PoWR

The Preservation of Web Resources

Digital preservation issues relevant to the UK HE/FE web management community

DRAFT HANDBOOK as at 10/09/2008

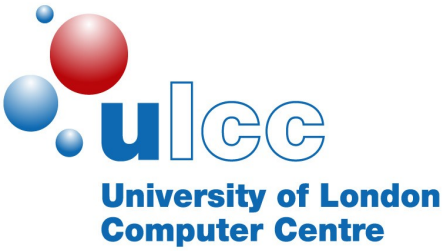

Table of Contents

University of London Computer Centre,
20 Guilford Street, London WC1N 1DZ
www.ulcc.ac.uk
UKOLN, The Library, University of Bath, Bath, BA2 7AY
www.ukoln.ac.uk

1 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 the 27 June (London), 23 July (Aberdeen) and 12 September (Manchester). The workshops 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 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 JISC PoWR 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 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, legal issues, 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. The content has been conceived as a narrative, starting with basic, familiar ideas and issues, and moving towards more complex issues. The Handbook is divided into four main sections, one dealing with Institutional policy issues, the second with Web resources, the third suggesting Actions and approaches, and the fourth outlining a range of legal issues.

There is a separate Appendix for Records management, as it covers things that may span several issues. We have also added some detailed explanations of issues and practical illustrations of a point, by using case studies and use cases. The intention is that the Appendices constitute background material, which may not necessarily be read by everyone.

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

This early draft is being exposed at the 12 September Workshop, as part of the ongoing process of peer-review and quality assurance.

Institutional web use

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

The reason for this is that we see these systems as hermetic and essentially self-managed by their professional user base (e.g. librarians, accountants). A preservation policy, or an Information Asset policy, must encompass all web resources, including these; but the data in these systems are generally be less at risk than less strictly controlled content on the web.

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.

PART A: INSTITUTIONAL POLICY AND PRESERVATION

2 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, 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 to the University and are held in digital form. They may increase in value through sharing and repurposing.

1. To protect your institution

University 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 University 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 University'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' university. 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.

3 What do we mean by preservation?

Below are summary outlines of certain definitions (which shade into suggested approaches and solutions), which will be explored in fuller detail in subsequent sections of the handbook. 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 preservation. The handbook will demonstrate how best to appraise your web resources, help to determine which approaches are most suitable for each resource, and which collection methods to use.

Preservation is possible!

Institutions when faced with the task of preservation projects of any sort can find it so daunting that, in the end, nothing gets 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.

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. Records management approach may help to enact "preservation" for business records or legal reasons, even if you don't intend to keep resource beyond its expiration.

Protection

Protecting a resource from loss or damage, in 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

What we mean by this is digital preservation as defined by the OAIS model, which is published and widely accepted (internationally) as a feasible model for DP. For web resources, we have to assume that an institutional decision has been made to keep the resource **permanently**.

What web resources need to be preserved?

This question will be dealt with throughout the handbook, but as a starting point we propose three main classes of web resource that are nominated for preservation. These are:

1. Records
2. Publications
3. Artefacts

Definitions, characterisations and examples of these classes will be given elsewhere.

We also want to be clear about *what* we are preserving: which could be web content, appearance, function and behaviour, or access and location; or a combination of all of these. KA sees "a distinction being made between preserving an experience and preserving the information which the experience makes available. Both are valid preservation approaches and both achieve different ends."

4 Web Preservation Scenarios

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 University 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 university 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 university 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 Guy

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

The Web Guy

"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."

5 Responsibility for preservation of web resources

Ownership of the problem

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

There's probably many sides to '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 University website, the web-publication programme, web resources that have value, and preservation of these. 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 may be assigned some degree of responsibility for the creation, management and storage of some of their web resources. Some of these people 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, CVs etc.

Implied 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.

Those involved in providing institutional Web services may not be interested in issues related to preservation. This translates into a potential risk for the Institution. Web managers may not have a particularly strong interest in this topic. If this is the case, it will be difficult to persuade them of the need to invest resources in this area and to gain the necessary commitment from senior managers and policy makers.

Resourcing web preservation

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 a technological problem exclusively. The solution does not lie in buying new software or more software. There is no single technological 'solution' that will fix everything. There is not one single tool that addresses all possible web preservation issues: behaviour, dynamic content, scripts, versioning, emerging standards, etc.

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 is the least of our worries.

Sponsorship for web preservation

The espida project in Glasgow (http://www.gla.ac.uk/espida/) offers a good 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 will 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 paraphrase espida slightly:

> "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."

See the espida Model Handbook, at http://www.gla.ac.uk/espida/documentation.shtml.

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. This Handbook will endeavour to bring together institutional stakeholders who might not otherwise encounter each other, such as records managers and web managers. Collaboration is the key.

> "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."
>
> (Stephen Emmott, from *Preservation of Web Resources: Making a Start*: ARIADNE issue 56 July 2008)

6 Institutional Strategy

Shaping of policy

What do you want to achieve? The Handbook will, without being proscriptive, suggest areas and resources you can target to help you define and decide what it is you want to get out of preserving your web resources. This includes adapting methods which can help you identify and measure the value of these resources.

