, appearance, function and behaviour, or access and location; or a combination of all of these? The Handbook makes a particular distinction between preserving an *experience* and preserving the *information* which the experience makes available. Both are valid preservation approaches and both achieve different ends. See also Chapter 6: Web Capture: what and how?

Priorities

Prioritisation is fundamental to successful preservation - keeping everything is rarely possible. Without policies, practitioners have little to guide their decisions about what must, should, could and won't be preserved, let alone how. In considering what to preserve and what not, you could adapt the **MoSCoW** method, which classifies requirements as one of:

- M: Things your institution **must** preserve.
- S: Things you **should** preserve, if at all possible.
- C: Things you **could** preserve, if it does not affect anything else.
- W: Things you **won't** preserve.

Preservation *is* possible!

When faced with the task of preservation projects, institutions can find it so daunting that, in the end, nothing might get done. We hope to demonstrate that the enormity of website preservation and web resources preservation is not as daunting as it might appear, and for these reasons:

1. Preservation will not apply to all your web resources, because PoWR will recommend a *selective* approach.

2. It won't necessarily mean preserving every single version of every single resource.

3. Preservation may not always mean 'keeping forever', as permanent preservation is not the only viable option.

4. Your preservation actions don't have to result in a 'perfect' solution.

Chapter 2: What's on your web?

In this chapter we outline the things we think are likely to appear on Institutional websites, and the types and location of other web-based resources. We make suggestions for the sort of information which, ideally, you would like to have available to help you start preservation activities; and suggestions for how you might collate that information.

Contents of Institutional websites

If we consider the website as a major tool of the Institution as an organisation and/or business, it is likely to contain:

- 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 outputs
- Research outputs
- Teaching outputs
- Project outputs
- Evidence of other activities (e.g. conferences)

In fact very few activities don't require a web presence, whether it is a single line or page, or a conference booking system. Many resources may already exist within a well-established managed environment, like VLEs and Institutional Repositories, but creating and maintaining a list of web-based resources is essential.

What systems have we got?

The kind of systems we would expect to find most HE institutions using are (in no particular order):

- Systems for managing assessments and examinations
- Online libraries
- Online teaching courses and course content
- Digital collections used for study
- e-learning objects and teaching materials
- e-portfolios
- Systems for managing assessments and examinations
- Blogs
- Wikis

Many will be on institutional web servers, but some may be hosted elsewhere. Some of these may contain interactive, social software, or transactional elements.

As you start thinking about ways to characterise the resources, it is important to distinguish between the following:

- Resources that are simply being accessed or delivered by a web browser. These may not be deemed web resources as such, because they are probably being managed already. The web element here is simply one of access or delivery. For example, an image collection of JPEGs, or a periodical collection in PDF form, may be accessible and delivered to students using an online catalogue with hyperlinks that connect to the resource and render the

resource onscreen. Neither the JPEGs nor the PDFs in this instance are web resources which need to be managed.
- Interactive or social software elements, which may result in outputs which require some form of preservation. This needs to be considered carefully.
- Transactional elements, which may result in outputs which require some form of preservation

Why have we got it?

As you begin to identify the web resources and various pages of the website, you may start to ask questions about who is using them and what they are doing. This divides into two pertinent questions:

(1) **Whose is it?**: Identifying relevant stakeholders: Students, academic staff, tutors, Institutional administrators, researchers, and the general public may all be making use of web resources. We will need to consider the use they are making of the resources, but also if they have a stake in the management, storage and retention of these resources.

(2) **What use are they making of the resources?**:

- What's the purpose of the activities?
- Are they creating original materials?
- Are they creating and storing records?
- How do they create the resource?

Where is it?

As already mentioned, while many resources ought to be found on institutional web servers, and in the institutional domain (usually .ac.uk), others may not be - increasingly the case since the advent of Web 2.0 and the growth of web-based cloud computing. For each resource identified, consider:

- How many domains do you have?
- Where is the Institution's web content?
- How did it get there?
- What URLs are being used?
- How many servers?
- Are backups being made?
- What Content Management Systems are we using?
- Do we have resources with external dependencies?

Most Institutions will operate more registered domains or sub-domains than just the main Institutional website. It might help to conduct a survey to establish all the URLs and domains currently being used or associated with the Institution. Some possibilities:

- Staff and student intranets
- Student portals
- VLE domains
- Separate domains for funded projects
- Museum domains

While some institutions require registration for all new websites created, it's also likely that departments and individuals are empowered to build websites as they are needed, sometimes with scant attention paid to things like corporate aims, consistent design, or record-keeping. From the first PoWR workshop, we sensed there was a general lack of centralised awareness about the number of websites and web resources in any given

Institution: "We don't know what we've got, or what people are using it for; and we don't know what to archive."

Ways of finding out

There are various ways for how you could start to whittle away at this big list of unknown quantities.

- **Conduct a survey.** This would involve approaching webmasters and stakeholders, including creators and owners of the resources. See Chapter 20: Information Lifecycle Management: Creation. It could take the form of a physical survey, visits to departments, meetings with people, a questionnaire, or extensive research. Or a combination of all of these.
- **Approach your Institutional hostmaster or Domain Name Server (DNS) manager.** This person should be able to inform you about all the URLs, domains and sub-domains which are owned, used and managed by the Institution, some of which may not be immediately obvious to you.
- **Compile an Information Asset Register (IAR).** IARs have a history in central government, where departments compile inventories of their information assets which have value to themselves, or through sharing with other departments. This is probably more of a longer-term approach than a quick win, but it is a good way of selling the idea of website and web preservation to senior management. It works from the assumption that the website and web-based resources are *assets* which have tremendous value to the Institution, hence are worthy of protection and preservation; you would be working towards bringing such resources in line with an Information Asset Management strategy.

Chapter 3: What risks and issues are peculiar to websites?

In discussions at the PoWR Workshops, and on the PoWR blog posts, the following risks and issues were identified:

Frequency of change

From the first workshop it is clear that many stakeholders in Institutions are aware that their website has changed quite dramatically in the last 5-10 years. But they all lack evidence of the changes. Agents of change can include:

- Corporate or institutional rebranding
- Move to a Content Management System
- Content provider change
- External consultancy

Quantity and range of resources

The quantity and range of resources potentially needing preservation may appear daunting. There are at least two sides to the problem: (1) knowing what there is and where it is, an issue which is partially addressed by Chapter 4: What are your web archiving requirements? and Chapter 2: What's on your web?. (2) knowing what to do about the resources, for which see Chapter 5: Selection.

Continuity

- Persistence of resources at a given URL
- Persistence of resources within a domain

Because of the ease with which websites and pages can be edited and changed, often by just one person, the possible impact on users expecting 'continuity' in web resources is easily 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. Is it even possible to support versioning across a whole site, so that old versions of a page link to contemporary versions of other pages?

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.

Ownership

- Web resources may be managed by many different departments, faculties, or members of staff
- Sub-sites may be temporary / ad hoc (for example, a project site)

Databases and deep websites

- Preserving an underlying database may not preserve user's experience on the web
- Database-driven websites are not always easy to capture by remote harvesting

Streaming and multimedia

- Quantity and quality of data; and see third-party websites, below.

Personalised websites

- Some websites offer users customisable features. Should we (even if we can) preserve every possible combination, or every user's custom view?

Third-party websites

- Groups on Facebook or Google, blogs, wikis - the content is hosted elsewhere but it constitutes valuable institutional material. How best can this be retrieved? Who 'owns' it? Is login authentication required to access some or all of the information? See Chapter 12: What about Web 2.0, and some of the case studies in Chapter 13: Scenarios and case studies.

Selection

- How to decide what pages, sites, subsites, web apps, to keep (or what bits of them)?
- Is capturing and storing everything an option?
- How to decide whether user experience (web interface) must be kept, or just underlying database/information
- Quality control/censorship

Providing access

- How to provide access to archived web resources
- IPR issues and ownership

Resources for preservation

- Personnel to undertake preservation work: preservation work can be an overhead on day-to-day web management.
- Storage space to store old versions of the websites: how can we estimate how much is required?

Resource issues apply to all digital preservation objects, and are not exclusively connected with web resources.

Chapter 4: What are your web archiving requirements?

What should be included?

Deciding on a managed set of requirements is absolutely crucial to successful web archiving. It is possible that, faced with the enormity of the task, many Institutions decide that any sort of capture and preservation action is impossible, and it is safer to do nothing. PoWR proposes that the task can be made more manageable by careful ***appraisal*** *of the web resources, a process that will result in* ***selection*** *of certain resources for inclusion in the scope of the programme. It will also help you identify those resources which can either be excluded from the programme, or at least assigned a lower priority for action.*

Appraisal and selection are disciplines borrowed from the archival and records management professions, and if successfully adapted can assist enormously in the process of decision-making. Appraisal decisions will be informed by:

- Knowledge of the Institutional structure and its aims
- Awareness of the policies and drivers for preservation
- Sound understanding of legal record-keeping requirements
- Use made of web resources
- Awareness of the stakeholders and their needs
- Potential re-use value of resources

In short, you need to understand the usage currently made of institutional websites and other web-based services, and the nature of the digital content which appears on these services. You will need to consider:

- Should the entire website be archived, or selected pages from the website?
- Could inclusion be managed on a departmental basis, prioritising some departmental pages while excluding others?

You will also be looking for unique, valuable, and unprotected resources, such as:

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

How to characterise your resources

One way to determine what kind of approaches to adopt, when considering web resources for preservation purposes, is to consider which of the following three categories best describe an object. Particularly if it is a record or a publication, it 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." (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 you need to refer to again;
- It is evidence of a transaction;
- It needs to be kept for legal reasons.

Some examples:

- Website contains the only copy of an important record. How do you know it's the only copy? If you don't know, then it shouldn't be removed or deleted carelessly unless you can establish this is the case.
- Website, or suite 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.
- Website is in itself evidence of the publication programme, or has such evidence embedded within its systems. If you need to provide, as evidence, that the Institution published a particular document on a certain date, then the logs in the CMS constitute an evidentiary 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 doesn't just mean money has changed hands). If these are transactions that require keeping for legal or evidentiary 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 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 www.nla.gov.au/services/ldeposit.html)

A web resource might be considered a publication if:

- It's a web page that's exposed to the public on the website;
- It's an attachment to a web page (e.g. a PDF or Word Document) that's exposed on the website;
- It's 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, we mean it's 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 to the original source.
- 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;
- It's an example of a significant milestone in the Institution's technical progress, for example the first instance of using a particular type of software

Artefacts preserved could include:

- Image collections
- Moving image collections
- Databases
- e-Learning objects
- Digitised objects
- Research objects

What resources can be excluded?

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, and your preservation actions are best concentrated directly on that resource, rather than on the web as a means of accessing it. This class might include:

- Digitised images
- Research databases
- Electronic journals
- Ebooks
- Digitised periodicals
- Examples of past examination papers
- 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; 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. Neither IRs nor objects stored in them need be included in the scope of your programme.

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 you ascertain that the 'somewhere else' is already being preserved, then you may not need to keep the website copies.

Web-based resources that have little or no value

Institutional Web-based applications which deliver a common service. 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 common services 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 might include Twitters and Instant Messaging, unless evidence can be found of a strong Institutional driver to retain and manage such outputs.

Chapter 5: Selection

This chapter proposes three main approaches to selection of web resources, and discusses the possibility of inscribing any decisions made within the framework of a written collection policy. The chapter also outlines the main differences between the capture of information and the capture of the web-based experience.

Among the National Libraries engaged with large-scale web archiving projects in their own country, three main approaches to selection have developed. These approaches can feasibly be adapted and scaled down to match the requirements of an HFE Institution, enabling you to decide which selection approach, or approaches, is best suited for you. (See for example T. Hallgrimsson (2008): 'International Approaches to Web archiving'.)

1. Bulk/domain harvesting

This could mean harvesting the entire website, and/or all its associated domains (which could mean targeting more than one URL).

TNA's guidance (Adrian Brown, *Archiving Websites*) would call this an "Unselective approach". It involves collecting everything possible. 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's no point in capturing 'everything' if you have already established that there are significant quantities of web resources in your Institution that do not even need capture, let alone preservation. In running a frequent domain-wide harvest of your own networks, you run the risk of creating large amounts of unsorted and potentially useless data, and commit additional resources to its storage.

2. 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 (especially if relevant to a field of research associated with your Institution)
- All resources that affect students only
- All resources that affect staff only
- All funded projects with web-based deliverables
- All resources thought to be at risk of loss
- All records
- All publications
- Resources that would most benefit an external user community (e.g. former alumni, historians)

TNA's guidance would characterise this as a 'Selective approach'. In the library and archive-based approach to web archiving, the selective approach is seen as the 'most narrowly-defined method'. Faced with the possibility of selecting external websites from the entire world-wide web for preservation in its collection, the Repository wishes to

narrow its scope by identifying very specific web resources for collection. This approach does tend to define implicit or explicit assumptions about the material that will not be selected, and therefore not preserved.

TNA also describe 'Thematic selection', which is a semi-selective approach. Selection could be based on pre-determined themes, so long as the themes are agreed as relevant and useful and will assist in the furtherance of preserving the correct resources.

The selective approach is a very library/archive based model with some form of curation implied; there is no guarantee that thematic selection alone will meet your Institution's business needs or information compliance requirements. On the other hand, it may be a good way of initiating a pilot project to find out what is possible, and achieve something manageable, with tangible results delivered quickly.

3. Event-based

National Libraries may tend to focus on world-wide or national events of importance, such as elections or disasters, which result in websites that respond to that event and thus may be (a) updated very frequently and (b) taken down after only a few weeks or months.

For an HFE Institution, the principles are broadly the same. 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 of term
- Beginning of term
- New academic year
- Appointment of a new senior official
- Departure of a professor or senior academic
- Completion of a major piece of research
- Publication of the new prospectus
- Purchase of new authoring software which affects web content
- Corporate or institutional rebranding
- Formation of a new department

Creating a collection policy

There may also be some value in scoping out a collection policy to decide which aspects of web resources need to be collected and preserved. A collection policy is not the same as defining your retention requirements, nor about assessing the value of resources, which are covered elsewhere (see Chapter 11, for example). But a collection policy could feasibly dovetail with other information management policies (e.g. records management, archival selection) within the Institution.

TNA's advice on a collection policy is in two parts: (a) devise a selection policy and (b) build a collection list. This could feasibly be scaled down and adapted to work in an HFE environment. Note that TNA's advice does not explicitly include any records management requirements as selection drivers.

Policy definition: defining a selection policy in line with your institutional preservation requirements. The policy could be placed within the context of high-level organisational policies, and aligned with any relevant or analogous existing policies. (See Chapter 16: Responsibility for preservation of web resources and Chapter 18: What policies exist?).

The policy will result in a collection list, which provides the basis for undertaking collection of the web resources. The boundaries of the resources on the list should be defined, and appropriate timing and frequency should be specified.

How to decide what *aspects* of web resources must be captured

It is possible to make a distinction between preserving an *experience* and preserving the *information* which the experience makes available. Both are valid preservation approaches and both achieve different ends. Putting it very simply:

- **Information** is all meaningful content (including words, figures, images, audio)
- **Experience** means the experience of accessing that content on the web, which includes all its attendant behaviours and aspects.

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

A few examples:

Evidential and record-keeping: As well as the content, you would need to preserve some form of change history, with as much contextual information as possible. This may not apply to all the web resources, just to ones which are needed for legal purposes, to protect the Institution, where decision-making is involved, etc. For such resources you would want to capture and preserve:

- An audit trail of changes
- A change history
- Contextual information about people - who wrote it, used it, added to it
- Contextual information about dates and times - date written, date changed, date published, date removed, how long a page was published and exposed, when it was taken down
- The content of the resource
- The appearance of the resource
- The 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 of the resource
- The appearance of the resource
- The 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 tell you 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, you would want to preserve the appearance of the resource so as to demonstrate how home pages looked five years ago.

Chapter 6: Web Capture: what and how?

What elements do we capture and preserve?

Capture is not preservation: simply capturing a resource in any form or format does not guarantee its survival. But it is a necessary step. In this chapter we explain certain capture strategies which are possible from a technological point of view. What approach to capturing you choose may be dependent on the content, and the policy decisions that your institution has already made governing records and publications. It can also be considered as a selection question: for more information on selection, see Chapter 5: Selection (How to decide what aspects of web resources must be captured).

The elements of web resources that need to be considered are:

Content

If content is all you want, just capture the content. This involves simply extracting content of a page, e.g. just the words. No links, no behaviour, no framesets, no stylesheets, no images - just plain text.

Appearance

If appearance is important as well as the content, then tools exist for capturing just that. Some snapshot-type capture tools (see Chapter 7: Tools for the job) can turn a web page, or sometimes even an entire website, into a static PDF file. This preserves some elements, including some navigational links. But certain things, including web behaviours, will disappear.

Behaviour

If you need to preserve content, appearance and behaviour, the job becomes more complex. Many websites and web pages are still essentially a simple mixture of text, images, and links. But others have more dynamic and 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, offering live linked information about users;
- Recent comments, which can change and update even when the main content remains static;
- Behaviours connected with management of the blog, such as site administration;
- Bookmark and tagging features, which may connect out of the blog to other related services.

