

Introduction

Many of JISC's programmes and initiatives support repositories and the retention, reuse and sharing of digital assets. Higher and further education institution websites might also count as repositories and may contain evidence of institutional activity, unrecorded elsewhere, that constitute an important record of digital initiatives over the last 15 years. Yet these websites are not being systematically preserved. There is a need to harmonise institutional web and digital preservation practices to prevent valuable collections of research and other materials being at risk.

There are two issues facing institutions in 2009:

- They may not be doing enough to preserve the valuable resources that have been created up to this point
- New resources are being created, especially through emerging Web 2.0 applications, that we don't even know how to capture yet, let alone preserve

What are web resources?

Institutional websites contain valuable records, publications, prospectuses, research and teaching outputs, projects and evidence of other activities. But beyond these, there are also assessment systems, centralised administrative systems, online libraries, digital collections, Virtual Learning Environments (VLEs), collections of e-learning objects, delivery systems, blogs and wikis, all potentially containing or generating unique content.

Preserving online resources is important

There are institutional benefits to preserving web resources. Considerable time and money has been invested in the creation of digital outputs and content, and in their storage and maintenance. Although there are costs

associated with launching a web preservation programme, it's also money wasted if resources aren't preserved.

Institutions have responsibilities to: students and staff, who may make serious choices about their academic careers based on website information; and researchers and scholars, who may need to use the university's resources in the future. Ensuring that the wider community has long-term access to research materials will be broadly beneficial.

There is also the matter of protecting institutions. Many risks are faced by organisations that choose to ignore web preservation. An institutional record may be required for the checking of strategic, legal, financial and contractual information, or simply for the day to day continued efficient running of the organisation. But there are external threats too. These include: data loss; loss of records and loss of resources; a failure to be information compliant (through not meeting Freedom of Information requests); risks of breaching copyright; and even risk of litigation from students or the public. Consider if a legal action were brought against an institution as a result of certain information that was exposed two years ago, and has since been taken down. Could the institution provide evidence, such as an audit trail, in court?

Websites need to be protected from careless editing, changing, deletion or removal – whether it's by a malevolent hacker or a well-intentioned member of staff. This includes everything from the removal of valuable research materials which appear to have expired because the pages haven't been visited or updated in two years, to the removal of an entire wiki full of useful content, occasioned by the automatic deletion of the account of a retiring staff member.

Who should be doing it?

JISC has identified a concern that IT communities within institutions, including webmasters, are not sharing practices with the digital preservation world. Managed

Preservation of Web Resources

March 2009

backup is not the same as preservation, which is about providing permanent access to a resource.

All institutional information experts need to become more pro-active and contribute their skills, working to existing models which can be adapted, and sharing the responsibilities for web preservation. Taking the collaborative approach means bringing together the interests and skill sets of information managers, webmasters, IT specialists, systems administrators, archivists and records managers.

It is no longer enough to leave all this in the hands of the webmaster, who may not even understand what the information lifecycle is. Software and system vendors have borrowed the lifecycle concept from records management to describe the process by which 'information' (data would perhaps be a better word) moves through a storage hierarchy depending primarily on the frequency with which it is used – online when in constant use, near-line when used occasionally, off-line when not used at all. But the final destruction/disposal phase still needs to be based, as in the records lifecycle, on assessed business need rather than on frequency of use.

People who create and use the resources should also have a voice. It is important to identify the correct stakeholders, finding out why an institution has these web resources, what use is being made of them and by whom.

Embedding behaviour and practice in policies helps with preservation. Extant policies and procedures (and they need not be IT or web-based) can be harnessed to support a web-archiving programme. JISC's 2008 study on Digital Preservation Policies and their implementation shows that web preservation can support the institutional mission. Archived web resources can support strategies in research, learning and teaching, information, libraries and records management. Harnessing web content in service of an institution's aims can become a major business driver.

How can it be done?

The process needs to be selective, and web resources need to be managed before they can be preserved. An important stage of any preservation activity is capturing it in the first place. An agreed institutional collection policy is essential, as is deciding which aspects of the materials need to be captured. It may only be important to capture

the content of web pages but equally, their appearance and the way they behave may also be significant.

There are three possible stages at which to perform capture:

- Within the authoring system or server
- Capturing at the browser
- Harvesting content with a crawler

Web 2.0

Many institutions will have been affected by the Web 2.0 phenomenon – that is, the recent explosion of interactive and personalised web services and applications, from blogs and wikis to online services such as Flickr, Twitter and SlideShare. Applications producing web resources include social bookmarking, media sharing, social networking, collaborative editing and syndication technologies, and even Instant Messaging.