Web preservation is a big topic and we're not even pretending to deal with all of it. The aspect that we care about - 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.

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 on *Selection*)
- Strategic approach - a comprehensive web-archiving programme over many years, affecting the entire Institution

Further suggestions in *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

PART B: WEB

7 What's on your web?

In this section we outline the things we think are likely to appear on University 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 University websites

If we consider the website as a major tool of the university 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 an 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 have we got?

- 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
- Systems for managing assessments and examinations
- Blogs
- Wikis

Some of these may contain interactive, social software, or transactional elements.

In this area, you may want to start thinking about some form of characterisation of 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 web site, you may start to ask questions about who is using them and what they are doing. This divides into two pertinent questions:

Whose is it?

Identifying relevant stakeholders: Students, academic staff, tutors, university 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.

What use are they making of the resources?

- What are they doing?
- Are they creating original materials?
- Are they creating and storing records?
- How do they create the resource?

Where is it?

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

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

From the first workshop, we sensed there was a general lack of centralised awareness about the number of web sites 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." 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.

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 our chapter on Information Lifecycle Management. It could take the form of a physical survey, visits to departments, meetings with people, or a questionnaire. Or a combination of all of these.
- Research.
- Approach your Institutional hostmaster or Domain Name Server (DNS) manager. This person should be able to inform you about all the URLs, domains and subdomains 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.

8 What are the main risks and issues peculiar to websites?

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

Frequency of change

Quantity and range of resources potentially needing preservation

Continuity

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

Because of the ease with which web sites 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 change. 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

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

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?

Ownership

- Web resources may be managed by different departments, faculties etc.
- Sub-sites may be temporary / ad hoc

Databases and deep websites

- Preserving underlying database may not preserve user's experience on the web

Streaming and multimedia

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 - hosted elsewhere but containing valuable institutional material. How best can this be retrieved? Who "owns" it? Is login authentication required to access some or all of the information?

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

9 Web Content Management Systems (CMS)

The section 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 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.

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, but the 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 whether the rollback functions and version control tools produce 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 truly restore one page to the state it was a month ago, we 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. Is it possible to migrate a website from one product to another? CMS internal management of content, data and metadata are application-specific, and 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 system; 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; nor can we 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 seems to be the recommended approach to dealing with these issues, but not if the CMS is going to defeat the crawl engine.

10 Web 2.0

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.

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. HE institutions are simply never going to be able to keep up with the relentless innovation that Web 2.0 represents, but their staff and students are impatient to use exciting new ways to work. See JISC-PoWR on Preservation and innovation.

In general, Web 2.0 applications fall into one or more of the following categories:

Third-party services for hosting content - often multimedia, streaming

Examples: Flickr, Slideshare, YouTube.

Uses: Galleries of images and videos; sharing presentations

Advantages: Quick, easy and free to set up.

Preservation issues: Third-party hosted applications are an inevitable cause for concern when considering preservation. Data and content at held at the provider's server, and the user usually has whatever access the provider considers necessary through a web interface. It can be extremely easy to add content, but not usually so easy to extract it in a reusable format. Of course, some Web 2.0 services are merely storing and presenting content that has been created elsewhere.

Preservation approaches: Adding something to Flickr, Slideshare, YouTube or a podcasting host presupposes that a photo, presentation, video or audio already exists. If material merits preservation, one would expect to use the original master as the basis for the preservation copy. See also http://widwisawn.cdlr.strath.ac.uk/issues/vol6/issue6_1_4.html for ideas about YouTube. JISC-PoWR on Slideshare

Collaborative/social web-based tools

Examples: Wetpaint, Ning, wikis, blogs, instant messaging, Skype, Twitter

Uses: Collaborating to create websites, documents, hypertexts and online journals; synchronous and asynchronous discussion

Advantages: Quick and easy and free to set up; powerful communication and collaboration features

Preservation issues: Collaborative tools are generally third-party hosted, and this creates the problem that the content is generally created directly in them - whether posts on Blogger, web pages on Wetpaint, or groups and discussions on Facebook. (See the Wiki post...).

Preservation approaches: Self-hosting of blogs and wikis potentially gives an institution much greater control over their use, management, etc. See JISC-PoWR on Twitter, JISC-PoWR on Wikis.

Personalised news, portal and aggregation services

Examples: Netvibes, Delicious, Technorati

Uses: Collating and aggregating news items and bookmarks

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

Preservation issues: Very personalised views, but views of content existing elsewhere. Is it necessary to preserve and keep records of individuals' personalised environments? Sometimes these tools may be used by groups, on projects or in departments, for example, for sharing news items or links of common interest: is there a value in preserving these resources. Who owns the content?

Preservation approaches: Harvesting may work. To capture information only, RSS feeds can be archived or used to create a database of objects and/or links.

Google apps and "cloudcomputing"

Examples: Google Docs, Google Sites, Google Groups

Uses: Online creation, collaborative editing and management of documents, spreadsheets, websites and newsgroups/bulletin boards

Advantages: Unlimited storage, largely free to use, powerful features.