It may not be feasible or desirable to capture all of these features, for example, one would not usually expect website administration pages to continue in the archived instance. Therefore it is important to specify which are most significant for preservation.

How do we capture web resources?

It is possible to imagine three points in the journey of a web page from server to user, where its capture is likely to be most feasible and fruitful. Each of these, inevitably, has advantages and disadvantages.

Capture within the authoring system or server

This would involve retrieving web pages directly at their point of origin, usually the Content Management System, or the server on which web pages are held. Doing this obviously requires access to the originating web server. But can the CMS-generated content be readily saved? You can certainly get all the content (HTML, CSS, GIFs, JPEGs, etc), but how will this relate to the way end-users experience it? Increasingly, web pages are formatted 'on-the-fly' to suit the specific needs of the browser that is requesting them (e.g. Firefox or Internet Explorer, small screen or large screen, desktop device or mobile phone): which of these possible versions should be captured?

- ✓ Likely to be easy to perform, if you own the server
- ✓ Works in short to medium term, for internal purposes
- × Captures raw information, not presentation
- × May be too dependent on authoring infrastructure or CMS
- × Not good for external access

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.

- ✓ You get what you see (but you don't necessarily know why)
- ✓ It's relatively simple for well-contained sites
- ✓ Commercial tools for doing it exist
- × Treats web content like a publication: frozen
- × Loses behaviour and other attributes

Harvest content with crawlers

Using a crawler is going to resolve some of the problems of other methods, but not all of them: crawlers are unlikely to succeed totally. They miss other external sources such as document servers, databases and datafeeds. Internal databases, subscription databases, file management platforms, and website content management systems may present problems to crawlers. 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 you don't need, by blindly following links. There are ways of setting exclusion filters to prevent this, but its behaviour can still be unexpected.

- ✓ Most widely-used
- ✓ Defers some access issues:
  - ✓ Link re-writing
  - ✓ Embedded external content: from archive or live?
- × Lots of work, tools and experience necessary
- × Presents many problems for capture: often don't get everything, or they get too much.

Chapter 7: Tools for the job

Although this Handbook cannot offer a complete how-to guide for these tool, this chapter describes some of the tools that exist, with a brief outline of their key functions and features. Netpreserve.org maintains a list of available tools; Harvard University Library have also published a good list of web preservation tools (and many other related resources).

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: Web capture: what and how?

Heritrix

Heritrix is a free, open-source, extensible, archiving quality Web crawler. It is used by the Internet Archive, and is freely available for download and use in your own web archiving projects, under the terms of the GNU LGPL [http://crawler.archive.org/license.html]. It is implemented in Java, and can therefore run on any system that supports Java (Windows, Apple, Linux/Unix). It was developed by the Internet Archive with the Nordic National Libraries.

- 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 [www.httrack.com/page/5/en/index.html]. Distributions are available for Windows, Apple, and Linux/Unix.

HTTrack allows you to download a World Wide Web site 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: www.httrack.com/
- Download: www.httrack.com/page/2/en/index.html

Wget

GNU Wget is a free software package for retrieving files using HTTP, HTTPS and FTP, the most widely-used Internet protocols. 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 avaiable under the GNU GPL: versions are available for Windows, Apple and Linux/Unix.

GNU Wget's features include

- Converting absolute links in downloaded documents to relative, so that

downloaded documents may link to each other locally
- Using filename wild cards, and recursively mirroring directories
- Resuming aborted downloads
- Multilingual message files
- Support for cookies, proxies and persistent HTTP connections
- Using local file timestamps to determine whether documents need to be re-downloaded when mirroring

- More information: www.gnu.org/software/wget/
- Download: www.gnu.org/software/software.html

DeepArc

DeepArc was developped 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 key in 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.

- More information: http://bibnum.bnf.fr/downloads/deeparc/
- Download: http://sourceforge.net/projects/deeparc/

Workflow systems or curatorial tools

Workflow and curatorial tools are generically tools 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 a tool 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 workflow system and a curatorial 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.

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 Archiving Consortium. (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 following an unsuccessful attempt to find an off-the-shelf system (or systems) to provide an integrated, web-based, web archiving management system. The need for such a system was evident as the scale of the Library's archiving activity increased and if the best possible efficiencies were to be achieved in building a collaborative, selective and quality assessed web archive. It was also necessary to enable other PANDORA participants to contribute to the Archive from various geographic locations. PANDAS was first implemented in June 2001, and a second much-enhanced version was released in August 2002. PANDAS version 3, a completely re-engineered and enhanced version of the software was deployed on 27 June 2007.

Like WCT, PANDAS is a workflow system; the actual crawling is done by HTTrack. PANDAS was used by the UK Web Archiving Consortium 2004-2007, a process which revealed a lots of its flaws. PANDAS 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. Potential users may wish to bear in mind that the tool has a very strong bias towards library models (it was built for the National Library, and treats websites and web pages as titles that have authors and subjects).

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.

- More information: http://pandora.nla.gov.au/pandas.html

NetarchiveSuite

NetarchiveSuite is a curator tool allowing 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.

Developed by the Royal Library and the State and University Library in the virtual organisation netarchive.dk

- More information and download: http://netarchive.dk/suite

Snapshot tools

Adobe Acrobat web capture tool

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 you to convert the currently displayed web page to a tagged Adobe PDF file.

The Internet Explorer Adobe PDF toolbar preferences determine only whether converted files open in Acrobat automatically, and whether you are prompted to confirm the deletion of files or addition of pages to an existing PDF file.

The Acrobat web page conversion settings, which are available only in Acrobat, let you set more advanced settings, including the creation of bookmarks and tags. After you set

the Acrobat web page conversion settings as desired, you need to use the Create PDF From Web Page feature in Acrobat at least once before the settings take effect in the Internet Explorer web page conversion feature.

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. PDF's print-oriented format isn't a good match to some sites, much as some sites don't look good when you try to print them. 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.

More information:

- www.document-solutions.com/accessibility/AdobeAccessChapter3a11.html
- www.wap.org/journal/acrobat4capture.html
- www.planetpdf.com/enterprise/article.asp?ContentID=6057

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. "I attended a presentation on a.nnotate at the Repository Fringe event where they suggested they had used OO to provide platform-independent, collaborative annotation of web pages." (Knight 2008)

A.nnotate

A.nnotate will let you do web page capture. "You can enter a URL or use a bookmarklet to take a snapshot of a web page and store a copy of the HTML in your private space on the a.nnotate.com site - which can be useful for archiving, as you get the page at a particular point in time. Currently it does a shallow copy (i.e. just the HTML) - the images are left on the original site, so it would need to download those too if you wanted to use it for archiving. The A.nnotate server is also available for local installation (with an API) if you want to integrate it with some other CMS. You can also upload PDFs to A.nnotate (we use OpenOffice to convert from the various Office formats to PDF first) and these get converted to images and rendered in the browser using pure HTML / AJAX (without any dependency on Flash or Adobe reader)." (Howell, Fred 2008)

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.

- More information:
  http://graphicssoft.about.com/od/screencapture/gr/snagit.htm

Chapter 8: Web Content Management Systems (CMS)

The chapter is intended to provide a little insight into the way a Content Management System behaves, and consequent preservation issues. Not all CMS systems are the same; some have change and edit histories built into them, some don't. Web managers will already be aware of what the CMS does and what's in it, but preservation-related features are not likely to feature high on the list of essential features.

CMS in institutions

A CMS is essentially 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 fiddly code in HTML headers, Javascript and CSS files. CMS can manage pages in a hierarchy, and generate page content dynamically by combining content with templates.

CMS can be implemented and administered in many different ways. While the overall structure and design of a website is generally managed by the Web and Marketing Teams, It is common, for example, to allocate 'sections' of the website to departments or other major institutional functions, and appoint 'sub-editors' responsible for the content in those sections. However, in some institutions, all changes may be routed for approval through a single Web Editor.

There are so many CMS available that it would be impossible to consider them all from a preservation viewpoint, whether commercial offerings (e.g. Red Dot, Sitecore, Sharepoint) or open source systems (e.g. TYPO3, Drupal, Joomla).

CMS and preservation

With digital preservation drivers in mind, we should consider generically what features they may offer. Of particular value would be:

- 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 CMS offer one or more of these features, but do these features work? The extent to which they can be easily used, to reinstate older versions of a website, or easily 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 rigorously.

Preservation issues presented by CMS

Page names and numbers

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

Results will also depend on which harvesting engine you decide to use: in some circumstances a gather may end up incomplete as it misses pages, and the crawler may get stuck in a 'loop' as it constantly requests pages.

Rollback function is limited

In TYPO3, for example, a rollback is not the same as restoring a full snapshot, and you can't use it to view the content of the old page. The content is held in the database, as layers of time-stamped pages. To access content from a CMS would require a script to retrieve it from the database. In short, it's not clear to what extent the rollback functions and version control tools produce useful, tangible outputs that could be captured, managed, or preserved.

Rollback will tend to focus on a particular page or content element, but not its entire context. Web pages unfortunately rarely stand in isolation, and many objects that they relate to - for example embedded images and stylesheets, or other pages that they link to - may also change. Therefore in order to fully restore one page to the state it was a month ago, we may have to ensure that related objects are also in that same month-old state, otherwise we may merely have created a meaningless hybrid of old and new content.

Lifespan of system

This may be something to consider as a preservation issue, as indeed it is an issue with any other database of any form of software. Valid questions include:

- How long will it be supported?
- Will the new version be compatible with the old version?
- Are you confident you can migrate your old website content into the new system?
- A lot of website designers see a new version as an opportunity to start from scratch. But what about the content?

There should be a recognised Institutional responsibility to maintain the CMS.

Compatibility between systems

This could also be an issue for preservation purposes. How easy is it to migrate a website from one CMS system to another? CMS internal management of content, data and metadata tend to be application-specific, making it a non-trivial task to move from one CMS to another: moving large quantities of interlinked website content between CMS packages is likely to be a manual and intensive process.

Summary

- CMS is a database full of content, but simply backing up the database (though backup should certainly be performed) 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 useable collection of page snapshots, or an archived website.
- The change history metadata would be extremely useful for records management and preservation purposes, but access to that metadata is not guaranteed: we would need to be able to export it in a form that would be preservable.
- If there is a known issue with lack of forward compatibility in your CMS, this would seem to put most, if not all the content at risk.
- Remote harvesting of the website itself is likely to be the best way to capture the website and circumvent these issues (but not if the CMS is going to defeat the crawl engine).

Chapter 9: What approaches and techniques can you use?

Successful web preservation is most likely to be achieved by a mix of skills from Information Management professionals: some asset management, some records management, some archival awareness; input from webmaster; awareness of storage costs. A variety of approaches are described below, drawing on aspects of information professions - records management, archival preservation and management, and information lifecycle management.

Two main classes of approach are listed:

1) **What to do now.** This includes quick win solutions, actions that can be performed to achieve effective results in the short-term, and to rescue or protect resources identified as being most at risk. Actions include domain harvesting, remote harvesting, pilot projects, and use of 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 approaches.** This includes longer-term strategic 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 approaches which involve working with external organisations to do the work (or some of it) for you. 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, rather than something which requires reactive responses or constant maintenance, both of which can be expensive, resource-hungry and risky methods.

What to do now

Domain harvesting

This could refer to two possible approaches:

1. The Institution conducts its own domain harvest, sweeping the entire domain (or domains), using appropriate web-crawling tools.

2. The Institution works in partnership with an external agency to do domain harvesting on its behalf; see Chapter 21: Can other people do it for you?

Domain harvesting is only ever a partial solution to the preservation of web content. Firstly, there are limitations to the systems which currently exist (see Chapters 6 and 7). You may gather too much, including pages and content that you don't need to preserve. Conversely, you may miss out things which ought to be collected - hidden links, secure and encrypted pages, external domains, database-driven content, and databases. Secondly, simply harvesting the material and storing a copy of it may not address all the issues associated with preservation.

From the first workshop, we heard some support for the idea of 'indiscriminate harvesting' on a regular basis, with the expectation that somebody would sort out the harvests later. An analogy was made with boxes of paper collected from the offices of

retiring academics. But the analogy doesn't work for web resources: selective collection of managed web resources is preferable to indiscriminate harvesting, if only because the sheer volume of unsorted resources quickly becomes unmanageable.

Pilot projects

Instead of trying to solve the web resource problem for an entire Institution, initiate a pilot project. The attraction of doing this is that you get can get a visible result quite quickly. The results may make it more persuasive for other departments to participate in web archiving, and add credibility to the programme. Pilot projects can also generate useful reports about lessons-learned, that will prepare you for pitfalls and make the next project even more successful. Another advantage of a pilot project is that fewer stakeholders may be involved, if you scope the project tightly enough, thus saving time on consulting users and owners of the content. See also Chapter 17: Institutional Strategy and Chapter 5: Selection.

Pilot projects could be targeted at, for example:

- One Department in the Institution, archiving their suite of pages, or project sites, or all web resources owned by that Department;
- One type of resource, such as those clearly identifiable as records, or as publications;
- Resources known to be orphaned, unmanaged or unprotected, and therefore at greatest risk;
- Selected contents of the CMS, for example public-facing pages only.

Migration

Migration of resources is a form of preservation. Migration means moving resources from one operating system to another, or from one storage/management system to another. 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?

We don't yet know how feasible it is to use an EDRMS for management of web resources. ERM 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. (See also Marieke Guy, 'When Do We Do Fixity', on the PoWR blog.)

We know that's 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.

See Appendix B: Records management: A guide for webmasters.

Strategic Approaches

Information Lifecycle guidance

Information Lifecycle Management (ILM) involves recognised professional standards and practices, leading to better management of information, and is one possible approach. If we can apply a lifecycle model to web resources, they will be created, managed, stored and disposed of in a more efficient and consistent way; it can 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's a conceptual framework to help ensure consistency within an organisation. It can be especially helpful when introducing new systems, or reviewing existing ones.

There is lots of literature available. Beginners could do worse than look at the JISCInfoNet published guidance, *Managing The Information Lifecycle*, which is geared towards the HFE sector.

Information moves through a series of phases over time. JISC's approach to ILM proposes four distinct phases:

- Creation
- Active use
- Semi-active use
- 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. You can adapt or vary a model according to your institutional needs. The model should have a chronological structure, clearly defined phases, user identification, and consistency. See also Chapter 20: Information Lifecycle Management: Creation.

Adapting records management approaches

Records management is suggested as another possible approach. If we can apply a records model 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. It will help the Institution with information legislation compliance. Work with your records manager. Use the JISCInfoNet published guidance, for example; and the guidance from Edinburgh University's Records Management Department. See also Appendix B: Records management: A guide for webmasters.

Continuity and maintenance

The **Web Continuity project** at The National Archives is a large-scale and Government-centric project, which includes a "comprehensive archiving of the government web estate by The National Archives". Its aims are to address both 'persistence' and 'preservation' in a way that is seamless and robust: in many ways, 'continuity' seems a very apposite

concept with which to address the particular nature of web resources. Many of the issues facing departmental web and information managers are likely to have analogies in HE and FE institutions, and Web Continuity offers concepts and ways of working that may be worth considering and may be adaptable to a web archiving programme in an Institution.

A main area of focus for Web Continuity is *integrity of website links*. Their use of digital object identifiers (DOIs) can marry a live URL to a persistent identifier. To achieve persistency of links, they use a redirection component which is derived from open-source software. It can be installed on common web server applications, e.g. Apache and Microsoft IIS. 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.

They are building a centralised registry database, which is growing into an authority record of Government websites, including other useful contextual and technical detail (and can be updated by Departmental webmasters). It is a means of auditing the website crawls that are undertaken. Such a registry approach would be well worth considering on a smaller scale for a Institution.

Their sitemap implementation plan involves the rollout of XML sitemaps across government. XML sitemaps can help archiving, because they help to expose hidden content that is not linked to by navigation, or dynamic pages created by a CMS or database. This methodology may be something for HFE webmasters to consider, as it would assist with remote harvesting by an agreed third party.

The intended presentation method will make it much clearer to users that they are accessing an archived page instead of a live one. Indeed, user experience has been a large driver for this project. UK Government want to ensure that the public can trust the information they find and that the frustrating experience of meeting dead-ends in the form of dead links is minimised. Further, it does something to address any potential liability issues arising from members of public accessing - and possibly acting upon - outdated information.

Protection / maintenance

Protection must include protection from careless or wrongful destruction, deletion, or removal of the resource. The danger of deletion or removal may arise when a website is rebranded or relaunched; when certain pages appear to lack owners who might defend them; when academic staff move on to other jobs or positions; when pages are apparently no longer being accessed; or when administrators have a spring-clean of the hard drive.

Chapter 10: Who wants to keep what, why, and for how long?

This involves an understanding of retention requirements, the needs of stakeholders, and internal and external drivers within your Institution. The traditional archival skills of appraisal and selection can also help. The more we know about retention, the more it helps you determine which solution or approach, or combination of them, is appropriate for your web resources.