All these applications are known to be in use in higher and further education in 2009 and their use often results in diverse collections of resources that are not kept on institutional servers and are outside the main domain. All Web 2.0 applications thus share the same problem. The issues are ones of ownership and responsibility. In an academic context these applications rely on the individual to create and manage their own resources. The academic, staff member or student creates and manages his or her own external accounts in Flickr, SlideShare or Wordpress, but they are not institutional accounts.

It is thus possible with Web 2.0 applications to conduct a significant amount of institutional business outside any known institution network. This situation could be managed by a central policy decision regulating the approved use of such applications, who is authorised to use them, for what purposes and under what circumstances.

There are dangers in relying too heavily on sites like SlideShare and YouTube to manage resources such as PowerPoint slideshows, moving images or audio presentations. Such sites can certainly be used for effective dissemination and delivery, but they should not be trusted to preserve or back up resources. It is essential for institutions to retain 'master copies' of such resources, ideally working in line with existing digital archive programmes.



The international web archiving picture

The **Internet Archive** is the most conspicuous initiative that crawls the web and takes snapshots of web pages. But it cannot realistically guarantee to capture all of an institution's web-based assets, nor preserve all of its scholarly material in perpetuity. The quality of the capture may not be ideal, with missing images, style sheets or even content. There are also problems with depth of capture, database-driven sites and dynamic content. The Internet Archive collection policy leads to gaps in temporal continuity; there can be large gaps between capture dates.

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. It also offers ways for anyone to submit a website to be included in the Archive.

The Internet Archive lacks an explicit preservation principle or policy, and has no real mandate to capture websites beyond a societal desire to see it happening and

to share the results with the public. This lack of policy may cause severe problems for institutions; it is unlikely that the Archive will cover everything an institution needs to do within its remit.

The **UK Web Archiving Consortium** (UKWAC) has been gathering and curating websites since 2004. Among its members are the national libraries, the National Archives, the Wellcome Trust and JISC. To date, UKWAC's approach has been very selective, and determined by written selection policies which are in some ways quite narrow. JISC, for example, have made it their remit to collect websites of projects that they have funded. That remit has expanded to include the websites of certain central and regional higher and further education organisations, but to date no complete snapshots of institutional websites in the UK have been collected.

It is possible to nominate a website for capture with UKWAC, thus resulting in a snapshot at a certain date and time. However, certain resources will be beyond the reach of the Heritrix crawler (including databases, secure and passworded pages and hidden links). If the website depends heavily on server-side architecture, then remote capture may fail.

Preservation of Web Resources

March 2009

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 the preservation of all web resources.

The **European Archive** describes itself as a digital library of cultural artefacts in digital form. This non-profit foundation wants to achieve 'access to knowledge' and is especially interested in partnerships with libraries and other institutions. By partnering with the Internet Archive, the European Archive hopes to lay down 'the foundation of a global Web archive based in Europe'. It includes web collections alongside digital recordings and moving images. Selected UK government websites are captured and hosted there, in weekly and six-monthly harvests. The European Archive Foundation offers free storage and bandwidth for a collection, although institutions presumably still have to undertake the harvesting themselves.

The **International Internet Preservation Consortium** (IIPC) won't help to harvest an institution's website, but it is an internationally recognised body of excellence for website preservation. The mission of the IIPC is to acquire, preserve and make accessible knowledge and information from the internet for future generations everywhere, promoting global exchange and international relations.

In addition there are many international initiatives, mostly library-based and sponsored at a national level, which are aiming to complete selective web collections, often based on the aim of archiving the entire 'national' domain. These include MINERVA (Library of Congress) and PANDORA (National Library of Australia).

Related JISC initiatives

The JISC Preservation of Web Resources (PoWR) project issued a handbook in 2008 describing the institutional benefits of preserving web resources. It outlines the tools and processes that are needed and how a records management approach may be appropriate.

The 2004 web-archiving feasibility study correctly identified the urgent need to start gathering JISC-sponsored project material from universities, much of which had already vanished. Although the gap is partially addressed by the resulting Archiving JISC Websites initiative, institutions still need to inculcate some sense of responsibility for preserving their own resources.

Toolkits such as: the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA); the Data Audit Framework (DAF); and Assessing Institutional Digital Assets (AIDA); will help institutions assess their capability for managing and preserving digital resources including web resources and research outputs.

This briefing paper was written by Ed Pinsent, ULCC.
Alternative formats of this briefing paper can be found at:
www.jisc.ac.uk/publications

Further Information and Resources

JISC Digital Preservation Policies Study
www.jisc.ac.uk/publications/publications/jiscpolicyfinalreport

JISC PoWR Project
www.jisc.ac.uk/whatwedo/programmes/preservation/2008powr

Internet Archive
www.archive.org

UK Web Archiving Consortium
www.webarchive.org.uk

The European Archive
www.europarchive.org

International Internet Preservation Consortium
<http://netpreserve.org>