Preservation issues: Data stored and managed by Google: only available as long as, and whenever, Google is. File formats of contents. Authentication/login a potential barrier to harvesting material not made completely open online. Potential IPR, ownership and Terms Of Use issues.

Preservation approaches: Use existing management systems where appropriate, e.g. DMS for documents and spreadsheets.

Approaches

It is worth noting though that in all cases, the web applications are designed to create a web-based view of the information entered in them. They are designed to create something that can be viewed as an ultimately static page in a web browser. Therefore they are all potentially susceptible to remote harvesting, which will create a local, browsable image of the remote resource, containing much, if not all, of the content and presenting it as it was presented on the web. However, there may be issues relating to user account login, streaming content, and terms of use policy.

Do we need to preserve News feeds?

Relevant material at:

- JISC-PoWR on Preservation and innovation
- JISC-PoWR on Twitter

PART C: ACTION

11 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.

- University 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 University 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 University, or advises on policy and current development activities.

Assessing your policies

You could 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

From the first workshop, we learned that none of the Institutions attending has web material included in their retention schedules. Nor did they admit to 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.

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 treat web resources.

12 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

What classes of resource to preserve

JISC PoWR suggests three discrete classes to which we could assign web resources for preservation purposes.

1) RECORD

Traditional description of 'record'

"Recorded information, in any form, created or received and maintained by an organisation or person in the transaction of business or conduct of affairs and kept as evidence of such activity." (http://www.recordsmanagement.ed.ac.uk/InfoStaff/RMstaff/RM_framework.htm)

How is a web resource a 'record'?

- If it constitutes evidence of business activity that you need to refer to again
- If it is evidence of a transaction
- If it is needed to be kept for legal reasons

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 University 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 (eg automated email responses), which may in turn needs to be captured out of the process and stored in a place where it can be retrieved an accessed.

2) PUBLICATION

Traditional description of a publication

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

How is a web resource a publication?

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

Examples

- Website contains the only copy of an important publication. See above.
- Web page constitutes 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, navigation elements that make it different to the original source.
- Web page constitutes a mix of published information. For example, a page of original University material combined with an RSS feed from outside the University.

3) ARTEFACT

How is a web resource an artefact?

- If it has intrinsic value to the Institution for historical or heritage purposes
- If 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

Examples

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

What resources can be excluded?

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 might include:

- Digitised images
- Research databases
- Electronic journals
- Ebooks
- Digitised periodicals
- Examples of past examination papers
- Theses

Institutional repositories. IRs are a similar example. For example, 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.

Resources that have no value

University 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.

In other cases, the service may not even generate any informational material or records of 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.

13 Selection

How to decide what web resources to capture

TNA on 'Selection'

This section on selection will help you decide how to scope out your collection policy in a logical and managed way, and 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 we have covered in the two previous sections.

Adrian Brown's chapter on Selection from *Archiving Websites* may have some useful concepts and approaches. However, his advice isn't addressed to webmasters or Universities; rather it makes the assumption that you are a digital repository or digital library, working to build a web-archiving programme or a themed collection of websites. That said, the principles are applicable; as part of your web-archiving programme, you are selecting websites, web pages, and other web resources for preservation. The concepts can be tailored to build a selection policy that suits any HFE Institution.

The advice 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 this Handbook's sections on Institutional buy-in and existing policies).

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.

A discussion of selection methods

Unselective approach. This 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, 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 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. For these reasons, PoWR do not recommend an unselective approach.

Thematic selection. TNA regard this as 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. This approach could feasibly be mapped to a University structure, with selection based on such things as subject matter, the collections of a creating department, the genre of the resource, or by domain.

Once again we should stress this 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.

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 Library or Repository wishes to narrow its scope by identifying very specific web resources for collection, such as a single web publication or website. This approach does tend to define implicit or explicit assumptions about the material that will not be selected, and therefore not preserved.

In the HFE world, that exact model does not quite apply, but PoWR recommend a selective approach based on the needs of your Institution, and the agreed framework for your programme that has determined what

resources need to be selected for preservation. If you have been successful, your selection policy will have been determined by the internal and external drivers already described in "What are the drivers for your web-archiving project?"

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 = content (which could be words, images, audio…)

Experience = the experience of accessing that content on the web, which all its attendant behaviours and aspects

Deciding which aspects of your web resources to capture can be informed to a large extent by the Institutional drivers, and the agreed reasons for retention and preservation. See the Handbook chapters on 'What are the drivers' and 'Why would we want to preserve what's on the web?'.

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, who used it, who 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.

14 What approaches and techniques can you use?

We propose that success is likely to arise from a mix of skills from IM professionals - some asset management, some records management, some archival awareness; input from web master; awareness of storage costs. Many different approaches are suggested, which take from 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 class includes quick-win solutions, actions that can be performed now to get results, or to rescue and protect resources that you have identified as being most at risk. Actions include domain harvesting, remote harvesting, use of the EDRMS, use of the Institutional Repository, and '2.0 harvesting'. These actions 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 class includes longer-term strategic solutions which take more time to implement, involve some degree of change, and affect more people in the Institution. These include approaches 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 resource-hungry 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 the chapter 'Getting other people to 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 Section 2). 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 (an archivist?) 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.