The Handbook recommends a *selective* approach. 'Keeping everything' is not a recommended option. Even 'capturing everything' is difficult enough when it comes to websites. Superficially, the cost of digital storage appears cheap: managing it, however, is not. The cost of storing accumulated copies of snapshots of your website will quickly become prohibitive, particularly if you haven't been selective, and opt for a domain harvest on a regular basis. The more material that is kept, especially if the storage is not managed, then the harder it becomes to locate and retrieve important information when needed.

It is worth bearing in mind the records management perspective. The more you keep, the more material you may have to disclose under the Freedom of Information Act. If you manage to get your retention decisions agreed at a senior level as part of the web management policy, and align them with records management schedules where appropriate, you will be in a much stronger position as regards FOI compliance, and your storage will be more cost-effective.

Internal drivers

Internal drivers for keeping web resources include:

- Operational, organisational and institutional considerations
- Resources are needed for the purpose they were created for
- Resources are needed for reuse and repurposing
- Resources are needed for heritage and historical value
- Money spent on their creation is wasted if you destroy carelessly

Consulting stakeholders

It is also helpful and good practice to identify stakeholders who have a legitimate interest in retaining the web resources. As part of the surveying process or general picture you are building up of the Institutional web collections, find out:

- What resources stakeholders want kept
- Why stakeholders want the resources kept
- How long they want them kept for

This will help get people on board; staff will feel less alienated if you can align web management with the things they are actually doing and working on. It will embed the notion of good practice and web management within the Institution, and start to get preservation included in the workflow.

While it is very important to ask other stakeholders for their view on retention, it's sometimes important to maintain a healthy scepticism. Many administrators consider everything they do to be important; and they assume you're going to keep it forever. Conversely, it's all too easy to sweep things away simply because the stakeholder isn't around to defend their own interests any more. Web resources are particularly

vulnerable. A project manager's website is taken down - and all its contents deleted - because the project ends, or because the project manager has left the institution. A researcher's blog which represents two years of accumulated wisdom from the researcher and their students, is thrown simply because when the researcher leaves, their computer accounts are dismantled.

External drivers

External drivers for keeping things include the legal and regulatory requirements for retention. Speak to your records manager about the Statute of Limitations, and how it affects many aspects of record-keeping. Protecting the Institution's reputation is an important consideration, and this also shades into risk management. Consider once again your website Publication Scheme, and ask yourself if you would benefit from being able to access the history of exactly what was published on the website, and when. It may protect the Institution from liabilities.

Chapter 11: How do we appraise the value of a web resource?

Chapter 10: Who wants to keep what, why, and for how long? deals with stakeholder requirements and internal and external drivers for retention. But in cases where there are no clear and compelling reasons for retention, you need to assess the value of web resources in an objective way, thus ensuring that the value of the resource justifies the costs of continued retention.

Some questions to help you decide:

- Is the resource needed by staff to perform a specific task?
- Has the resource been accessed in the last six months?
- Is the resource the only known copy, or the only way to access the content?
- Is the resource part of the Institution's web publication scheme?
- Can the resource be re-used or repurposed?
- Is the resource required for audit purposes?
- Are there legal reasons for keeping the resource?
- Does the resource represent a significant financial investment in terms of staff cost and time spent creating it?
- Does it have potential heritage or historical value?

DPC decision tree

Another potentially useful tool is the Decision Tree produced by the Digital Preservation Coalition. It is intended to help you build a selection policy for digital resources, although we should point 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.

> *"Clearly defined selection policies will enable cost savings in terms of time taken to establish whether or not to select and also potential costs further down the track of needing to re-assess digital resources which are either in danger of becoming or are no longer accessible. This Decision Tree may be used as a tool to construct or test such a policy for your organisation. The decision process represented in the tree should be addressed by your policy for selection of digital materials for the long-term."*

> *"Assuming a digital resource is being considered for selection, the questions and choices reflected here will assist the ultimate decision to accept or reject long-term preservation responsibility. The flow of the questions represents a logical order of evaluation. If the response to early questions is not favourable there is little point in accepting preservation responsibility for the resource or continuing its evaluation, for example if the content does not meet your collection policy then the response to questions on the technical format will be irrelevant. The structure of the tree aims to reflect this process." DPC 2008.*

Archival appraisal

Traditional archival appraisal remains one of the core skills of the professional archivist. The usual aim of archival appraisal has been to identify and select records for permanent preservation. Quite often appraisal has taken place at the very end of the lifecycle process (although records managers intervene where possible at the beginning of the process, enabling records of importance to be identified early).

Appraisal looks for records which will build a comprehensive 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
- a member of local, national and international communities
- a community in itself

The records 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 of its impact locally and in the wider world. The selection process should also facilitate the survival of records which contain unique information incidental to their main purpose or function but which, nevertheless, might have research value. This approach is not unique to HEIs but is common to all organisations and similar records have the same value in all organisations, irrespective of what they were set up to do.

In simple terms, appraisal of HEI records for permanent preservation should focus on:

- substantive functions (i.e. Teaching, Research, Academic Award Administration)
- 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*)

Chapter 12: What about Web 2.0?

Overview

We have become increasingly familiar with the term Web 2.0, referring in a very general way to the recent explosion of highly interactive and personalised web services and applications, from blogs and wikis to online services like Flickr, Twitter and Slideshare. Collaboration and social networking are a key feature, for example through contributing comments (blogs, Flickr, Facebook) or sharing write access and collaborating (wikis, Wetpaint, Google Docs). Highly tactile and responsive interfaces (using AJAX) are also a common feature of Web 2.0 applications. Colleagues can create, share and store information in any number of web-hosted packages. The information is not held on the Institutions' servers.

Many of these applications have now crossed the threshold between private, personal use and applications in business and education; others are falling over themselves to do so. Web 2.0 tilts the balance of power in relation to information away from the organisation towards the individual. Students and staff continue to innovate new, individualistic, extramural ways of working: a research, study or project group could quickly and easily create a personalised environment using Delicious or CiteULike? to store bookmarks and bibliography, Google Docs or a wiki to collaboratively develop and edit documents, Skype or Instant Messaging for discussions, and a blog to keep a journal of activities and progress. Anyone can be an author, a contributor, or a commentator. All of this without any reference to institutional IT provision.

Some HE institutions have begun installing Web 2.0 applications in their own domains (Warwick Blogs, for example), which would potentially allow preservation activities to be targeted at the hosting server and application. However IT strategy is always going to lag behind the pace of innovation on the web at large, and staff and students will continue to be impatient to use new and exciting ways to work, on third-party hosted services. Where a preservation need is identified, 'off-air' archiving (i.e. capture by remote harvesting), though imperfect, is likely to be the most effective approach (as long as it is not inhibited by factors such as user account login, streaming content, and terms of use agreements).

Types of Web 2.0 application

In a recent briefing paper for JISC, Mark van Harmelen defined seven types of Web 2.0 application:

1. Blogs
2. Wikis
3. Social bookmarking
4. Media sharing services
5. Social networking systems
6. Collaborative editing tools
7. Syndication and notification technologies

In addition, Instant Messaging, while nothing new as a standalone application (IRC, Windows Messenger, etc.), has taken on a 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). Nevertheless, this is a useful model for considering the differing application requirements and preservation issues.

1. Blogs

Examples: Blogger, Wordpress, Edublogs, Warwick Blogs

Uses: Publishing online journals for a wide variety of information dissemination purposes. Often readers can leave comments on individual entries.

2. Wikis

Examples: Mediawiki, Wetpaint, Tiddlywiki

Uses: Collaboratively creating online hypertexts, electronic research or reference resources, mini-websites, class projects.

Advantages: Quick, easy, free to set up. Enable collaboration and long-distance working. Flexible regarding the management of attachments. Built-in version control and rollback features. Some wikis can feasibly be used as a form of EDRMS.

3. Social bookmarking

Examples: Delicious, CiteULike, Connotea

Uses: 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.

4. Media sharing services

Examples: Flickr, Slideshare, YouTube. (Scribd, DeviantArt)

Uses: Galleries of images and videos; sharing presentations. Sharing multimedia resources for teaching and research, including podcasts and videos of lectures etc.

Advantages: Quick, easy and free to set up. Obviates the immediate need to find server space for large resources that need to be shared quickly. Many of these services allow tagging and comments on the resource.

5. Social networking systems

Examples: Facebook, Ning, Elgg, Crowdvine, LinkedIn

Uses: Allowing 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 conference, can continue online, and vice-versa.

Advantages: Quick and easy and free to set up. Brand recognition leads to widespread use.

6. Collaborative editing tools

Examples: Google Docs

Uses: Collaborating to edit documents and spreadsheets, allowing users in different locations to collaboratively edit the same document at the same time.

Advantages: Quick and easy and free to set up; change-tracking; location and software independent. Ease of retrieval and migration / export to other formats / locations.

7. Syndication and notification technologies

Examples: Netvibes, Technorati

Uses: 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.

Advantages: Creating a dynamic, personal online environment embedding or linking to commonly used web resources. Quick and easy and free to set up.

8. Instant Messaging

Examples: Facebook Chat, Google Chat, Skype, Jabber, Windows Messenger

Uses: Informal messages for conducting informal business. However, it is possible that a thread starts as an informal chat and develops into something more formal along the way. It's possible to decide in advance that you're going to use IM for formal work, thus obliging a record-keeping step. See also the Case study on Preservation and Instant Messaging.

Creation and preservation issues

Some of these applications and services listed above are still at an 'experimental' stage and (at time of writing) being used in Institutions primarily by early adopters of new technologies. But it is possible to discern the same underlying issues with all these applications, regardless of the software or its outputs. The two most important ones are ownership and retention.

Ownership and responsibility

Quite often in an academic context these applications rely on the individual to create and manage their own resources. A likely scenario is that the academic, staff member or student creates and manages his or her own external accounts in Flickr, Slideshare or Wordpress.com; but they are not Institutional accounts. It is thus possible with Web 2.0 application for academics to conduct a significant amount of Institutional business outside of any known Institution network. The Institution either doesn't 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.

Because of the nature of the relationship between naive but enthusiastic users and external service providers, undesirable outcomes may result. For example, by joining and uploading material to Facebook, many users are unwittingly accepting the following agreement that apparently permits Facebook to do whatever it wants with it - a far cry from Creative Commons:

> *By posting User Content to any part of the Site, you automatically grant, and you represent and warrant that you have the right to grant, to the Company an irrevocable, perpetual, non-exclusive, transferable, fully paid, worldwide license (with the right to*

> *sublicense) to use, copy, publicly perform, publicly display, reformat, translate, excerpt (in whole or in part) and distribute such User Content for any purpose, commercial, advertising, or otherwise, on or in connection with the Site or the promotion thereof, to prepare derivative works of, or incorporate into other works, such User Content, and to grant and authorize sublicenses of the foregoing.*

By contrast, one would expect blogs and wikis hosted by the institution to offer more acceptable terms of use, in line with existing policies on rights, retention and reuse, as expressed in IT and information policy, conditions of employment or matriculation, etc.

Retention of 'master copies'

Third-party sites such as Slideshare or YouTube are excellent for dissemination, but they cannot be relied on to preserve your materials permanently. If you have created a resource - slideshow, moving image, audio, whatever it be - that requires retention or preservation, then someone needs to make arrangements for the 'master copy'.

Ideally, you want to bring these arrangements in line with the larger web archiving programme. However, if there is a need for short-term action, and the amount of resources involved are (though important) relatively small, then remedial action for master copies may be appropriate. Some possible remedial actions are:

- Store it in the EDRMS
- Store it on the Institution website
- Store it in the IR
- Store it on a local networked drive

It's important to ensure that materials stored in these locations are being preserved, or managed in some way. For that reason, don't use untrusted storage methods, such as storing the resource on your C Drive (where it won't be backed up) or burn it to a disk. As suggested, this kind of remedial action goes against the underlying intentions of the Handbook, and we recommend that retention and preservation activities are carried out within an agreed records management or asset management framework. Failing that, create a local written policy for your external activities that explains what you and your department are doing, and which could be used to demonstrate some form of IM compliance. (See MacGlone, 2008)

In the case of blogs, wikis and collaborative tools, content is created directly in them, and access is entirely dependent on the availability of the host and the continued functioning of the software. Users of such tools should be encouraged and assisted to ensure significant outputs of online collaborative work are exported and managed locally.

Chapter 13: Scenarios and case studies

Scenario: Home page history

***Description**: Your institution is about to commemorate an important anniversary (10 years, 50 years or 250 years since it was founded). Your VC wants to highlight the fact that the institution is actively engaging with new technologies, and 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 has been taking 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 the Internet Archive's WayBack Machine to illustrate how it had changed between 1997 and 2007: an animated visualisation of the changes, linking to the IA's snapshots, is available at 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 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 Archiving Consortium. See Chapter 21: : Can other people do it for you?
- 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, while 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, /virginia2000, etc.). Some years are missing: whether because the changes were insignificant, or no copy survives is not clear. Although there are many dead links in the archived sites, or anachronistic links to current versions of pages, the archived snapshots provide a valuable view of the evolution of the institution's web presence.

Scenario: Putting the prospectus online

Description: *The Institution currently provides prospective students with a printed copy of its prospectus, while only limited information on courses has been made available on the institutional website. This information is fairly general to avoid legal vulnerabilities and to minimise maintenance issues. There is growing pressure to improve the online prospectus content, and even make it comprehensive enough to supersede the print version. The Web Manager is keen to pursue this, but is unsure of how to proceed.*

Issues

- Who has responsibility for such a project? The creation, editing and publishing of an online prospectus will involve shared ownership, including for example Marketing, the Academic Office, the Undergraduate Admissions office, the Publications office, and the Web team.
- Creation and ongoing management of an online prospectus is likely to present not only technical challenges, but also necessitate cultural and administrative change. The content will be updated much more frequently by many contributors. How will old versions of the prospectus (or parts of it) be stored, managed and accessed?
- Students will be making decisions about their academic career based on what they find in this prospectus. It may be important for the Institution to know precisely when it was made available online, and precisely what content was viewable within that timeframe. Ideally, you would want to be able to access time-based snapshots of this part of the website, or previous (dated) versions of the prospectus.
- Discussions at the third PoWR workshop revealed that the distinctions between a printed prospectus and an online version are increasingly becoming blurred. Some years ago, it used to be common practice to copy and paste a static document into a web page; now it's more likely that the online version will be 'poured' into a printable document. Some say the online version has supplemental material, while others see it as subordinate to the printed version.

Approaches

- Creating an online prospectus may simply be matter of creating PDF versions of the document, not dissimilar to the paper versions, and attaching the PDFs to the website. This should be considered part of the Institution's publication programme (and ought to be declared in the FOIA Publication Scheme). It will need to be managed, and copies of this serial publication will still need to be retained (probably permanently) by the archivist.
- Another approach is to use an online publishing system, which may have automated version control and a facility for storing and retaining backup copies. Presumably it would also allow content to be dynamically updated. You may need a method to ensure that versions and changes can be captured, preserved, and subsequently accessed.

Scenario: Vanishing domain names

Description: *A project team in the Institution has purchased a domain in the .org TLD, outside of 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 now 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: if it did, then this will probably be a result of major changes affecting your institution.

If, however, you are using an alternative domain name (such as a .org, .org.uk, .co.uk or .com) then you will need to take care in managing the registration for your domain. If you have an annual re-registration for your domain, you will need to ensure that your internal administrative management procedures will ensure that the domain name is renewed prior to the expiry.

You may ask why you would wish to make use of a non-.ac.uk domain in light of such possible dangers. JANET, the organisation responsible for managing .ac.uk domains, does not sell off its domains to the highest bidder. It does, however, have strict eligibility guidelines that may not be met by short-term or collaborative or cross-sectoral projects and services, that may involve many institutions, some international. Equally within institutions, the allocation of fourth-level subdomains (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.

Scenario: Student blogs

Description: *Your Institution runs a blog service, that all students and staff can use for personal and social activities, or research and study projects. One enthusiastic alumna wants to migrate the extensive blog she has kept for three years, but your institution, like many others, 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 content of their blog (including embedded resources, comments, etc.)?*

Issues

Blogs may contain valuable discussions on important topics, as well as reflecting the intellectual and social life of the institution. The most-read blogs typically accumulate a large number of links to them, many from outside the institution's domain. These links will die if the blog is deleted, or if it is moved to another location (e.g. to an archive or alumni subdomain).

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?

Students are increasingly encouraged to use blogging as a way of reflecting on their experience, but why should anyone invest effort in the construction of an artefact, and believe that that effort is valued, if no thought has been given to its preservation and continuation. The absence of some kind of migration or continuity option might create motivational and validity issues that could undermine the value of the facility.

Does an institution have permission to archive the content of blogs (and make it available elsewhere)? This might include permission not only from the blog author (which might be obtainable in the general terms and conditions of registering with the Institution), but also third-party content: embedded quotes, images, audio, video. 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 as well informed about the issues of online copyright as they are expected to be about plagiarism, citation or photocopying regulations? 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 for its own students, or to use third-party providers such as Blogger.com or Wordpress.com? In the latter case, resources created can persist at the discretion of the student (and the third-party host), independently of institutional policy.

Approaches

- Decide that the issue is predominantly one of policy, not of self-hosting versus third-party hosting. If an educational institution is encouraging 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 members of the institution - such as 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 T&C and functionality very carefully to ensure they meet standards it can recommend to those in its charge. Blogger,

Wordpress.com and Facebook are very general 'tools', and a particular institution might legitimately want something more tailored - like Edublogs, ELGG, Club Penguin even - or something truly bespoke.

- Seek permission from the owners of the blog content before making copies. Investigate wider application of Creative Commons licences. Work towards resolving third-party issues.

Case study: The Vanishing Blog

Background

The e-learning team at the University of Bath set up a blog called Auricle in early 2004. The blog was hosted in the bath.ac.uk domain. Derek Morrison, head of the e-learning unit, was interested to explore the potential of new technologies, one example of this being the series of podcast interviews he recorded and made available on the blog in 2005.

During the course of the JISC-PoWR project, Brian Kelly searched for an article in Auricle. Being a very Google-friendly name, a Google search for "Auricle Bath" easily found links to the blog. However, the URL Google displayed for the Auricle blog at Bath led only to a 404 (Page not found) error message on Bath's web server.

It seemed highly regrettable that potentially valuable historical content giving views on the potential of the Web (including technologies such as blogs and podcasts) to enhance the quality of students' learning experiences was now no longer available. The Institution might reasonably have legitimate concerns about this loss of its intellectual endeavours demonstrating its own early endeavours in blogging and e-learning.

Why did the blog disappear?

The URL for Auricle blog (www.bath.ac.uk/dacs/cdntl/pMachine/morriblog.php) provides some clues. DACS is the Division of Access and Continuing Studies, and CDNTL is the Centre for the Development of New Technologies in Learning - but neither of these departments still exists. 'pMachine' is the blogging software, and morriblog clearly refers to Derek Morrison, who left the Institution a number of years ago to support the HE Academy's Pathfinder programme.

Following staff departures and organisational changes, support for learning at the Institution is now provided by the Learning and Teaching Enhancement Office (LTEO) with the e-Learning Team having responsibility for managing and supporting e-learning developments. As a result, it seemed that the Auricle blog got lost somewhere along the way.

Could any of the resources be retrieved?

Since the blog was public, the contents of the blog have been indexed by Google. Using a combination of search terms, such as "Auricle Bath", it is also possible to discover Web resources which cite the Auricle blog, for example a contemporary post on Stephen Downes's blog (Downes, 2004) citing Derek Morrison's views of the potential of blogs as "the basis for a distributed, not centralised, information and learning object system":

It was also discovered, through further Googling, that the Auricle podcast resources were still available, on the University of Bath Website. An RSS file also contains the publication dates, confirming that the podcasts were published in 2005.

Better still, further Googling at length revealed that the Auricle blog is alive and well. Its new address www.auricle.org/ (an improvement on the original URL; but see also Scenario on Vanishing Domain Names). The blog now uses Wordpress, and posts from the original pMachine implementation at Bath have been imported.

This doesn't mean the blog has been preserved. It might be more accurate to observe that it continues to be used and be useful. Nevertheless anyone who has bookmarked it

at its original address is going to have to be persistent if they want to find it.

The Lessons

What are the implications of this case study for the wider community? And what lessons can be learnt?

Web managers should be aware of the dangers of associating services too closely with departmental names and specific technologies. "It is the the duty of a Webmaster to allocate URLs which you will be able to stand by in 2 years, in 20 years, in 200 years. This needs thought, and organization, and commitment." (Berners-Lee, 1998).

Departments need to audit their networked services, and document their policies regarding the sustainability of such services. Such documented policies should be examined when departments change their names, or there are significant changes in personnel.

This case provides an interesting example of a service which has been driven by an individual, who feels personal ownership for the content. It was this personal and professional attachment that led to the blog being continued in a new location, and former content being migrated. But who owns the blog? And what would have happened if there had been a dispute over ownership of the blog content, or even its distinctive name? These are questions which will be relevant to many academics who make use of blogs to support their professional activities.

The Institution had originally provided the technical platform for a blog that was intended to offer some value to the wider sector. But without one or more champions to sustain the momentum and assume ownership once the original team left the Institution, Auricle (in its original form) became the victim of annual online account housekeeping.

Now that the issue has been identified, what should the Institution do about the dead hyperlinks? A page pointing to Auricle.org could be inserted; It would even be possible to effect forwarding from every old addresses in the Institution domain, to new addresses in auricle.org. A 'Page not found' error seems the least desirable option.

Another interesting issue is the value perceived by the different actors. The owner valued Auricle for its ability to help organise and rehearse material and arguments for public consideration, and others working and studying in the field valued it as a resource. But, in common with other HEIs, once personnel leave an institution, electronic resources associated with them - emails, server and web space - are indiscriminately deleted. They rarely seem to be analysed for the potential value of their content. This is unfortunate, because it represents a lost opportunity for ongoing collaboration, discourages future investment by blog or wiki authors in such institutionally-hosted resources, and forces a move to outside the institution (so reducing the chances of generating assets of future value to the institution). Nevertheless it should be recognised that supporting such resources represents an ongoing cost for an institution (albeit a relatively tiny one in Institution financial terms).

On the other hand, should a decision about value be left to an individual? It can be argued that information of significant value to the Institution should be deposited in shared or managed spaces, such as document management systems, records management systems, institutional repositories, etc. Are archivists and records managers aware that this material is being generated? Are their information and retention policies governing these systems sufficiently advanced to accommodate new breeds of online resource like blogs and wikis?

Case study: Capturing wiki contents

Scenario: *Wikis have come a long way since the first WikiWikiWeb, and now are at the online heart of innumerable projects - for teaching, research, publishing and business. Some wiki systems go considerably beyond the original concept of a simple editable hypertext and incorporate a raft of other Web 2.0 features, like blogs, comments, newsfeeds, discussion and messaging, to create a comprehensive collaborative environment (e.g. Wetpaint, Confluence).*

A Wetpaint wiki is free and easy to set up: JISC-PoWR used one as a collaborative space to record project workshop feedback. Once all the input was collated, the wiki was no longer any use to us. But what if we needed to capture the contents? There are many good reasons to do this: to migrate to another wiki system or CMS, as the shape and nature of the content evolves; or put it on a permanent, persistent footing by moving it into our own domain; or to back it up or take a snapshot; or pull out information for publication in a different form. With one or two pages, it might have seemed trivial; but what if you now have hundreds (to say nothing of comments, blog posts, revision histories, etc.)?

Issues

Unfortunately, just as exporting the information is often a secondary consideration for wiki content creators, so it also is for some wiki systems. Browsing the Wetpaint Wiki Support Discussions indicates that an export feature was a long time in coming (and its absence quite a blocker to adoption by a number of serious would-be users). And what was eventually provided does leave a lot to be desired. Wetpaint's 'backup option' lets you download your wiki content as a set of HTML files. Well, not really HTML files: text files with some embedded HTML-like markup. (Which version? Not declared.) Don't expect to open these files locally in your browser and carry on surfing your wiki hypertext (even links between wiki pages need fixing). The export doesn't include comment threads or old versions. Restoring it to the original online wiki (or a new one) is not possible. On the plus side, though, you have at least salvaged some sort of raw content, that might be transformed into something else with a bit of scripting.

Another impressively-specced, free third-party wiki is Wikidot. This has a backup option that creates a zip file containing each wiki page as a separate text file, containing wiki markup as entered, as well as all uploaded file attachments. However, according to Wikidot support "you can not restore from it automatically, it does not include all page revisions, only current (latest), it does not include forum discussion or page comments". To reconstruct your wiki locally, you'll, again, need some scripting, including using the Wikidot code libraries to reconvert its custom wiki markup into standard HTML.

A third approach is possible with a self-hosted copy of Mediawiki. Here you can select pages, from one to all, and have them exported as an XML file, which also contains revisions and assorted other metadata. Within the XML framework, the page text is stored as original custom wiki markup, raising the same conversion issues as with Wikidot. However, the XML file can be imported fairly easily into a different or blank instance of Mediawiki, recreating both hypertext and functionality more or less instantly.

There are also some script-based tools available that can convert some wikis (particularly) to HTML pages. Confluence has a UWC extension (Universal Wiki Converter) - but it only works to translate from popular wikis like Twiki, Mediawiki, Sharepoint to Confluence. The documentation makes it clear that further manual work is likely to be necessary after conversion.

In contrast to all these approaches, if you set a spidering engine like HTTrack or Wget to work remotely harvesting the site, you would get a working local copy of your wiki looking pretty much as it does on the web. This might be an attractive option if you simply want to preserve a record of what you created, a snapshot of how it looked on a certain date. It would also be unnecessary to keep running the wiki software (assuming you are hosting it). However, this will only result in something like a preservation copy - not a backup that can be easily restored to the wiki, and further edited - in the event, say, the wiki is hacked/cracked, or otherwise disfigured. For that kind of security, it may be enough to depend on regular backups of the underlying database, files and scripts: but you still ought to reassure yourself exactly what backup regime your host is operating, and whether they can restore them in a timely fashion. (Notwithstanding the versioning features of most wikis, using them to roll back a raft of abusive changes across a whole site is not usually a quick, easy or particularly enjoyable task.)

Approaches

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

- How long do we need it for?
- Will it need preserving at intervals, or at a completion date?
- Is it more important to preserve its text content, or its complete 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?

Questions like these are no less relevant when considering your uses of blogs and other social software.

Scenario: Preserving presentations on Slideshare

***Description**: Slideshare is a popular third-party service for providing access to copies of presentations. Presentations uploaded to Slideshare can be viewed there, commented-upon, and used to find related materials. They can also be embedded in other web pages. There is evidence to demonstrate Slideshare has considerable impact in maximising awareness of presentations of all kinds - materials for teaching, research, business and marketing - as well as allowing presentations to be reused, delivered online, and referred to after the main event. Discussions and social networks can grow around one or more presentations.*

Issues

Slideshare is not a Trusted Digital Repository, as defined by the OAIS. Slideshare's predominantly a broadcasting/dissemination tool - it's clearly not any kind of system for managing institutional records or digital assets, or long-term preservation and storage.

Are there risks associated in making use of a third party service in this way? What will happen if Slideshare ceases to operate? Might this risk be addressed if Slideshare's functionality could be provided in house? In the case of Slideshare an in-house solution would not only be costly to replicate its functionality, but it would also be unlikely to provide the global impact and popularity which Slideshare has.

Approaches

One challenge is to assess possible risks and to explore mechanisms for managing such risks. An approach is to look at the popularity of the service and its user community (an approach, incidentally, which has also been recommended when selecting open source software), which may indicate something about its potential longevity. The Techcrunch service can be useful if providing information on the financial background to many Web 2.0 companies and its information on Slideshare seems reassuring, with a post in May 2008 described how Slideshare had secured $3M for Embeddable Presentations.

- Store a managed master copy of the slides on institutional systems (CMS, DMS, IR) and ensure that links to this resource are provided on Slideshare. If institutional policy permits, the Institutional Repository could be a suitable location; if not, a parallel repository system could be established, much as Southampton University set up an Eprints system dedicated to managing materials for its Open Repositories 2008 conference.
- Provide a Creative Commons licence for the resource, which seeks to avoid any legal barriers to future curation of the resource and allow the resource to be downloaded from the Slideshare site.

This approach aims to ensure that the master resource is kept at a stable managed location, allows users to make a copy of the resource (if, for example, the Slideshare service suffers from performance or reliability problems) and allows uses to bookmark or cite the managed master version of the file.

Scenario: Institutional Use of Twitter

Description: *Twitter is a 'microblogging' service which can be used to create a brief (up to 140 characters) blog post and broadcast it in several different formats. The fact that Twitter updates, called 'tweets', can be sent and received as mobile phone text messages, as well as via a web interface, considerably expands its potential uses - although this has been undermined by the recent withdrawal of the free SMS feature for users in the UK.*

The Institution might choose to set up an Institutional Twitter account, which it uses to disseminate news on institutional 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 in the Institution have quickly become early adopters of the technology, and are 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.

Issues

There may be a need for institutions to consider the preservation and management implications of their tweets, even if it would be inappropriate to have heavyweight policies on personal use of micro blogging or instant-messaging technologies.

Approaches

Technically, capturing Twitter tweets is straightforward, since 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 (there is a Wordpress theme available, 'Prologue', which demonstrates how one might create a virtual clone of the Twitter interface.

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 record manager?
- Scholarly record: can it be demonstrated that tweets are part of the scholarly record that isn't captured in other forms?
- Legal reasons: Is Twitter being used to deliver learning? Is there a legal requirement to record what has been sent to whom, 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 Universities 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 mediums may take over the role of others.
- As a recognised official Institution publication, the posts need to be subject to QA and editorial processes, which includes 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.

Case study: Preservation and Instant Messaging

In this brief case study we describe the use of instant messaging to support communications between two institutions. The case study will attempt to draw out some of the general policy issues which should be applicable more widely.

Use of IM for the QA Focus Project

This example describes the approaches taken to the use of instant messaging to support communications between the project partners for the JISC-funded QA Focus project which was launched in January 2002. The project partners were UKOLN (based at the University of Bath) and, initially, ILRT, University of Bristol. However after the end of the first year of the project ILRT withdrew form the project and were replaced by AHDS, who were based in London.

In order to minimise the amount of travel and to help to provide closely integrated working across the project partners it was agreed to make use of instant messaging technologies. As well as enabling the team members to have speedy contact with each other it was also recognised that official project meetings could be held using the technology. It was appreciated that in this context there was a need to have a slightly formal protocol for managing the meetings, to compensate for the limitations of online meetings. And in addition to the best practices for managing the online meetings it was also agreed that a record of the transcript would be kept, and that this record would be copied across to the Intranet along with other formal documents.

After AHDS replaced ILRT as project partners we decided to change our IM client from Yahoo Messenger to MSN Messenger. It was either during this change of IM tools or whilst making use of another IM client that we noticed that different IM applications work in slightly different ways. This includes whether a transcript of dialogue is kept automatically and whether new participants to a group chat will see only new discussions or discussions which have taken place previously (which has the potential to cause embarrassments at the least).

The experiences we gained in use of IM led the project partners to develop a policy on use of IM (which covered issues such as the possible dangers of interruptions, as well as keeping records of formal meetings held on IM). The policy also clarified use of IM in an informal context, with their being no guarantee that records would be kept.

The policy stated that:

- IM software may be used for formal scheduled meetings. In such cases standard conventions for running meetings should be used. For example an agenda should be produced, actions clearly defined, changes of topics flagged and a record of the meeting kept.
- IM software may be used for direct communications between individual team members. For example it may be used for working on particular tasks, to clarify issues when working on collaborative tasks and to support team working. IM may be particularly suited for short-term tasks for which no archive is needed and other team members need not be involved - for example, arranging a meeting place.
- Highly confidential information will not be sent using IM, due to the lack of strong encryption.

The Web 2.0 environment has a strong emphasis on communications between individuals and not just one-way publishing. This pattern of usage places additional challenges for institutions wishing to ensure that records are kept of the dialogue which takes place. These challenges may well need to be addressed within the context of policies on the preservation of Web resources as increasingly digital communications technologies will have Web interfaces.

Issues

The general issues arising from this case study include:

- The need to ensure that the users of the IM technologies and those involved in developing policies related to its use have a good understanding of how the technologies work together with an understanding of the differences between different IM systems.
- The need for simple documented policy statements

Part II: INFLUENCING THE INSTITUTION

Chapter 14: What are the drivers for web archiving?

There are many drivers for undertaking website and web resource preservation within a Higher Educational institution: institutional policy, legal requirements, and research interests are just a few. Web preservation should be part of the Institution's operational work, and not a project in and of itself. Rather, the project is to embed Web preservation into operational work. Below is a summary of some of the internal and external forces that can act as drivers to doing this.

We need to consider preservation needs at all levels of the institution, wherever the web is used. However we can initially draw useful conclusions even by just considering the issues at perhaps the highest institutional level. The website may also contain digital assets and electronic resources, which is to say assets which may be of continued value and are held in digital form. They may increase in value through sharing and repurposing.

1. To protect your institution

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 are taken. This loss could be construed as a threat to what might be called the business continuity of the organisation. If you do not record or protect certain information you are in danger of failing to comply with legal acts such as FOI and DPA, you may be breaking contractual and auditing obligations, and putting your institution at risk. This risk management approach has been taken with other digital resources in an institutional context (for example, management, selection and preservation of emails), and it is only a matter of time before it is a standard approach to websites.

The Institution is an organisation with business continuity and interests that need to be protected. It is an employer with statutory obligations. Reference to archived copies of institutional websites may be required for the checking of strategic, legal, financial, contractual or scholarly information. An Institution needs to be sure that its resources are trustworthy and reliable. This is all part of the evidentiary value of web resources.

The Institution's reputation can be put at risk by poor website management and a bad publication programme. Users can be frustrated by poor web continuity, dead links, and missing resources.