JISC PoWR maintains that the selective collection of managed web resources is preferable to indiscriminate harvesting. See our guidance on selection, records management, etc.

Migration

Migration of resources is a form of preservation. Migration is moving resources from one operating system to another, or from one storage 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. (And see http://jiscpowr.jiscinvolve.org/2008/07/14/when-do-we-fixity/).

We know that's technically feasible to capture Instant Messaging outputs as HTML pages which could

conceivably be saved into an EDRMS. The question is 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 the *Records Management Appendix* for 'What is an EDRMS'?

Strategic Approaches

Information Lifecycle guidance

Information Lifecycle Management (ILM) involves recognised professional standards and practices, leading to better management of information. It's recommended by PoWR as 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.

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. (See http://www.jiscinfonet.ac.uk/infokits/information-lifecycle).

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.

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.

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 Chapter 16 for more on ILM.

Adapting records management approaches

Recognised professional standards and practices leading to improved management of records in the website, or web-based record material.

Records management is recommended by PoWR[?] 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.

Use the JISCInfoNet published guidance, for example; and the guidance from http://www.recordsmanagement.ed.ac.uk/.

Records management, including many of the technical terms, is explained in more detail in our Appendix on Records Management.

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 analogues 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 a University.

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, eg 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 University.

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.

15 Getting other people to do it for you

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 the JISC. To date, UKWAC's approach has been very selective, and determined by written selection policies which are in some ways quite narrow. The 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 UK University websites have yet been collected. (The National Library of Wales are taking snapshots of Welsh University 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 (eg 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.

Internet Archive

The **Internet Archive** (http://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 (http://www.archive-it.org/). 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 University assets are missed out by the UKWAC or IA approach. For example library catalogues, image collections, e-prints collections with a database, and interactive teaching materials.

HANZO

Hanzo Archives is a commercial web-archiving company. See http://www.hanzoarchives.com/. They claim to be able to help HFE institutions archive their websites and other web-based resources.

They offer a software as a service solution for web archiving. This is based on Hanzo Enterprise software, which provides for the full end-to-end requirements for web archiving: crawl management, archive management, full-text search, archive browsing, retention, etc. The service includes aspects of quality assurance, crawl engineering and support on top of the software, to make sure the archive content is of the required quality. The software includes an advanced archival web crawler that can capture a very wide range of web content. In addition, they use APIs and scripts to archive more interactive web-resources. For FOI and DPA enquiries, the archive browsing and full text search are discovery features which can operate on individual archive resources or across all of them simultaneously, for a specific archive time or across a range of times.

To provide an estimate of costs, Hanzo need to construct a set of scenarios. For this, you would need to consider the following questions, with numbers or a range of numbers, which can be estimated against.

1. Average number of websites and web-resources, etc.

2. Frequency of capture (weekly, monthly, quarterly, etc.)

3. Retention period in years

4. A sample of websites and web-resources, to review the technological topology of the sites

Further observations on collaborative approaches

It's possible for ownership to be shared at multiple levels; for instance, one can depend on a national infrastructure or service to do the actual preserving, but still place responsibility on the creator or the institution to make use of that national service. That's effectively the situation with social science datasets and the UK Data Archive - it exists because of national decision-making and national funding, but material only ends up there if the creators deposit it (and if the archive accepts it.)

Lots of Copies Keeps Stuff Safe (LOCKSS) - http://www.lockss.org/lockss/Home. This is primarily a digital library initiative, and intended for materials like e-journals, but the community-based model is attractive.

16 Information Lifecycle Management: Creation

In this section we consider lifecycle management in more detail. 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."

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 http://www.jiscinfonet.ac.uk/infokits/information-lifecycle).

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.

Website links need to be reliable too; see the section on the TNA Continuity Project.

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. 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. (See elsewhere for website as a publication). 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 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 http://www.loc.gov/standards/mets/
- PREMIS (Preservation Metadata)

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) http://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

17 Approaches to Preservation

This section could include, for example:

- Protection of web resources from loss, removal, damage (i.e. short-term preservation)
- Records management (mid-term preservation with a view to eventual disposal or destruction) - already explored
- Digital preservation (long-term to permanent preservation)
- An explanation of some elements of the OAIS model - eg Ingest and Access stages, the creation of SIP, AIP and DIP
- The meaning of digital curation
- Managing preserved content - access, storage
- Presentation of archived copies - making it clear they they are archived copies and not the live site
- Strengths and weaknesses of existing approaches to web preservation

Definitions of certain terms

Digital preservation is defined as a "series of managed activities necessary to ensure continued access to digital materials for as long as necessary" (Digital Preservation Coalition, 2002).

Digital archiving. 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.

Digital curation. A third term, digital curation, has recently gained prominence. This 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.

What we want is to "maintain access to digital materials beyond the limits of media failure or technological change". This leads us to consider the longevity of certain file formats, the changes undergone by proprietary software, technological obsolescence, and the migration or emulation strategies we'll use to overcome these problems.

By **migration** we mean "a means of overcoming technological obsolescence by transferring digital resources from one hardware/software generation to the next."

In contrast, **emulation** is "a means of overcoming technological obsolescence of hardware and software by developing techniques for imitating obsolete systems on future generations of computers."

The Digital Preservation Coalition (DPC) is a consortium of many leading institutions working in the field, including The British Library and The National Archives. Its online handbook contains much excellent information and includes a useful glossary. Many quotations and definitions above are taken from the DPC's online handbook.

The best source of information about digital curation is the Digital Curation Centre, based at Edinburgh University.

18 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.

JISC PoWR are recommending a *selective* approach. 'Keeping everything' is not a recommended option. As we have shown up to this point, even 'capturing everything' is difficult enough when it comes to websites.

Storage costs for digital materials may be cheap, but just think of the cost of storing accumulated copies of snapshots of your website, particularly if you haven't been selective enough and you 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 retrieve important information when needed.

Always bear in mind the records management perspective. The more you keep, you 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 for keeping things

- 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

Remind yourself of the Institutional drivers for web resource preservation.

Use guidance from your archivist, and your records manager, who will also help you with selection and appraisal.

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 University. A researcher's blog, which represents two years of accumulated wisdom from herself and her students, is thrown simply because when she leaves, her computer accounts are dismantled.

Example from the first workshop: There is often a conflict of requirements among stakeholders. One good illustration is the online tutorials at Open University, which (according to one rule of behaviour) can be 'removed' from the website six months after completion. The trouble is that they are also being thrown away at

that time. Whoever throws them away isn't asking the archivist (or anyone else) if copies need to be kept. Already the archive of tutorials has gaps in the collection.

External drivers for keeping things

These 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 University's reputation is a consideration. 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 University from liabilities.

19 How do we appraise the value of a web resource?

The previous Handbook chapter deals with stakeholder requirements, 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.

A few questions to help you decide:

- Is 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."

See http://www.dpconline.org/graphics/handbook/dec-tree.html.

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).

Taken from *Guidance on Archival Appraisal*, JISC InfoNet[2], January 2007. For the full document see http://www.jiscinfonet.ac.uk/partnerships/records-retention-he/archival-appraisal.

PART D: LEGAL MATTERS

20 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 for 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 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 universities.
- 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 University 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.

This 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 University 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 University holds information and gives the University 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); and
- 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 (which we will address below) 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. <http://www.wipo.int/about-ip/en/iprm/index.html>

UK Intellectual Property Office <http://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 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

Intro 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* <http://www.opendefinition.org/>
- *Creative Commons* <http://creativecommons.org>
- *Open Source Initiative* <http://www.opensource.org>
- *Free Software Foundation / GNU project* <http://www.fsf.org/> <http://www.gnu.org/>
- *Open Knowledge Foundation* <http://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 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 labeled 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 http://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* <http://www.opendatacommons.org>
- *Public Domain Dedication & Licence* <http://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 university 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* <http://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 <http://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 <http://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.

APPENDICES

21 Records Management: A guide for web masters

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."

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 University 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 University 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 archives 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, 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 University 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 - eg PDF files 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 University 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 University 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 University Records Management programme. This may entail a certain amount of interpretation as well as co-operation. University 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 University 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."

Data Protection and Freedom of Information can be two very strong drivers for records management.

A University 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 a University 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."

> "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."

> **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 - ie the permanent preservation - of those records which have permanent value to the University, 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."

All quotations from the **University of Edinburgh Records Management Policy Framework**. See http://www.recordsmanagement.ed.ac.uk/InfoStaff/RMstaff/RM_framework.htm.

22 CASE STUDIES AND USE CASES

Use Case: Home page history

Scenario

Suppose 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.

Questions

- How has your institutional home page changed over time? Have you kept records of the changes and the decisions which were made?
- If the above scenario took place in your institution, do you feel you would be able to deliver a solution?

Solutions / options

- The Internet Archive may have snapshots of your home page from 1996. In order to illustrate how an institution's home page may change over a period of over 11 years, the Internet Archive's WayBack Machine was used to view the first occurrence of the University of Bath home page in every year from 1997 until 2007. See a visualisation of the changes at http://www.ukoln.ac.uk/web-focus/experiments/experiment-20080612/bath-web-tour.html
- Building a compiled history is another approach. Details of 14 year's history for the University of Virginia's Web site from 1994-2008 can be seen at http://www.virginia.edu/virginia2008/archive/index.html. The page provides details of the Web usage statistics in the early years, with screen images shown of major changes to the home page from 1997 (unfortunately no screen images are available for the first three years of the service).

Use case: Online prospectus

Scenario

The University currently provides prospective students with a printed copy of their prospectus, and 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.

Supposing the Web team would like to move towards creating an online driven model prospectus, but are unsure of how to proceed.