2. To be forward-looking

Starting a Web preservation programme will make you look like a ‘forward thinking’ Institution. You could be one of the first to start an official ‘Web preservation’ programme which will be great marketing fodder. (Remember the first UK Universities to offer blogs to students (Warwick), launch a YouTube channel and offer downloadable lectures using iTunes (University College London)? How about the first to get sued by a student for changing the course specification and having no record of the previous entry? Universities have already been sued over website accessibility, copyright of material on their site and allowing plagiarism to take place.) Embedding Web preservation strategies will also help you think about the continuity of resources, dead links etc.

3. It could save you money

Web resources cost money to create, and to store; failing to repurpose and reuse them will be 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.

4. Responsibility to staff and students

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

5. Responsibility to users

You have a responsibility to the people who may need to use your resources in the future. Many of the resources which your institution publishes are unique, and deleting them may mean that invaluable scholarly, cultural and scientific resources (heritage records) will be unavailable to future generations.

Chapter 15: Some personal perspectives on web preservation

The following statements are intended to typify many common web issues (ownership, responsibility, reuse, publication, records, value) as they are perceived by the many stakeholders a web resource is likely to have. They are based on comments and observations made at the JISC-PoWR Workshops and on the project blog.

The User (Staff, Student)

"I know of some resources on the web, associated with my institution, that I think should be considered for preservation. They don't seem to be in scope of any institutional system for managing records or publications, and they may be in danger of being lost if the website they are part of is redesigned. Who do I tell, and what can they do?"

The User (member of the public)

"There was a really useful project page on that Institution website just a few weeks ago, and I was using it to help my research. I even put in a link to it on my blog. Now it's gone 404. I can't even find it using the search engine. Where has it gone?"

The Web Manager

"I get the feeling that I am expected to preserve some things about our websites, but I don't know what. If I knew what needs keeping, and why, I might be able to work out how."

The Records Manager

"I have the uneasy feeling that there are Institutional records being stored on our website and perhaps I need to do something about it. But this sounds technical and I'm a paper person. I have enough trouble trying to preserve hard copy records without having to worry about the web. I can see the value in theory, but in practice it's too huge."

The Library Manager

"I know about the books and periodicals in my digital library, but is anyone collecting a series of digital publications from the website, like the prospectus? Or the research materials published online by Departments?"

The Registrar

"Ever since we installed that online student registration software, I've never understood where we keep the records of students who have registered. I'm sure they must be somewhere in the system."

The Information Manager

"We have many systems for storing and managing different sorts of information: Email, Website, Intranet, CMS, VLE, Document Management System, Institutional

Repository, shared drives. Suddenly students and staff are using all sorts of things - blogs, wikis, Facebook, Google, Wetpaint, Ning, Twitter, Flickr, Slideshare, Second Life. How do we deal with that - let alone preserve any of it."

The Marketing Team

"We're relaunching the website tomorrow. I expect the web team will keep a copy of the old one on a CD somewhere in case we ever need it."

The Web Team

"No one's asked me to keep a copy of the old website when the new one goes live tomorrow. I suppose if anyone needs to see the old one again, there's always The Internet Archive."

Chapter 16: Responsibility for preservation of web resources

Ownership of web preservation

Effective web preservation needs to be policy-driven. It is about changing behaviour, and consistently working to policies.

There are many aspects of ownership. The Institution ought to take ownership at the highest level with a clear policy that states the importance and value of its web resources, and makes it clear why some of them are being preserved. There ought to be a sense of corporate ownership of the Institution's website, the web-publication programme, web resources that have value - and therefore of the need to preserve them. At senior level, there should be an in interest in operational efficiency and compliance.

As part of that institutional ownership, the Institution may be motivated by legal and records-management reasons to protect and preserve web resources. Records managers and other information professionals will own the problems associated with those web resources that are classed as records. Their interest will be in legal compliance, long-term RM goals, retention, disposal, and classification.

Individuals, authors, academic staff, administrative staff and even students are also stakeholders with varying degrees of responsibility for the creation, management and storage of web resources. Some will have an interest in visibility, accountability, compliance and control. Authors, for example, may wish to retain a copy of papers, articles and other written works for future use in CVs, assessment exercises, etc.

Implicit in the above is ownership of the preservation problem; and other issues associated with making the resources preservable in the first place, for example, capture, storage and management. This implies ownership and responsibility at all levels.

Unfortunately, those involved in providing institutional Web services may not always be interested in issues related to preservation. This presents a potential risk for the Institution. If, for example, Web Managers are not interested, it may be difficult to persuade them of the need to invest resources in preservation, and to gain the necessary commitment from senior managers and policy makers.

Resourcing web preservation

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 another indicator that technology alone does not have the answer. Any programme of work associated with web archiving needs to be properly resourced, with team-based and collaborative approaches drawn from across more than one discipline.

Policy-making and resources for implementation should originate outside the IT department. IT departments are there to effect policy, not make it: the implementation responsibility lies with IT, but IT should not own the web archiving programme or project. Web archiving is not primarily a technological problem: the solution does not lie in buying new software or more software. There is no single technological 'solution' that will fix everything, nor is there any single tool that addresses all possible web preservation

issues (behaviour, dynamic content, scripts, versioning, etc.).

Sponsorship for web preservation

The espida project in Glasgow (www.gla.ac.uk/espida/) offers a useful methodology which could be used to quantify the value of web archiving. It takes a pragmatic view of the way that HFE Institutions operate in the real world. espida understands that Universities aren't geared towards preservation, and that preservation activities will continue to vie with other services for funds. Quite often any digital preservation projects that do take place are given short-term funding, which is at variance with the nature of the problem. *"Now that for the most part the technological solutions have been, or can be, solved, the focus has to be on creating an environment where digital longevity is an organisational goal."* (espida)

espida can help you:

- Demonstrate the value of websites and web resources
- Communicate the intangible benefits of web archiving and web resource preservation to your potential sponsor, and articulate those benefits
- Make a case for a web archiving and preservation programme, based on a formalised and transparent communication process between the proposer and the funder
- Identify costs and benefits of web archiving and preservation, using scorecards and cost templates
- Produce a decision-making process that is transparent and based on all relevant information

At the end of the process you will be empowered to present a business case which not only answers the question "how much does web archiving cost?", but also "why do we need web archiving?" and "why should we spend money on web archiving, rather than on the primary business of the organisation?"

To quote espida:

> *"Strategic thinking is not driven by cost and financial issues alone. It is driven by vision and insight with organisations taking risks when investing in new ideas in order to develop. The espida project is seeking to ensure that where required, organisations recognise the value of their web resources and have the foresight to see that their persistence should be a matter of decision rather than technological determination. This requires an explicit recognition of the value of web resources...The challenge is in expressing value in terms that senior managers understand. If web resources can be shown to bring value (which is multifaceted) in strategic terms, then there is a greater chance of receiving resources for their retention, so as to capitalise on that value." (espida Model Handbook)*

A PoWR programme

Success in the preservation of web resources will potentially involve the participation and co-operation of a wide range of experts: information managers, asset managers, webmasters, IT specialists, system administrators, records managers, and archivists. Through its workshops and activities, the JISC-PoWR project has endeavoured to bring together a variety of institutional stakeholders who might not otherwise encounter each other, such as records managers and web managers. Collaboration is the key.

A collaborative approach is not the only path to success. The intention here is not to recommend the formation of a 'virtual committee', whose views have to be taken into

account on every single issue and every single time a change is proposed, which might tend to obstruct progress. Rather, it is a question of accessing the expertise as needed, in order to initiate change.

> *"We can be confident that the archive is the responsibility of the archivist; the Web site the responsibility of the Web manager. However, Web resources which should be in the archive, and under archivists' control, are not. This creates a significant additional burden for Web managers with an ever-expanding Web presence to manage, and a dilemma for the archivists in that a significant proportion of the archive is no longer under their control. For the organisation, and especially the records manager, this introduces a variety of risks, such as the persistence of personal data without good reason. Once published, a Web resource will be retained- i.e. kept in current use - until it is either deleted or archived. The default position of retention is not tenable. Existing policies - such as an organisation's retention policy - need to be translated or adapted to guide the management of these processes. Moreover, this requires a virtual team of practitioners – especially the archivist, records manager, IT manager, and Web manager – to develop and implement them." (Emmott, 2008)*

Chapter 17: Institutional strategy

Shaping of policy

Web preservation is a big topic and we're not even pretending to deal with all of it. The aspect that we care about, and that JISC believes the community is looking for help with, is fairly well-defined. We want to help institutions make effective decisions about preserving web resources, and help them implement those decisions in a way that is cost-effective and non-disruptive.

What do you want to achieve? In this chapter, we suggest areas and resources you can target to help you define and decide what it is you want to get out of your web preservation programme. This includes adapting methods which can help you identify and measure the value of these resources.

Options

As you gather more information and arrive at a deeper understanding about your web resource collections, a number of options will become open to the Institution. For example:

- Business as usual - nothing needs to change
- Policy review is needed (see below)
- Quick wins - actions that can be performed now to get results, or to rescue and protect resources that you have identified as being most at risk
- A finite, selective web preservation solution - targeted at one department or many departments; or at one particular collection, or type of resource (and see Chapter 5: Selection)
- Strategic approach - a comprehensive web archiving programme over many years, affecting the entire Institution

Further suggestions can be found in Chapter 9: What approaches and techniques can you use?.

Policy Review

Reviewing policies and procedures is vital. 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.

Cycle of policy review

Policy review can also be embedded as a continual-review action within the PoWR process itself. Because we are promoting an Information Lifecycle approach, and a selective approach, there are clearly-defined stages in the PoWR approach when decisions are being made. For example:

- The scope of what you will include in the collection
- What you will exclude from the collection
- Suitable approaches to preservation
- Decisions about the identification of records, publications and artefacts
- Who wants resources kept, why, and for how long?

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

Managing Decisions

Making effective decisions

At its simplest level, this means deciding what to keep and what not to keep. There may be many drivers for these decisions - institutional policy, legal requirements and research interests are just a few. The decisions need to relate not just to what is to be kept, but *why* and *who for*. That's because those requirements may have a bearing on how you choose to go about the job, or whose responsibility it is to carry it out. Not everything needs to be kept, and even when it does, it may not be your institution's responsibility to keep it.

Implementing those decisions

Carrying out your decisions - keeping things, throwing things away, or ensuring that other people keep things - can be the trickiest part of the process. You may know you want to preserve the prospectus for past years, but can you be sure that your CMS, or the Internet Archive, or some local use of web harvesting tools is going to do this job effectively for you? You may be being told that some part of your web infrastructure would be easier to preserve if you avoided the use of certain features, or used a different authoring system. Is that true, and if it is, what are the negative consequences of such decisions?

This aspect of strategic planning and thinking will involve:

- Making decisions that are consistent with policy
- Making decisions that are consistent with regulation
- Making decisions quickly
- Making decisions cheaply
- Making decisions that are reusable, long-lived and implementable
- Tying in decisions with high-level responsibility and individual ownership of resources
- Decisions about behaviour of individuals, when creating, using and storing resources
- Information lifecycle-type decisions, about when things are supposed to happen in the cycle

Chapter 18: What policies exist?

Finding policies and procedures

It is unlikely that any Institution will have a single stand alone policy or mission statement that governs everything we would like to see happening with regards 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.

That said, we suggest the following sources are investigated and studied as they may prove helpful. Your Institution may not have policies or guidelines for all of these.

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

You may also want to locate the Minutes of any Committees or Advisory Groups in your Institution who formulate web development strategies for the Institution, or advises on policy and current development activities.

Assessing your policies

Once you have located all relevant policies, you should ask the following questions:

- Do any policies refer explicitly to web resources?
- Do the policies refer to our three proposed web resource classes (Publications, Records, Artefacts)?
- 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 you to carry out preservation actions?

- Would these policies prevent you from carrying out preservation actions?

Interpretation of policies and procedures

At the first JISC-PoWR Workshop, we learned that none of the Institutions attending had web material included in their retention schedules. Nor were they aware of having a web preservation strategy.

This may not matter. A records manager's retention schedules, for example, may not explicitly mention web resources by name. Retention scheduling is just one approach to digital asset and web resource management; and records managers tend to identify the content of the record, rather than describe the form it is in.

Such policies are typically generic – for instance, talking in terms of 'information' – and therefore need to be translated, or adapted, to address the preservation of Web resources. This entails all Web resources and must stand the test of time without the need for endless revision.

Therefore even without that explicit identification of web materials, it will still be possible for us to turn the RM policy into something that will enable us to deal with web resources.

Chapter 19: How can you effect change?

It became apparent at the JISC-PoWR workshops that there was concern amongst delegates about instigating or establishing web preservation activity within their Institution. Some comments suggested that it was going to require a lot of difficult preparatory work. Some fears included:

- We need to consult lots of other people from different backgrounds, some of whom I don't know very well, and we don't have a shared language
- We'll need to form a virtual committee or task force to get this done
- We need to take it to the very top of the Institution and they probably won't listen
- We need to write a persuasive case
- It's going to make me very unpopular
- We'll need to buy, implement, and learn new software
- It's going to cost a lot of money
- People will have to change what they're doing
- We need to gather evidence of who would be affected by data loss, and why
- We need to do risk analysis
- We need to do change management

The answer to some of these reservations is 'not necessarily'. In Chapter 9 we have already described ways in which a quick win might be achieved in the short-term, if a long-term strategy can be deferred. For example:

- You can get quick results by running a pilot project, especially if it's very selective
- Web harvesting software is open-source (i.e. free)
- You could set up a regular harvest of the Institution's website with little or no disruption
- Consulting a few people is an easy way to get results, and not the same as establishing a virtual committee

However, in order to look again at some of the issues of effecting change, we suggest, below, the espida model, change management, risk analysis, and the ways in which records managers and webmasters need to rethink some attitudes.

espida

The issue is one of clearly and effectively communicating value and benefit in terms that senior management can understand. The espida approach does not pit the proposer against the decision-maker, but encourages them to enter into a dialogue on investment opportunities. (See, for example, Chapter 3 of the espida guide, 'Building a good business case: the espida Approach'.) This means supplying answers to the following questions that management may ask:

- How much do you want?
- What do I get for it?
- How will I know that I've got it?
- How likely am I to get it?
- What determines success or failure?
- How will you manage it for success?

Change management

This is a large topic and beyond the scope of the PoWR Handbook. However, JISC have published change management and project management guidance as part of their infoKit series.

The Change Management infoKit (JISC 2008) was developed out of a HEFCE Good Management Practice Project led by the University of Luton entitled 'Effecting Change in Higher Education'. The project team consulted widely on aspects of change in the sector and put together theories, approaches and tools that resonated with them and with those they talked to about their experiences of the practical difficulties of managing change.

These approaches have the advantage of coming from HE environments, and as such they will be tailor-made for your Institution. The 'Effecting Change' team summarise their findings by the following observations:

- There are no easy solutions.
- Adapt processes to suit the change intended.
- Change requires teamwork and leadership (and the two are related).
- Work with the culture (even when you want to change it).
- Communicate, communicate, communicate.

Web preservation can support the Institutional Mission

What does preservation of web resources do to support the Institutional mission statement? This idea comes directly from the Neil Beagrie study on *Digital Preservation Policies and their implementation*, commissioned in 2008 by JISC to address the gap caused by the lack of discrete DP policies in the UK HFE sector. The study aims to provide an outline model for digital preservation policies and in particular to analyse the role that digital preservation can play in supporting and delivering key strategies for Higher Education Institutions in areas such as research and teaching and learning. (See www.jisc.ac.uk/fundingopportunities/funding_calls/2008/01/dppolicy.aspx)

Preservation of web resources can be seen as legitimate digital preservation activity. As such, you can strengthen the case for doing it 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 don't explicitly mention digital resources by name, for example 'Attracting a wide variety of students'.

Preservation of web resources can develop into a means to an end, rather than an end in itself; and harnessing web content in service of the Institution's declared aims can become a major business driver.

Risk management and risk analysis

Possible risks associated with websites and web resources are identified in various chapters of the Handbook, including for example:

- Data loss
- Loss of records
- Loss of resources
- Failure to be information compliant

- Risk of litigation from students or the public
- Risk of breaching copyright

Some of these were discussed at the third JISC-PoWR workshop, along with the suggestion that bringing web preservation into line with disaster planning may help to change institutional practice. There is also the view that some of these constitute low risks: is it likely that a student will sue an Institution because of a change to a website's content? There is the general feeling that Data Protection and FOI legislation have not quite caused the massive information management upheaval in Institutions that was originally anticipated.

The risk associated with possible IPR infringements are put in perspective by Charles Oppenheim's lightweight risk formula: R = A x B x C x D, where:

- A: Probability that you're illegal
- B: Probability that you're 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.

See also Risk Management in Appendix A: Legal Matters, Chapter 14: What are the drivers for web archiving?, Chapter 15: Some personal perspectives on Web Preservation, and Chapter 18; What policies exist? Risk analysis and risk management are large areas and beyond the scope of this Handbook. If you decide a risk management strategy is needed to enable web preservation, we suggest you start with the JISC infoKit on Risk Management. The kit takes the view that Risk Management is an essential part of project management.