Issues

- What challenges will an online prospectus present? Consider both the technical and cultural challenges. Who has responsibility for such a project?
- Which of the challenges have preservation implications? What are the implications?
- What potential solutions are there for the preservation implications?

Solutions / options

- 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 reckoned as part of the University publication programme; 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 are accessible, captureable and preservable.
- Students will be making decisions about their academic career based on what they find in this prospectus. It's important for the University 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.

Use case: Persistent domains

Scenario

A project team in the University has purchased a .org domain name, outside of the main University domain, 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 has failed to renew the domain name, and it has now been purchased by a third party. This third party is now requesting a significant fee to repossess the domain name.

Issues

If your resources are held on the main institutional Web site, particularly one with a .ac.uk domain name, then your domain is unlikely to disappear - and if it does, then this will probably be a result of your institution being shut down.

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 (JA.NET, the organisation responsible for managing .ac.uk domains, does not sell off its domains to the highest bidder). There may be legitimate reasons for this - the domain may be used for an EU-funded project, for example, or the domain may be used to support a cross-sectoral service for which it is not possible to obtain a .ac.uk domain.

Solutions / options

- 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.

Use case: Student blogs

Scenario

Your University offers a large-scale student blogging service. One enthusiastic user wants to migrate their blog after they leave the University. But your institution systematically deletes files and accounts held by students once they graduate.

How should the institution respond if students wish to maintain their blog content, embedded resources, comments, etc?

Issues

How should an institution go about providing a blogging service for its students? The traditional approach which has been taken to the provision of an IT service for members of an institution has been to evaluate the range of products and select a solution which satisfies the user requirements, taking into account the resource and support implications.

In a Web 2.0 environment, however, other options become available. Rather than installing software locally, services which are available on the network can be used - and blogging services such as Wordpress and Blogger are very popular blog hosting services.

What are the preservation aspects associated with the provision of a student blogging service? One might feel that the locally-installed application must be preferable, since management of the software and data is under the control of the institution. But what happens when students leave the institution? The normal policy in many institutions has been to delete student accounts and their data shortly after they leave. But is this desirable from the student's perspective? And what if they wish their data - their blog posts - to still be available after they leave the institution?

Rights issues

> It would make great sense to archive conversations that take place on blogs, and recognise the role that this medium has within education. Do you have permission to do so? By their very nature blogs are a communal activity, even if one author does provide the bulk of the content. This collaboration – both a strength and a weakness - makes it very difficult for the 'owner' of a blog to grant permission to archive. Often blogs contain third party material, most commonly images, video or chunks of text copied from other authors. All blog 'comments' are of course third party content. Again, the blog owner probably has no rights in the material and cannot properly grant permission.
>
> We could take a risk management approach. What is the likelihood that a large multinational will sue because images it has rights over have been misappropriated? The likelihood may be large or small, but we are well aware that when third party material is included in a blog, 'innocence' remains no defence.
>
> We could remove content that we deemed to be the copyright of someone else. But why would we deliberately archive only a part of a blog? Removing text, images, video or audio would be time consuming and problematic – what if we 'missed' a bit? Or what if copyrighted material were added in the future? We could create a 'dark archive' with no public access, but who would thank us?
>
> In any event we would still need to contact the blog owner to seek their permission to archive. What if they 'gave' permission, not understanding the rights implications and then feel aggrieved that their site had not been archived due to its third party content?
>
> Ideally blog creators would work under a Creative Commons licence, and everyone would respect copyright. Maybe. But is this really the only option?
>
> Dave Thompson, adapted from his post dated 08/01/2008 to the Digital Curation Centre forum. See http://forum.dcc.ac.uk/viewtopic.php?t=237

Options / solutions

- Decide that the issue is predominantly 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: Migration of a blog

The e-learning team at the University of Bath was one of the early adopters of blog technologies to provide a forum for reflecting on e-learning in a Web context. The blog was set up by Derek Morrison when he was head of the e-learning unit. Derek had an interest in exploring the potential of new technologies, with one example of this being the series of podcast interviews he recording and made available on the blog back in 2005.

The name of the e-learning team's blog was Auricle, which has an advantage of being a very Google-friendly name, and a Google search for "Auricle Bath" finds links to the blog itself and various pages which refer to the blog.

The trigger for this case study was when it seemed that the blog no longer existed - following a link to the blog's home page yielded a 404 error message:

The web address http://www.bath.ac.uk/dacs/cdntl/pMachine/morriblog.php was not found. It may have moved, or it may no longer be available.

It was highly regrettably 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 the student's learning experiences was now no longer available. The University of Bath could have had some legitimate concerns about this loss of its intellectual endeavours and the role that the University had in being one of the early adopters of blogs by an e-learning team.

Why Did The Blog Disappear?

The URL for the Auricle blog provides an indication of some of the reasons for the disappearance of the blog: dacs refers to the Division of Access and Continuing Studies and cdntl to the Centre for the Development of New Technologies in Learning - but neither of these departments still exists. Following staff departures and organisational changes, support for learning at the University of now provided by the Learning and Teaching Enhancement Office (LTEO) with the e-Learning Team having responsibility for managing and supporting e-learning developments.

In addition to these organisational changes, the pMachine part of the blog's domain name refers to the pMachine blog engine and morriblog clearly refers to Derek Morrison, who left the University a number of years ago to support the HE Academy's Pathfinder programme.

Following such changes and the influx of a large number of new staff in the e-Learning Team, it seemed that the Auricle blog got lost somewhere along the way.

Could any of the resources be retrieved?

The study asked if it would be possible to retrieve any of the blog posts and related resources. Likewise any details about the blog, such as when it was launched, the number of posts published during its lifetime, how popular it was and, perhaps, the impact that the blog may have had.

Since the blog was public, as opposed to a blog which was restricted to members of the University of Bath, the contents of the blog have been indexed by Google. And using a combination of search terms, such as "Auricle Bath", it is also possible to discover Web resources which cite the Auricle blog. This helped to find a blog post on Stephen Downes's blog on The Weblog as the Model for a New Type of Virtual Learning Environment? in which Stephen (a high profile Canadian e-learning guru) clearly acknowledged the importance of Derek Morrison's views on the potential of the blog as providing "the basis for a distributed, not centralised, information and learning object system":

The author of Auricle nails it. "In the weblog, however, the announcements, articles, stories are the raison d'etre' so much so that, not satisfied to present articles from one source, the weblog has the temerity, due to the adoption of the RSS standard, to receive syndicated stories from other sources and, in turn, offer it's own portfolio of articles for use by others. For example, a blog supporting a programme or module could be the vehicle by which faculty post date and time-stamped short articles relevant to the course but which also link to related, but distributed, learning resources which are presented via RSS feeds. Such feeds can be static or dynamic so that updated RSS formatted information will be reflected in whatever application is displaying it, e.g. a la Auricle's RSS Dispenser. Here then is the basis for a distributed, not centralised, information and learning object system." Derek Morrison, Auricle, February 27, 2004.

The date of Stephen's post (27 February 2004) indicates that the Auricle blog was available in early 2004.

With some further use of Google, it was discovered that the Auricle podcast resources are still available on the University of Bath Web site. The RSS file also contains the publication dates, which show that the podcasts were published during 2005. We seem to have unearthed some further information about the Auricle blog.

Rediscovering the blog

A Google search for "Auricle Morrison" revealed that the Auricle blog is alive and well. It is now hosted at http://www.auricle.org/auriclewp/ (an improvement on the original URI). As well as providing access to the original posts (although with a new look-and-feel, as the blog is now based on the Wordpress blog software) the blog is still active, with Derek using the blog to support his Pathfinder work at the HE Academy. As Chris Rusbridge pointed out, phrases like "long term accessibility" or "usability over time" are more meaningful in this context than the process-oriented phrase "digital preservation". It is also an example of how the Auricle blog has been preserved by continuing to still be used and accessible to its user community.

The Lessons

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

We should be aware of the dangers of associating services with departmental names and specific technologies. This has been well documented, including Tim Berners-Lee's article on "Cool URIs Don't Change!" - although this is clearly easy to say, but more difficult to implement in practice.

There is also a need for departments to audit their networked services and to document their policies regarding the sustainability of such services. And such documented policies should be examined when departments change their names or there are significant changes in personnel.

And this case study provides an interesting example of a service which has been driven by an individual - Derek Morrison. As Derek clearly felt ownership to the Auricle blog, he was motivated to migrate the content of the blog to a new platform and, at a later date, to continue to contribute to the blog, although not as frequently as previously. This probably saves the e-Learning Team at Bath from having to retrieve backup copies of the blog posts and provide an archived copy of the resource. But who owns the blog? And what would have happened if there had been an ownership dispute over the blog and the name of the blog? These are questions which will be relevant to many academics and support who make use of blogs to support their professional activities.

While the original Auricle represented a major investment in personal time, reflection, research, and overall effort, the owner felt this was worthwhile because it provided a vehicle for reflecting on what are an increasing number of difficult issues relating to the use (and abuse) of technology in Higher Education. The University of Bath generously provided the technical platform for developing 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 university, Auricle was inevitably caught up in annual online account housekeeping.

The owner was heavily influenced by the ethos of the LOCKSS Programme, i.e. Lots of Copies Keep Stuff Safe (http://www.lockss.org/lockss/) and rescued what could be rescued before the virtual "Grim Reaper" arrived. The main core of the content authored at Bath still lives in its new home at auricle.org, although there is further work to do to rejuventate apparently dead hyperlinks to resources that were originally hosted at the University of Bath.

The core 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. Parts of the sector seem to have valued it as a resource, with Auricle sometimes making a small contribution to someone's course or module. But, in common with all other HEIs, once personnel leave an institution such blogs, wikis etc associated with them are not analysed for the potential value of their content but treated more as an email account, i.e. something to be deleted. This is unfortunate, because it represents a lost opportunity for ongoing collaboration, discourages future investment by blog, wiki authors etc 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).