Records management vs Web Management

Records managers:

- Need to recognise that the web is a potential place where records can occur. They need to identify where, how, when and by what agency this is happening.
- Need to rethink some of their traditional models. Records are now occurring in new ways. They are not static any more. "Centralised control over records is no longer possible, and insisting that records should be managed in the same way regardless of format is wishful thinking. New methods will be needed for retention and appraisal." (Bailey 2008)
- Need to overcome fear of IT, and forge relations with people like the webmaster, sys admin, 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 ways to better capture and manage the content.
- Consider preservation-friendly software for your next purchase of web tools.

JISC resources

The JISC have also published additional infoKits which may help with aspects of change

management:

- *Project Management*: www.jiscinfonet.ac.uk/infokits/project-management
- *Programme Management*: www.jiscinfonet.ac.uk/infokits/programme-management
- *System selection*: www.jiscinfonet.ac.uk/InfoKits/system-selection
- *Implementing an EDRM*: www.jiscinfonet.ac.uk/InfoKits/edrm
- *Risk Management*: www.jiscinfonet.ac.uk/InfoKits/risk-management.

There is also the JISC Guidance on *Functional Disposal Schedules* for records management in HEIs. Some of these functions might align with the way you are using and managing your web resources:

- www.jisc.ac.uk/whatwedo/themes/eadministration/recordsman_home/srl_structure/srl_intro.aspx

Chapter 20: Information Lifecycle Management: Creation

In this chapter we consider lifecycle management in more detail. The implementation of ILM can be considered as part of a strategic approach to managing web resources; and it can be considered as one possible way of effecting a change within your Institution. This guidance on web resource creation is adapted heavily from the JISC infoKit, which proposes four stages to the lifecycle: Creation, Active Use, Semi-Active Use, and Final Outcome. (See www.jiscinfonet.ac.uk/infokits/information-lifecycle).

According to the Digital Preservation Coalition, *"The major implications for life-cycle management of digital resources is the need actively to manage the resource at each stage of its lifecycle and to recognise the inter-dependencies between each stage and commence preservation activities as early as practicable."*

JISC-PoWR recommends close examination of the first stage, on creation. The creation stage raises many pertinent questions which apply to websites and web-based resources. At creation stage, you should address questions to ensure that 'the information created is fit for purpose and that it is actually capturing appropriate and reliable content'. This means getting involved with the functions of web resources; ensuring the right people are involved in the creation; reliability and trustworthiness of resources; formats; and the creation and management of metadata.

Creating the right resources

Web resources may be created for many purposes, including:

- Information through content
- Record of a process
- Publication and dissemination
- Teaching and learning

Ideally we want a consistent approach to resource creation across many operations, and all parts of the Institution. Departments should not be creating things in different ways. Everyone should work to agreed practices for web publication; if no such practices exist, define them. It is possible to create too many resources, just as it is possible to use a web-based method for doing something that could easily be done another way.

Creating reliable resources

Web resources should be trustworthy. Resources should be current and up-to-date, if that's part of their purpose. If the resources are published, consider the wrong decisions that can be made if unreliable resources are published to the web. Define responsibility for who should be updating the content, how frequently it happens, and when. Metadata helps with reliability and audit trails. Not just dates, but other metadata should be controlled, such as name of creator, or name of department. Fitness for purpose is important. Use pick-lists from databases wherever possible to ensure data quality for the resource. Version control can be managed by using features in the CMS.

The right people creating resources

This requires some understanding of the structure of the organisation and roles and responsibilities within it. Proof of provenance and authentication is important, especially for record-keeping and publication. We need to know where the resource comes from, and who created it.

Content Management Systems can create and organise content, yet they can also restrict the task of creation to a few authorised people, sometimes to the exclusion of those best placed or qualified to manage content. Conversely, open-source web authoring software like Wordpress is increasingly allowing more users to create and manipulate information quickly.

Social software (blogs and wikis) means unfettered access; many people, staff and students alike, can contribute content to the resource. But users of the content need to be aware of the mixed origins of the content, otherwise we have another 'reliability' issue.

Resources created in the right formats

Sometimes the format of the resource can be overlooked in favour of a concentration on content, or delivery of the resource. Issues concerning reuse and expected longevity may get overlooked. In fact, web-based applications may not always be the best solution for the resource.

Web resources are very easy, cheap and quick to publish. Sometimes the decision is made to opt for online publication only. But 'what if decisions are being made against the content of that publication?' What if 'it suddenly becomes necessary to know exactly what [the publication] said at a particular point in time some months or years ago?' Does your CMS have 'sufficient capability to track changes and roll back to how it appeared at a certain date?' All these questions raise further questions about authenticity and reliability. Failure to address them may leave the Institution at risk and liable.

Social software services are often externally-hosted. Content is created and stored there. This applies to certain wikis, blogs, online photographic storage services, and Second Life. If learning materials are being created and stored this way, you need to ensure you can continue to access them and preserve them. There may be a risk of the company which provides the service going bankrupt, or withdrawing the service. There may also be intellectual rights issues regarding content hosted by Second Life, or any third-party provider.

Format choices need to take into account longevity, protection, and preservation. If resources are not needed beyond five years, then questions about formats need not be a problem.

Metadata

Consider metadata requirements, especially when building a new web resource or website. Metadata is required not just for location and retrieval of content, but for many other purposes. Metadata can tell us about the audit trail of the resource, its intended use and purpose, its technical application, its retention or preservation requirements, for example.

According to MANDATE, this sort of metadata is fundamental:

- Administrative (date of creation, which department)
- Legal (copyright, digital rights, retention requirements)
- Preservation (metadata on format, software)
- Technical (formats, size)
- Structural

Other useful metadata standards are:

- DublinCore: http://dublincore.org/
- METS: www.loc.gov/standards/mets/
- PREMIS (Preservation Metadata): www.loc.gov/standards/premis/

Enacting metadata and its automation

- Be selective. Not all the proposed metadata listed above is needed for every single resource.
- Use automated metadata extraction where possible.
- Use picklists and keywords from a master source.
- Work towards consistency of date formats.
- If creators are entering metadata, enable ways for it to be as consistent as possible.

See also:

Gail M Hodge, 'Best Practice for Digital Archiving: An Information Life Cycle Approach (*D-Lib magazine* Vol 6 No 1, January 2000) www.dlib.org/dlib/january00/01hodge.html

Jane Greenberg et al, 'Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization' (*Journal of Digital Information*, Vol 2, No 2 2002) http://journals.tdl.org/jodi/article/view/jodi-39/45

Chapter 21: Can other people do it for you?

Internet Archive

The **Internet Archive** (www.archive.org) was founded in 1996. It is also called the 'Wayback Machine'. Bewster Kahle is the American director of the company and it is based in San Francisco. Although not exactly of Trusted Digital Repository status, and perhaps open to the accusation that it does not support many international standards (e.g. OAIS), the organisation is unique in that it has been gathering pages from websites since 1996. As such, it holds a lot of web material that cannot be retrieved or found anywhere else, and would otherwise have been completely lost.

The Internet Archive has always offered ways for anyone to submit a website to be included in the Archive. The simplest method is to register on the site, and submit a URL for inclusion via the 'Archive That!' service. The most recent development is the Web archiving service called Archive-It (www.archive-it.org/), which is a subscription service. The advantage of Archive-It is that you can create distinct Web archives called 'collections', containing only the content you are interested in harvesting, at whatever frequency suits your needs. The collections created with Archive-It can be catalogued and managed directly by the subscriber. The assumption is that you will make your archived copies 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's 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, you are expected to share your resource publicly and make use of Creative Commons to protect your resource.

A few caveats about the suitability of the Internet Archive solution to HFE Institutions in the UK:

- To date, IA lacks any sort of 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 lack of policy may cause severe problems to HFE Institutions; it is unlikely that it will cover everything your Institution needs to do within its remit.
- There is potential for legal difficulties and litigation. IPR issues may not be adequately dealt with by the Creative Commons and the IA's 'notice and take down' approach.
- IA may not have a sustainable funding model. Financial supporters come and go, and its continuance is largely dependent on the generosity of Brewster Kahle.

There are additional caveats about the technical failings of the Wayback Machine:

- IA won't capture all your web-based assets
- They can't guarantee capture to a reliable depth, or reliable quality
- They cannot capture any site or service that depends on a database, or a login
- Dynamic content can't be captured reliably
- Their cyclical gathering method leads to gaps in temporal continuity; there can be large gaps between capture dates
- There may be broken links or missing pages in the archived pages (no quality-assurance is undertaken, unlike with UKWAC who do a lot of curation)
- There may be missing images in the archived pages

- The image assets in IA are always smaller than archive quality copies
- IA may not be preserving the resources they capture (or at least not to OAIS standards)
- There is little in the way of contextual information in their catalogues

If all this is true a number of Institutional assets are missed out by the IA approach. For example library catalogues, image collections, e-prints collections with a database, and interactive teaching materials.

UKWAC

The **UK Web Archiving Consortium** (UKWAC) has been gathering and curating websites since 2004. Among its members are the National Libraries, The National Archives, The Wellcome Trust, and JISC. To date, UKWAC's approach has been very selective, and determined by written selection policies which are in some ways quite narrow. JISC, for example, have made it their remit to collect websites of HFE projects which they funded or helped to fund. That remit has expanded slightly to include the websites of certain central and regional HFE organisations, but to date no HFE Institutional websites in the UK have yet been collected. (The National Library of Wales are taking snapshots of Welsh Institutional sites.)

It is possible to nominate your Institutional website for capture with UKWAC. Bear in mind the following features:

- 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 passworded pages, hidden links)
- Similarly, if your website depends heavily on server-side architecture, then remote capture may fail

If you undertake the nomination and your website is selected by UKWAC, it will involve a few practical things:

- Signing a permissions agreement that states you agree to remote harvesting and copying
- Agreeing to having the archived copy made publicly available
- Allowing the remote harvester to ignore your robot exclusions

UKWAC, whilst demonstrating the economies of scale that can be achieved in web archiving, preserve only what their curators select. An UKWAC solution is better than nothing but there are limitations, and it may not constitute a quality solution to preservation of all your web resources.

The International Internet Preservation Consortium (IIPC)

The IIPC won't help you harvest your 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.

The goals of the consortium are:

- 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

See www.netpreserve.org/about/index.php.

Other external initiatives of interest

These initiatives, mostly library-based and mostly sponsored at a National level, are aiming 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 you with the archiving of your website. Some of the National Library collections are not open to the public.

MINERVA, Library of Congress (www.loc.gov/minerva/). 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/index.html). National Library of Australia with nine other Australian libraries and cultural collecting organizations.

UK Government Web Archive (www.nationalarchives.gov.uk/preservation/archivedwebsites.htm). A selective collection of UK Government websites, archived at regular intervals from August 2003, and developed by The National Archives using the services of the Internet Archive.

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

DAMP (Digital Archive for Web Publications), 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 (www.webarchiv.cz/index-e.html)

APPENDIX A: Legal Matters

Legal issues

Preservation of web resources places the Institution in a similar position to a publisher. Additionally, preservation activities always require **copying** of the resource. 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.

Legal issues that can arise when preserving web resources include:

- Freedom of Information (FOI) legislation, which entitles the public to request recorded information from public authorities, including universities
- Data Protection Act (DPA) rules governing the use of personal information
- Intellectual Property Rights (IPRs), particularly copyright
- Criminal and civil laws that relate to the content of the resource, such as defamation, obscenity, or incitement to racial hatred
- Contractual obligations such as Terms of Service (ToS?) for third party websites, particularly in the Web 2.0 space (such as Facebook or Slideshare)

Managing legal issues

Naturally the above list does not exhaust all of the potential legal issues, and each preservation project will have different risks and legal obligations. When examining the potential legal issues on a particular project, it might be useful to break down the issues into the following:

1. **Preservation of a resource because of a legal requirement.** Such requirements could be taking place in a records management context, in order for the Institution to comply with FOI legislation.

The 'legal requirement' area could be further divided into hard requirements: laws that explicitly state a resource must be retained or preserved, and soft requirements: self-imposed rules to avoid exposure to some legal risk. One example of a soft requirement might be keeping a copy of a website's terms and conditions as they evolve, in order to prove what terms governed at each exact time.

2. **Legal requirements not to preserve a 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."

3. **Preservation of content for a non-legal reason but for which legal issues must be addressed.** This could include any number of reasons, such as for cultural heritage.

Risk Management

The notion of risk management rather than absolute risk avoidance does act as an overall umbrella to these three areas. Clearly rules that firmly require information to be retained or not must be complied with. Concentrating on the possibility of legal liability too much for every area in-between does run another kind of risk - losing the resource.

- Think risk management: not total risk avoidance
- Don't get so caught up on legal issues that you totally miss the chance to preserve the web resources!

- If you don't preserve, then the resources may disappear before all the issues are resolved.

Assess the risk, and if it is low, then look into going ahead and doing it.

Identify the risks - What are the risks for your activities?

- Risk Avoidance - Pick what you preserve.
- Risk Reduction - Find ways to reduce the liability on what you preserve.
- Risk Acceptance - Accept that the risk is low and go ahead.
- Risk Transfer - Insurance for preservation activities.

Freedom of Information

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

The Freedom of Information Act requires Universities to adopt, maintain, publish and review from time to time a **publication scheme**. A publication scheme is a list of the classes of information that the Institution will make available on a routine basis. You may need some awareness of the Scheme, and whether it affects web resources you wish to preserve.

A website is one of the ways that the publication scheme (and some of the information made available through it) can be accessed. Users may be able to browse and search the scheme through links. In some cases the scheme may point to information that is available on the Institution's website, and where this is the case it may link to the appropriate web page.

Data Protection

The Data Protection Act 1998 gives rights to people about whom the Institution holds information and gives the Institution responsibilities regarding that information.

- The Data Protection Act covers almost all material which is associated with an identifiable living person
- The Institution must register its holdings and its use of them
- It must protect information
- It must not hold it for longer than is required
- You must give access to the subject
- You must allow errors to be corrected

In terms of how this impacts on web preservation activities, the Data Protection Act may prevent you 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

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

Technical Protection Measures (TPMs)

Technical Protection Measures (also often referred to as DRMs or Digital Rights Management) are the generic names for technological mechanisms which restrict what can and can't be done with a digital work. This can include copy protection and print protection. It may also cover material locked to a particular system or software, or possession of licence key. Circumvention of TPMs can be against the law when not covered by a specific exception to circumvention.

How TPMs affect digital preservation:

- Can you make security copies?
- Can you migrate between formats?
- Can you combine content from different sources?
- Can you provide access - can you charge?
- Can you provide copies?

Problems of TPM: The technical restrictions limiting use by others can make the task of preservation difficult or impossible, and inappropriate use by your own organisation can cause you to fall foul of legal requirements in this area.

Intellectual Property (IP) Basics

Intellectual property as a term describes a variety of legal rights and concepts that at a very general level revolve around intellectual efforts rather than physical efforts, and applies to sometimes very different legal rights, including:

- Copyright
- Patents
- Trade marks
- Unregistered and registered designs
- Trade secrets (confidential information)
- Database rights

IP at its core is different than physical property like a house or your desk - it is something such as idea, expression, or symbol. It may be embodied in a physical object, but it is the ephemeral part that the law protects.

Another term used in this area is **industrial property** which usually refers to only patents and trade marks (leaving aside copyright).

What this appendix covers

This appendix concentrates on copyright, rights in databases, licensing, and their relation to the web. While other areas of IP, such as trade secrets, patents, and trade marks may also impact web preservation.

International aspects of IP

Intellectual property is essentially about national law, with international obligations. Within the UK, IP does not form a part of devolution for Scotland, Wales, or Northern

Ireland, and so UK IP law and policy comes from Westminster. However, as a member of the European Union, the UK must implement EU law as it relates to IP (and in many other areas). Beyond the EU, international IP law generally adheres to two principles:

- *National treatment* - the idea that treaty members give the nationals of other treaty members the same rights as their own citizens.
- *Minimum standards* - International IP treaties often harmonise IP law by setting minimum standards that members must meet by implementing them in their own law (a floor, but not a ceiling, of standards).

The key questions in terms of 'on the ground' IP protection usually depend most closely on what the relevant IP law is in the relevant national jurisdiction and not on international law such as EU law or international treaties. This is because *how* a jurisdiction has implemented its obligation for complying with its international obligations (EU or otherwise) for minimum standards differs between jurisdictions.

So for example, the Berne Convention for the Protection of Literary and Artistic Works requires in Article 7(1) that members grant copyright protection for the life of the author plus fifty years after his or her death. Some Berne members leave protection at 'life +50', while others, such as the European Union (including the UK) and the United States set the term at 'life +70'. They are free to do this because Berne only sets life+50 as a minimum standard. National treatment means that the UK must grant their higher 'life+70' standard to all works produced in Berne member jurisdictions, regardless of where they are from.

The international IP system and the internet

There have been two large technological impacts on IP law recently:

- digital technology, which allows for unlimited and perfect reproduction; and
- the Internet, which allows digital copies to flow relatively effortlessly globally and across jurisdictional borders.