However, it's also necessary to recognise that supporting such resources do represent an ongoing cost for an institution (albeit a relatively tiny one in University financial terms).

Use case: Migration of a wiki

Scenario

Wetpaint wiki is just one of the many enticing, powerful, quick-fix web apps that have sprung up around Web 2.0 and Social Networking. Wikis have grown up a lot since the first WikiWikiWeb, and now are at the online heart of many educational projects at all levels, from classroom, to research and publishing.

Wetpaint's wiki feature could be used as a collaborative space for a project's workshop feedback. Once all the input for project outputs has been collated, in a few weeks it would probably be no loss to delete the wiki, or just set it adrift among all the other jettisoned flotsam in cyberspace.

But what's often given less serious consideration, in the excitement of using a third-party provider of wikis, blogs, Ning, etc., to get a collaborative hypertext project off the ground so quickly and easily - and without having to go cap or cheque in-hand to whoever guards your web space - is this key preservation issue: what happens when you want to get your painstakingly intricate web of hyperlinked pages out?

There are many good reasons why you might want to do this: you might want to migrate to another wiki system or CMS, as the shape and nature of your content evolves; or put it on a permanent, persistent footing by moving it into your own domain; you might simply want to back it up or take a snapshot; or you might want to pull out information for publication in a different form. When you had one or two pages, it might have seemed trivial; but what if you now have hundreds?

Unfortunately, just as exporting the information is often a secondary consideration for wiki content creators, so it also is for the wiki farm systems. The Wetpaint Wiki discussion boards indicate 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 leaves a lot to be desired.

Options / solutions

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 your online wiki is not possible. But, for what it's worth, you have at least salvaged some sort of raw content, that might be transformed into something like the wiki it came from, if hit with a bit of Perl script or similar.

Wikidot is another impressively-specced, free "wiki farm". Wikidot's backup option will deliver you a zip file containing each wiki page as a separate text file, containing your 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 non-standard wiki-markup into standard HTML.

A third approach can be seen with a self-hosted copy of Mediawiki. Here you can select one or more pages by name, 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 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.

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; or just in case a day should come when Wetpaint.com Inc., and the rest, no longer exist.

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.)

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.

Use case: Preservation and Slideshare

Slideshare is a popular externally hosted Web 2.0 service for providing access to copies of presentations. There is evidence to demonstrate its impact in maximising awareness of presentations, this might include both awareness of research activities, and marketing activities.

Issues

Are there risks associated in making use of a third party service in this way? What will happen if, for example, the Slideshare's business model is flawed and the company goes bankrupt? Rather than making use of a Web 2.0 service shouldn't we be providing Slideshare's functionality 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 impact and popularity which Slideshare has.

The challenge is to assess possible risks and to explore mechanisms for managing such risks. One 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.

Options / solutions

- Store a managed master copy of the slides on your own website, and ensure that links to this resource are provided on Slideshare. The URL could be included on the title slide and in the accompanying metadata. In addition the URL could also be included in the footer of the hard copy printouts.
- 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.

Comments

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.

Use Case: Institutional Use of Twitter

Scenario

The University has 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 University as a individual productivity and social tool. However, one Department in the University 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

Twitter is a microblogging application which can be used to create a brief (up to 140 characters) blog post. Although initially used by individuals to summarise feelings and thoughts, the ways in which the service is being used has evolved. It has the capacity to be used as a general chat facility, and so has some parallels with an instant messaging environment (with the added advantage that tweets can be delivered free-of-charge to mobile phones).

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 technologies.

Solutions / options:

This is simply a matter of making an Institutional decision about the value of these resources. Why keep Twitters?

- 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 the 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 university 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 this 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 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 by 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. And 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.

General 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

GLOSSARY

Accessibility

AJAX

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: http://www.learninfonet.ac.uk/partnerships/records-retention-he/archival-appraisal

Archiving (1)
The permanent preservation of those records which have permanent value to the University, or records which may be deemed to have historical, heritage, and research value to others. The University 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

Backup

Cardinality

Cloud computing

Content management

Content Management System (CMS)

Continuity

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.

Digital Curation Centre

Disaster recovery

Domain harvesting

Domain persistence

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.

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

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

Information lifecycle See Lifecycle management

Information vs Record

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

Internet Archive
Based in San Francisco, the Internet Archive is an open, online archive of digital material. It operates the [[Wayback Machine]], a [[remote harvesting]] system for mirroring websites.

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: http://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: http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/

Permalink

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:
http://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 University and preserving an adequate historical record.
Source:
http://www.recordsmanagement.ed.ac.uk/InfoStaff/RMstaff/RM_framework.htm

Remote harvesting
A method of [[web archiving]] which takes snapshots of websites by following all the links in each web page.

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

Risk analysis

Selection

UK Web-Archive 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.
http://www.webarchive.org.uk/

URI
See http://www.w3.org/TR/uri-clarification/

URI persistence

URL

URN

Wayback Machine
The [[remote harvesting]] [[web crawler]] operated by the [[Internet Archive]].

Web archiving

Web content management system

Web management

Web resource

Website