As noted scholar Carlos Correa puts it, these technologies allow for 'unauthorized, perfect and costless copies and the almost instantaneous and worldwide distribution of protected works through computer networks'. This has put a large amount of pressure on an international system of IP protection that is essentially based around national rules and respect for national borders. The transborder shift of the Internet has changed the interplay between users and copyright owners, particularly in relation to areas of participatory culture and fair dealing / fair use and has caused a shift to reliance on contract (such as Terms of Service and EULAs) and technological measures (such as TPMs/DRMs).

Further resources on IP

The WIPO Intellectual Property Handbook: Policy, Law and Use. < www.wipo.int/about-ip/en/iprm/index.html>

UK Intellectual Property Office < www.ipo.gov.uk/>

Copyright

Copyright is a property right that covers certain types of works, including most creative and artistic works such as paintings, sculpture, literature, films, television, and music. Copyright can also include broadcasts, typographical layouts, sound recordings, and databases.

Obtaining a copyright

Copyright operates automatically and so you do not need to register or apply for a copyright in any way. As we will see, automatically acquiring rights differs from many other types of IP such as trade marks, which require registration with a governmental body in order to subsist. Once you create a work that meets the legal requirements for having a copyright you instantly have a copyright over that work. This means that you hold copyright over work you've produced, including past school papers, letters and emails to friends and family, and other works.

This explains that copyright is automatic, but not every work produced meets the requirements for being copyrightable. The law can vary on its requirements for copyright to subsist in a work, but for literary, dramatic, musical, and artistic works the law generally requires that the work be *original* and that it be *fixed*.

'Originality' doesn't mean that the work has to have some great spark of imagination - only that you didn't copy the work from another and that there was some level of effort, skill and labour to its creation. In practice, questions as to originality will be very fact specific, but the threshold is generally rather low.

'Fixation' only means that the work must be tangible or 'fixed' in some way, such as recorded on video or audio. As a practical matter, all web resources will meet this requirement.

What copyright protects

Copyright grants a monopoly to the rights holder over doing certain acts with the work, including to:

- Reproduce the work (make copies);
- Distribute the work to the public;
- Rent or lend the work to the public;
- Publicly perform the work;
- Broadcast the work or include it in a cable television service; and
- Adapt the work or to do any of the above with an adaptation of the work.

The rights owner of a copyrighted work can thus prohibit others from doing any of the above acts, unless an exception or limitation to copyright applies. Use of the work in ways not covered by fair dealing or another exception requires permission from the copyright holder. Permission to use a copyrighted work usually comes in the form of a licence, which is a legal document outlining what can and can't be done with the work.

Fair dealing and other exceptions

We've seen the acts copyright covers, such as reproducing, publicly displaying, adapting, and distributing a work. These rights have a number of exceptions to them under what is known as 'fair dealing' ('fair use' in the United States).

Because registration is not required, this also means that any work that you come across on the internet (or elsewhere) is likely under copyright.

Copymyths

We've mentioned a few of these, but they are worth considering again:

- You do not need to register a copyright - you get a copyright automatically;
- Posting a copy of a work to yourself is not required to get a copyright. At best it is a weak way of proving the date of creation of your work.
- Copyright doesn't protect ideas - if you have an idea for a better mousetrap or the next amazing website start up, you'll need to think about other ways of protecting your idea.

Moral rights

Moral rights protect the rights of personal authors (human persons and not legal persons such as corporations) over certain aspects of their association with their work:

- *Paternity* -- the right to be identified as the author of the work;
- *Integrity* -- the right to object to derogatory treatment of the work;
- *False attribution* -- the right to not be incorrectly attributed as the author of a work; and
- *Privacy* -- the right to privacy over photographs or films commissioned for private purposes.

These rights differ from other (economic) rights such as the monopoly right over distribution of the work. 2006 saw the introduction of a new right called the *resale right* that combines some aspects of moral rights with the economic rights - authors of works of graphic or plastic art under copyright have a right to a percentage (a royalty) of certain resales of the work.

Copyright term

Copyright term can be confusing and almost always requires some research to clearly identify whether or not a work is out of copyright, and if it is in copyright, who holds the current rights and for how long. Copyright term also can vary on the type of right and type of copyrighted work. But for example:

- Copyright in Literary Dramatic, Musical, and Artistic works (LDMA works) | 70 years from the end of the calendar year of the death of the author
- Copyright in films | 70 years from the end of the calendar year of the death of the last of - the principal director; - the author of the film screenplay; - the dialogue author; or - the film music composer
- Copyright in sound recordings | 50 years from the making of the sound recording, or if it is released, 50 years from its release

Moral rights | Integrity and paternity rights last as long as copyright. The right to object to false attribution lasts for 20 years from the end of the calendar year of the death of the author.

Term changes from jurisdiction to jurisdiction, and you should seek professional advice or assistance on the rights clearance process.

International copyright

It is important to note that because of international treaty obligations, work produced here in the UK is likely to automatically have copyright in most jurisdictions throughout the world, and will almost certainly have copyright throughout the EU and in places such as Australia, Canada, and the United States. This is because most countries have signed up to the major copyright treaties and thus automatically give protection to work produced in another country. Contrast copyright with patents or trade marks, where you

can get protection only after registration, which must be applied for on a jurisdiction-by-jurisdiction basis.

The opposite is also true, so when examining resources produced in other jurisdictions, they will have a UK copyright as well.

Public Domain

Not every work has a copyright, either because it can't have one in the first place, because the copyright term has expired, or because the rightsholder gave up their copyright. Works that do not have a copyright often get described as being in the 'public domain'. Without a copyright, anyone can publicly distribute, copy, adapt, rent, or do anything with the work (subject to other laws, such as trade marks, privacy, etc). Public domain works would be available for web preservation (from a copyright angle) without permission because there are no rights to clear.

Orphan works

Orphan works are works likely to be in copyright - therefore requiring permission to use barring an exception - but whose author cannot be established or cannot be located. Because users still need a licence but cannot get one, they cannot be used. Orphan works cause particular difficulty for those people such as documentary filmmakers and museums and libraries doing archival work. At present UK law only provides for a very limited process before the Copyright Tribunal. Otherwise, the solution requires a legislative response.

In web preservation, you will undoubtedly run into orphan works when you try to locate the rightsholder to a particular copyright.

Open content licensing

Introduction to licensing

You can think of copyright (and other IP rights) as a bundle of sticks. Each stick represents an individual aspect of copyright, like the right to create an adaptation, or the right to distribute a work. What at first glance could be a really broad right such as distribution can be thought of as a bundle of sticks all in themselves. You can break up the right to distribute only via the internet (and not at physical retail outlets). You could license the right to distribute in physical form (such as CDs) worldwide or you could break this right up geographically by jurisdiction, such as distribution rights in Europe but not North America or UK only.

Licences are how these sticks get broken up and handed over to others.

Industry practice in many areas has collected certain aspects of licensing together with terminology that you won't find in the actual text of copyright law. So for example, the music industry refers to 'mechanical rights', which you won't find in the UK law on copyright (the Copyright Designs Patents Act). In terms of the actual law, 'mechanical rights' licenses copying the music, issuing copies to the public, and renting or lending copies to the public.

When discussing licensing, it is important to distinguish between a *licence* and an *assignation*:

- *Licence* - retaining ownership but granting rights to others to use it under certain circumstances.
- *Assignation* - transferring the entire ownership to another (handing over the whole bundle of sticks). After assigning a work, you would no longer own the rights to the work. This means that you could infringe a work that you created if you use it without permission from the new rightsholder.

Once an IP rights has been assigned, the original creator or owner has no rights over the assigned work, whereas if licensed then the original owner or creator can still retain certain rights.

Open content licensing

Copyright, as you've seen, grants exclusive rights to the rightsholder so that, unless covered by a specific exception, users of the work must ask for permission. Because copyright lasts for quite a long time (life of the author plus 70 years in the UK for some works), you must generally assume that a work has copyright and thus often need to seek permission before using it. Always seeking permission and negotiating a licence (outlining the scope of the permission) can be a burdensome process, especially as even just tracking down who to ask permission from can be very difficult (and sometimes even impossible).

Open content licensing generally grants a wide range of permission in copyright for use and re-use of the work via a copyright licence, whilst retaining a relatively small set of rights for the rightsholder. In contrast, to the 'permission principle' built into copyright law, open content licensing reverses this default and grants permission for a very wide range of uses, but asks that users seek permission only in a limited number cases. This approach is often known as a 'some rights reserved' model, in contrast to the familiar 'all rights reserved' copyright notice asserting control by the owner of all copyright.

Some important points about open content licensing to keep in mind:

- Open content licensing still depends on copyright to grant some (usually most) permissions but retain some areas where permission would still be required;
- This style of licensing, like any other, can only be used on works by someone who owns the rights over the work or otherwise has permission to do so.

Creative Commons

One major example of open content licensing is that of Creative Commons (CC). This organisation, founded in 2001, maintains a number of easy to use licences available via their website. These licences allow for further distribution and copying of the work without further permission from the rightsholder. The main set of CC licences all offer a series of 'baseline rights' together with four 'licence elements' that can be mixed and matched to produce a licence through a point-and-click web interface:

The baseline rights:

- to copy the work
- to distribute the work
- to display or perform the work publicly
- to make digital public performances of the work (e.g., webcasting)
- to shift the work into another format as a verbatim copy (format shifting)

The four ‘licence elements':

- Attribution (BY) - you must credit the licensor of the work;
- Non-Commercial (NC) - you can only use the work for non-commercial purposes;
- No-Derivatives (ND) - you may not create adaptations of the work; and
- Share Alike (SA) - you may create adaptations of the work, but these must be under the same licence as this work. Note that SA and ND are mutually exclusive because SA requires that you allow adaptations of the work.

Attribution now forms a part of all current licences, thus these four elements form the six basic CC licences, with their common abbreviations in brackets:

- Attribution (BY)
- Attribution | No Derivatives (BY-ND)
- Attribution | Non-Commercial | No Derivatives (BY-NC-ND)
- Attribution | Non-Commercial (BY-NC)
- Attribution | Non-Commercial | Share Alike (BY-NC-SA)
- Attribution | Share Alike (BY-SA)

The generic or 'unported' set of CC licences only reflect the rules present in international treaties on copyright and related rights and not the actual law of the various world jurisdictions. Legal teams in over 40 jurisdictions have therefore ‘ported' these licences to meet their jurisdiction-specific legal needs, including specific sets available for Scotland and for England and Wales.

Your situation in attempting to preserve web resources may be unique, but in general all six of the basic CC licences are compatible with web preservation (assuming that the use is non-commercial).

Other open content licences

Creative Archive. This licence operates the same as the Creative Commons Attribution Non-Commercial Share Alike (CC-BY-NC-SA) licence. It adjusts the language to UK law, as well as adds some additional restrictions, including:

- No endorsement - you cannot use the work to promote, among others uses, political purposes; and
- UK use only - the licence only gives permission for use within the United Kingdom.

The Creative Archive licence was developed by the Creative Archive Licence Group, which consists of the BBC, the British Film Institute, Channel 4, the Open University, Teachers' TV, and the Museum, Libraries and Archives Council. The project launched in 2005, about the same time as the England & Wales and Scotland CC licences.

GFDL. The GNU Free Documentation Licence or GFDL is used primarliy by the Wikipedia project. It is a content licence built around the use case of reference materials and instructional text to accompany Free and Open Source Software (FOSS) and has some specific requirements when printing. It is similar to the Creative Commons Attribution Share Alike (CC-BY-SA) licence in how it works. Similar to Creative Commons, these licences are generally compatible with web preservation, though your situation may be unique.

Further open content resources

- Open Definition < www.opendefinition.org/>
- Creative Commons <http://creativecommons.org>
- Open Source Initiative < www.opensource.org>
- Free Software Foundation / GNU project < www.fsf.org/> < www.gnu.org/>
- Open Knowledge Foundation < www.okfn.org/>

Open data licensing

Open data

Data and databases are not a 'rights free' area where no intellectual property rights apply. International trade agreement TRIPs, for example, requires that members of the World Trade Organisation (WTO), including the EU, the US, and the UK, provide legal protection for databases. Rights covering databases can include:

- *Copyright* - both for the selection and arrangement of the database contents and over the contents of the database itself (the data), though factual information will generally not be protected by copyright.
- *Database rights* - The European Union's Database Directive requires member states to implement a "sui generis database right" covering the extraction and re-utilisation of the contents of protected databases.
- *Contract* - contractual obligations about what users can and can't do with a database and its contents can also be used to provide for protection.
- *Other rights* - rights such as trade secret and laws of unfair competition can also protect databases.

This rights thicket protecting databases and data can form a significant obstacle for the use and re-use of data, including for those wishing to preserve data made available on the web. Rights over databases will become increasingly important to web preservation as we move to a semantic web.

Science Commons Protocol

Science Commons was founded in 2005 and works on a variety of projects related to looking at rights issues related to scientific research, including legal issues surrounding data. Science Commons is a project of Creative Commons and is overseen by its board. On December 15 2007 Science Commons released their Protocol for Implementing Open Access Data. This protocol, written in the same style as a Request For Comment (RFC), outlines a legal standard for open access to data based on three principles:

- 3.1 The protocol must promote legal predictability and certainty.
- 3.2 The protocol must be easy to use and understand.
- 3.3 The protocol must impose the lowest possible transaction costs on users.

Guided by these three principles and Science Commons' experiences with Creative Commons licences and data, they arrived at an approach that calls for waiver of relevant intellectual property rights so that data could be treated as close to being in the public domain (no IP) as possible. Thus the protocol calls for waiver of:

- Copyright
- The sui generis database right in the European Union and similar protections
- Implied contract rights and rights in tort or delict such as unfair competition or trade secrets.

This protocol gets enforced through the use of a 'Open Access Data Mark', which will be managed by Science Commons and sister organisation Creative Commons. They will limit use of the mark to licensing schemes that comply with the protocol, so that users can be assured that the data labelled with the mark meets the criteria of waiving IP rights. The Science Commons protocol thus sets a standard that any licensing scheme can implement.

Open Data Commons

With the funding and support of information management company Talis, the Open Data Commons project was founded in the Autumn of 2007 to provide legal tools for sharing data. This project started through funding licence development by Mr. Jordan Hatcher (author of this guide for the JISC-Powr project) and Dr. Charlotte Waelde (University of Edinburgh). The eventual legal tool created, the Public Domain Dedication & Licence (PDDL), meets the Science Commons Protocol and is available to review at www.opendatacommons.org.

CC0 (CCZero)

Creative Commons has also implemented the Science Commons Protocol with their own public domain tool - CCZero or CC0 - based in part on their earlier work on their Public Domain Dedication tool currently available on the CC site. CCZero is at the same time an implementation of the Protocol for data and an expanded and clarified version of their public domain dedication. The CCZero tool applies to all types of content, and not just data. As of this writing, the CCZero draft legal text has not been finalised, but work is in process and available on the CC site. Regardless of the final text, because both CCZero and the Open Data Commons PDDL are (or will be) both compliant with the Science Commons Protocol any information covered by one licence can be fully integrated with information under the other licence because both place the work in the public domain.

Relevant links

- Science Commons Protocol <http://sciencecommons.org/projects/publishing/open-access-data-protocol/>
- Science Commons Protocol FAQ <http://sciencecommons.org/resources/faq/database-protocol/>
- CC0 Feedback and wiki <http://wiki.creativecommons.org/CC0_Feedback>
- CC0 legal text <http://labs.creativecommons.org/licenses/zero/1.0/legalcode>
- Open Data Commons homepage <www.opendatacommons.org>
- Public Domain Dedication & Licence <www.opendatacommons.org/odc-public-domain-dedication-and-licence/>

Further resources

Additional legal resources for web preservation

The Handbook and appendices have covered a wide range of legal issues. Your organisation or Institution will have a number of institutional resources and contacts on these legal issues as they are issues that they will already face every day.

In addition, there are several external projects and resources available to help you navigate the legal issues related to web preservation.

JISC Legal <www.jisclegal.ac.uk/> JISC Legal 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. Beyond the advice present on their site, they also offer an enquiry-based advice service.

Web2Rights project <www.web2rights.org.uk/> We've also addressed some of the legal issues surrounding use of 'Web 2.0' services such as Twitter, Facebook, Flickr, and others. 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 <www.oss-watch.ac.uk/> Though we haven't addressed free and open source software (OSS or FOSS) in depth, many of the computer programs used for web preservation may be open source. OSS Watch provides advice on their use and how to comply with their terms. They don't address open content (such as Creative Commons) or open data issues.

APPENDIX B: Records management: A guide for webmasters

What is 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 University and preserving an adequate historical record." (Edinburgh University, 2008).*

All Universities have corporate and business records, which need to be managed within the framework of a records management programme. Records include institutional and policy records, financial records, personnel records; departmental records, student records, academic records; health and safety records, records relating to building and estates management, and many more.

Records need to be kept for legal reasons; e.g. to protect against litigation, to prove title, to satisfy audit requirements, to protect property and business interests. *"Specific business functions and activities within the Institution may also be subject to specific legislation or to professional best practice or relevant ethical guidelines. For example, Finance activities are governed inter alia by the Finance Acts, the Taxes Acts, and the Pension Act 1995. Personnel activities are regulated by Employment Law."* Sometimes these reasons are governed by legislation and regulations that explicitly or implicitly require records to be kept, and in some cases can dictate the length of time records must be retained.

Increasingly records are needed by Institutions to comply with information legislation, such as data protection and freedom of information. (See below)

The task of Institutional records managers will be to identify, protect, store and manage these important business records, and ensure that they are retained and made accessible for as long as there is a continued operational and business need to do so.

> *"Records Protection and Security relates to measures taken to ensure that the vital records of the organisation are securely held in terms of physical and on-line access procedures and permissions, and that back up procedures are in place in the event of a disaster."*

Further, records management will ensure that the records are not kept for any longer than they should be, and carry out disposal and destruction in a timely manner (unless there is a requirement for permanent retention, or a historical / heritage need for retention, in which case records should pass into an archive for permanent preservation). This process is usually governed and carried out by means of agreed retention schedules (see below for fuller definition).

Within any paper-based record regime, records managers have been instrumental in preparing record inventories and survey lists, so that an Institution knows the whereabouts of all its current records. Records Surveys, Audits and Business Functions Analysis are all *"processes [which] aim to identify and compile an inventory of the main records series held by an organisation and map them to the various functions, activities and transactions carried out by individuals and business units."* Records management has also reduced costs by managing non-current or semi-current records in cheap offsite storage.

In the world of digital records and electronic file management, not all of these paradigms continue to apply, but traditional records management skills will continue to add value, and can underpin approaches to the management of websites and web resources.

How records management applies to web resources

For JISC-PoWR, a records management approach can help with some web preservation issues. Records management is not the only way to manage web resources, nor is it always the appropriate path. It will certainly be considered suitable when it is known that a website contains unique digital records, or if the website itself is considered a record that is worthy of capture.

Records management may also assist with certain classes of web-based resources which are not themselves part of the central website, but still require some form of managed retention and disposal, particularly if they are required for legal, audit, or business reasons.

The website as a record

It can be difficult to decide exactly whether and when a website should be treated as an authentic (and authenticable) record, a publication channel, or a publication itself - among other things. The Institutional website, as a site where information is frequently added, updated, removed and published by central and ancillary departments, could itself be viewed as a record of institutional activity. A case could be made for managing the website - or snapshots of the information held there - as a record.

This line of thinking could apply even if the digital copies of some information - e.g. PDFs of prospectus documents - are known to be copies, of which the 'original' or authentic copies are stored and managed elsewhere. What is more relevant is the fact that a certain version of a document was published and put online on a certain date, in a certain iteration of the website, and was available for consultation. This action of publication and dissemination could be said to constitute the record of an institutional decision.

The Institutional website could also be seen a potential place where unique records can be stored or generated. The website may also be a portal through which transactions can take place, which in turn generate further record evidence and audit trails of the transactions.

A records manager could legitimately ask if staff, students, or members of the public are making business decisions, or decisions about their academic career, based on the information they find on the website. A records manager would have a professional interest in the records of transactions, financial or otherwise, taking place over the website or via a web browser; whether record evidence is generated from such transactions; and whether the Institution needs to keep records of these transactions. In short, are there unique, time-based, evidential records being created this way?

Queensland (Australia) State Archives published a policy Managing Records of Webpages and Websites in 2004. They stated "This policy has been prepared to ensure that information made available on a public authority's websites, and the associated electronic transactions, are captured as public records and managed appropriately. Many Queensland public authorities have embraced web technology for use as an electronic 'shop front' to provide information to the public. Increasingly these websites are also used for electronic service delivery including electronic form lodgement and transaction payments for products and services online. It is critical that record-keeping

practices in the web environment comply with legislative, accountability, business, and historical requirements."

Including a website in a records management programme

A web manager could co-operate with the records manager (and vice versa) to the extent that the site, or parts of it, can start to be included in the Institutional records management programme. This may entail a certain amount of interpretation as well as co-operation. Institutional policies and procedures, and published records retentions schedules, will exist; but it is unlikely that they will explicitly refer to websites or web-based resources by name. Where, for example, institutional policies affecting students and student-record keeping are established, we need to find ways of ensuring that they extend their coverage to the appropriate and corresponding web resources.

The attraction of bringing a website in line with an established retention and disposal programme is that it will work to defined business rules and retention schedules to enable the efficient destruction of materials, and also enable the protection and maintenance of records that need to be kept for business reasons. The additional strength is that the website is then managed within a legal and regulatory framework, in line with FOI, DPA, IPR and other information-compliance requirements; and of course the business requirements of the Institution itself.

Information compliance

> *"The University will seek to ensure that its records management systems and procedures facilitate compliance with relevant legislation and University policies. Legislation of general relevance to the University as a whole includes the Data Protection Act 1998 (DPA) and Freedom of Information (FOI) legislation." (Edinburgh 2007)*

Data Protection and Freedom of Information can be two very strong drivers for records management. An Institution may be legally obliged to answer a request under the FOI Act, and make available certain data in recorded form. If the data or records cannot be found, or if they have been deleted, there may be consequences. A records manager will thus facilitate compliance by (a) ensuring that the records exist and can be easily retrieved and (b) by ensuring that reliable audit trails of records, including their location, their use, or their destruction under a retention schedule, can be produced as evidence of compliance. Under Data Protection, certain classes of records which contain personal information must not be retained for any longer than needed. Again, if an Institution has been managing and destroying its personal information records in line with the RM programme, then they will be complying with the DPA and will have evidence of their compliance.

What is an EDRMS?

An Electronic Document and Records Management System (EDRMS) is a safe, secure and governed information and record-keeping 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. Electronic Document and Records Management Systems have been emerging in the UK for many years, usually provided via commercial software vendors. Some of them just do Electronic Document Management (EDMS), some just do Electronic Records Management (ERMS) and some perform both functions.

An EDRMS usually sits outside a network, and applies automated records management rules to documents that are created and stored in it. The network can be the same as a Windows folder tree structure, for example, although many organisations implementing an EDRMS have taken the opportunity to reorganise their electronic filing to create a 'file plan'. Often, this has been built along shared and functional lines, instead of reinforcing departmental divisions.

Documents - usually word-processed files, spreadsheets, emails and other 'static' records - can be stored in the EDRMS as they are created. The EDRMS adds a profiling step to ensure that the correct contextual metadata is assigned. Best practice information and records management requires classification and metadata to be captured at the beginning of document creation, rather than at the perceived end of the life-cycle. A document once stored is then 'declared' as a record. This means the content is frozen, thus ensuring its authenticity, its reliability and security; its fixity is assured, and it cannot be changed by another user.

The EDRMS also manages automated retention and disposal scheduling, by applying rules and sets of rules to the collections of records. This task is made easier if the 'functional filing' approach is used. This feature applies timed reminders to related series of records, enabling their timely disposal or destruction by a system administrator. Most of these EDRMS functions described above are managed by its underlying database, which allow it to behave as a species of 'electronic registry'.

Retention scheduling

> *"All records have a life cycle from creation/receipt (birth), through into the period of active currency (youth), thence into semi-currency, e.g. middle-aged closed files that are still referred to occasionally, and finally either confidential disposal or archival preservation. In the digital age it is especially important to introduce conscious management at the earliest possible stage as this will determine the ultimate extent of control over electronic material." (Edinburgh op cit)*
>
> *"The University will develop a schedule for retention and disposal of records drawn up as a result of applied best practice i.e. based on records surveys, analyses, agreements with business units, etc. The preparation and maintenance of this will primarily be the responsibility of the records manager. Substantial input from the relevant officers will be required if the schedule is to reflect the business needs of the University corporately and of the individual departments and units." (Edinburgh op cit)*

Records Analysis and Retention Schedules are "processes which apply various 'appraisal criteria' such as legal, operational, administrative and historical requirements, to determine how long a particular series need to be retained."

The Records Disposal process "implements and documents the operation of the retention schedule recommendations, ensuring that records in all formats that are no longer needed are disposed of confidentially, or reviewed after formally agreed periods of time, or permanently preserved as the archival record."

Archiving / preservation

One final outcome from a records management programme is the archiving - i.e. 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 University aims to preserve those records designated as having permanent legal, administrative or research value at the earliest possible stage in the records life cycle. Given the rapid pace of technological change in the digital age and the vulnerability of digitally held information, archival status records held solely in electronic formats need to be designated as such soon after creation or receipt. The procedures required to achieve this aim will be developed in consultation with the University Archivist and will follow emerging professional practice in digital archives preservation." (Edinburgh op cit)

Bibliography

Chapter 5

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.

Chapter 7

Netpreserve.org. Downloads. www.netpreserve.org/software/downloads.php. Retrieved 8th October 2008.
Harvard University Library. Web Archiving Resources. http://hul.harvard.edu/ois/systems/wax/resources.html. Retrieved 8th October 2008.
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. www.dlib.org/dlib/may08/paynter/05paynter.html. Retrieved 22nd September 2008.

Chapter 9

Guy, M. (2008). "When Do We Do Fixity". On JISC-PoWR blog: http://jiscpowr.jiscinvolve.org/2008/07/14/when-do-we-fixity/. Retrieved 8th October 2008.
JISC Infonet (YEAR). *Managing The Information Lifecycle*. www.jiscinfonet.ac.uk/infokits/information-lifecycle. Retrieved 8th October 2008.
Edinburgh University. Records Management Department. www.recordsmanagement.ed.ac.uk/. Retrieved 8th October 2008.

Chapter 11

Digital Preservation Coalition (2006). Decision Tree. www.dpconline.org/graphics/handbook/dec-tree.html. Retrieved 8th October 2008.
JISC Infonet (2007). *Guidance on Archival Appraisal*. www.jiscinfonet.ac.uk/partnerships/records-retention-he/archival-appraisal. Retrieved 8th October 2008.

Chapter 12

Warwick Blogs. http://blogs.warwick.ac.uk
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
Retrieved 12th September 2008.
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
Retrieved 14th September 2008.

Chapter 13

UKOLN (2008). Bath Web Tour. www.ukoln.ac.uk/web-focus/experiments/experiment-20080612/bath-web-tour.html. Retrieved 17th September
University of Virginia (2008). "History of U.Va. on the Web". www.virginia.edu/virginia/archive/index.html. Retrieved 17th September 2008.
Information Commissioner's Office (2008). Background to publication scheme initiative. www.ico.gov.uk/what_we_cover/freedom_of_information/publication_schemes/new_publication_scheme.aspx. Retrieved 17th September 2008.
www.ja.net/
www.ja.net/services/domain-name-registration/register.ac.uk/eligibility-ac.html
Downes, S. (2004). "The Weblog as the Model for a New Type of Virtual Learning Environment?". www.downes.ca/post/6869. Retrieved 18th September 2008.
Berners-Lee, T. (1998). "Cool URIs Don't Change!" www.w3.org/Provider/Style/URI. Retrieved 19th September 2008
Wikidot (2008). Community Forum: Backup and other questions. http://community.wikidot.com/forum/t-10000/backup-and-other-questions. Retrieved 22nd September 2008.

Atlassian Software. Universal Wiki Converter.
http://confluence.atlassian.com/display/CONFEXT/Universal+Wiki+Converter.
Retrieved 22nd September 2008
Techcrunch (2008). "Slideshare secures $3m for embeddable presentations."
www.techcrunch.com/2008/05/07/slideshare-secures-3m-for-embeddable-presentations/.
Retrieved 22nd September 2008.
University of Southampton (2008). OR08 Publications. http://pubs.or08.ecs.soton.ac.uk/.
Retrieved 22nd September 2008.
The Guardian (14/8/2008). "Twitter drops text support in UK."
www.guardian.co.uk/technology/2008/aug/14/twitter.
Retrieved 22nd September 2008.

Chapter 16

JISC/University of Glasgow (2007). *espida: Making it happen by getting real*. www.gla.ac.uk/espida/.
Retrieved 9th October 2008.
JISC/University of Glasgow (2007). *espida Model Handbook*. www.gla.ac.uk/espida/documentation.shtml.
Retrieved 9th October 2008.
Emmott, Stephen (2008). "Preservation of Web Resources: Making a Start." *ARIADNE* issue 56 July 2008.

Chapter 19

JISC (2008). *Change management Infokit*. www.jiscinfonet.ac.uk/infokits/change-management. Retrieved 9th October 2008.
Korm, N. and Oppenheim, C. (2007). *IPR Risk Assessments, Rights Clearances And Rights Management: Practical guidelines for content creators within FE and HE. HEFCE*.
Bailey, S. (2008). TFPL blog. http://tfpl.typepad.com/tfpl/2008/06/steves-bailey-m.html.
Retrieved 23rd September 2008.

Appendix B

DublinCore: http://dublincore.org/
METS: www.loc.gov/standards/mets/
PREMIS (Preservation Metadata): www.loc.gov/standards/premis/
JISC. Information Lifecycle Infokit. www.jiscinfonet.ac.uk/infokits/information-lifecycle. Retrieved 9th October 2008.
Hodge, G. M. (2000). "Best Practice for Digital Archiving: An Information Life Cycle Approach." *D-Lib magazine* Vol 6 No 1, January 2000.www.dlib.org/dlib/january00/01hodge.html
Greenberg, J., et al (2002). "Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization." *Journal of Digital Information*, Vol 2, No 2 2002.
http://journals.tdl.org/jodi/article/view/jodi-39/45
Edinburgh University (2007). Records Management Policy Framework.
www.recordsmanagement.ed.ac.uk/InfoStaff/RMstaff/RM_framework.htm. Retrieved 9th October 2008.

Glossary

Appraisal
The aim of archival appraisal should be 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. The records 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 of its impact locally and in the wider world. The selection process should also facilitate the survival of records which contain unique information incidental to their main purpose or function but which, nevertheless, might have research value.
Source: www.learninfonet.ac.uk/partnerships/records-retention-he/archival-appraisal

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 life cycle.

Archiving (2)
Backup of digital resources. It's best to consider the scope of digital preservation as much broader than digital archiving, though the terms are often used interchangeably. Because, in computing generally, 'archiving' is the process of backup and offline storage of data, 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**

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 significant knowledge about or control over the technical infrastructure that supports them.

Content Management System (CMS)
A content management system (CMS) is a computer application used to create, edit, manage, and publish content in a consistently organized fashion. CMSs are frequently used for storing, controlling, versioning, and publishing industry-specific documentation such as news articles, operators' manuals, technical manuals, sales guides, and marketing brochures. The content managed may include computer files, image media, audio files, video files, electronic documents, and Web content.
Source: http://en.wikipedia.org/wiki/Content_management_system

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 your Institution's efficiency and effectiveness. If managed properly, assets can maximize 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 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
This 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 pro-active 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 Managment 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

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: 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: 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

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. For an Institution, this would include 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.

Lifecycle management
The information that your institution creates and uses can either represent an asset or a liability. Into which of these camps it falls is largely dependent on how it is managed. Put simply, the concept of information lifecycle management is about making sure you ask yourself the right questions at the right time 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.
Source: www.jiscinfonet.ac.uk/infokits/information-lifecycle/introduction/index_html

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; these support an extremely wide range of activities. For example, there are some initiatives specifically concerned with the development of metadata schemas for long-term preservation.
Source: www.dcc.ac.uk/resource/curation-manual/chapters/metadata/

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 organization 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. 'Long-term' is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. 'Long-term' may extend indefinitely. In the OAIS reference model there is a particular focus on digital information, both as the primary forms of information held and as supporting information for both digitally and physically archived materials. Therefore, the model accommodates information that is inherently non-digital (e.g., a physical sample), but the modeling and preservation of such information is not addressed in detail.
Source: http://en.wikipedia.org/wiki/Open_Archival_Information_System

Off-air archiving
In radio and television broadcasting, the capture of a copy of a programme as broadcast, for example on video or audio tape. Quality is inevitably lower than for copies made directly from studio masters; however many early broadcasts only survive thanks to off-air recording by enthusiastic amateurs. There are many similarities between off-air archiving of broadcast programmes and remote harvesting of websites.

Preservation
Digital preservation is defined as a “series of managed activities necessary to ensure continued access to digital materials for as long as necessary”.
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”. Records occur in all types of recording media.
Source: www.recordsmanagement.ed.ac.uk/InfoStaff/RMstaff/RM_framework.htm

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

UK Web Archiving Consortium (UKWAC)

UKWAC is a consortium of six leading UK institutions who got together in 2004 to explore the best way to collect and preserve web materials. Websites have been selected for inclusion in the archive by members of the UK Web Archiving Consortium. Each consortium member selects and captures websites relevant to their individual collection development policies. The archive includes a wide variety of websites including (but not limited to) e-theses, research papers, literary and creative projects, government websites, museum web pages, blogs and sites of cultural, historical and political importance.

See www.webarchive.org.uk/

Wayback Machine
The access tool operated by the Internet Archive.

Web harvesting
Web harvesting is the process of downloading web material.

Website
A collection of Web pages, images, videos and other digital resources, hosted on one or more web servers, usually forming some discrete and coherent whole.

INDEX