scripted

Volume 6, Issue 3, August 2009

Sustaining On-line Research Resources

Arshad Khan[*], *David Martin*[°] *and Jane Seale*[†]

Abstract

We have seen enormous growth in both the usage and creation of web resources in the last decade. Significant funds have been devoted to the creation of high quality academic web resources by both public and private sector organisation s. These have already benefited a large community of web users, researchers, students and teachers. In order to ensure continued access to this wealth of on-line resources, the web preservation community has already started making efforts to formulate and execute strategies aimed at collecting, processing and preserving today's web resources so that they can be accessed with tomorrow's technologies. This article reviews such initiatives, drawing a comparison between current web preservation practices and the ESRC-funded ReStore project, a sustainable web resources repository. Detailed consideration is given to issues including authorship of web page content (intellectual property rights, copyright), metadata generation and preservation, the selection of web resources, and accessibility to hidden pages on a web server. We present a possible short-medium term preservation model aimed at sustaining on-line research method resources developed as part of ReStore. The article considers the potential for evolution from the current rather disparate web preservation approaches to standardised "develop with a view to preservation" practices among web resource creators and the web preservation community.

DOI: 10.2966/scrip.060309.616

---

[*] Research Assistant (ReStore Project), ESRC National Centre for Research Methods, University of Southampton.

[°] Professor of Geography, School of Geography, University of Southampton.

[†] Senior Lecturer, School of Education, University of Southampton.

1. Introduction

The on-line information revolution has provided numerous opportunities to express, share, comment and communicate ideas more quickly and easily. The web has evolved into an enormously rich but largely unstructured source of data and information. The phenomenal growth of the web reflects not just widening access and developing technologies, but also the increased availability of really useful content. In the last decade in the United Kingdom we have seen significant investment by academic research councils and other funding bodies in projects involving the creation of web resources. Such projects have resulted in the creation of web-based knowledge domains that can be invaluable to the research community.

Unfortunately, deterioration of these web resources often begins immediately after funding ceases and teams disperse, just at the point at which the resource becomes most valuable to researchers. The content of the resource becomes outdated, its links break down and eventually it ceases to present appropriately on users' web browsers. The digital formats conventionally used within web resources change over time and some fall into disuse.[1] Live sites gradually change, by implementing various software upgrades, changing hardware platforms and perhaps even adopting new protocols. These are some of the challenges that led to the establishment of the ReStore project (www.restore.ac.uk), funded by the Economic and Social Research Council (ESRC).

In this article, we examine the ReStore project as a part of comparative analysis of digital repository initiatives. We will set out our approach to sustaining on-line resources through ReStore, and assess the advantages and disadvantages of a variety of other approaches prevalent in the web preservation community. We will consider the role of harvesting, metadata generation, deployment and exposure of web resources and their respective metadata in the improvement of cross-platform web searching and metadata harvesting. Drawing on ReStore experience, we will highlight issues relating to intellectual property rights (IPR), copyright and third party contributions prior to sustainable web preservation.

In the remainder of this article we identify the elements of major interest and then consider the purpose of web preservation. Section 4 outlines current approaches to preservation of web resources and section 5 explains the particular nature of the ReStore project. Section 6 and 7 contrast ReStore with the Open Archival Information System (OAIS)[2] reference model, emphasising the need for long term preservation. We then consider the most appropriate time and means by which to ReStore on-line resources and outline our approach to the selection of resource sites. IPR, digital repository networks and metadata issues are each reviewed. After identifying some limitations in the current approaches to web preservation, the paper concludes with a consideration of future directions and the specific role of ReStore.

---

[1] JA Smith, "Integrating preservation functions into the web server" (2008) available at http://www.joanasmith.com/node/47 (accessed 18 Nov 09).

[2] Digital Preservation Europe, "DPE: Digital Preservation Europe" (2008) available at http://www.digitalpreservationeurope.eu/video-training/prague-2008/?media=3 (accessed 18 Nov 09). OAIS defines a common framework in order to analyse and describe concepts and terminology for digital archives and repositories.

2. Web resources, web sites and preservation

Diversity of content that is largely unstructured and connected to insufficient or no metadata, poses a mammoth challenge to those involved with web resource preservation. The range of file types is immense and heterogeneity of formats makes sustaining a typical web resource a serious challenge to archivists, repository managers and researchers. After more than 10 years of web evolution,[3] however, HTTP[4], MIME[5] and HTML[6] form the foundation of the web, which is optimised for the "here and now". The publication of new web pages by someone having only basic Internet knowledge has never been easier.

As a result, the locus of knowledge is shifting from traditional libraries and archives to digital web-based resources, where growing technology poses a great challenge. In order to preserve, sustain and disseminate knowledge, some type of sustainable system for the preservation of digital web resources is unavoidable. We deliberately refer to our approach as "active preservation," that is different from merely capturing the content of a web page or web resources through the use of snapshot tools and crawling software. Active preservation might therefore be considered as one of a range of sustainable preservation practices.

Before proceeding, it is necessary to define various terms that will be used here. "The Web is designed as a network of more or less static addressable objects, basically files and documents, linked using Uniform Resource Locators (URLs)[7]". A resource is implicitly defined as something that can be identified, and identification serves the two distinct purposes of naming and addressing, the latter being dependent only on a protocol e.g. HTTP [8]. With this definition in mind, a web resource is a network of static and dynamic addressable objects, each having a unique URL and interlinked with other URLs through HTTP.

---

[3] JA Smith and ML Nelson, "Creating Best effort Preservation Metadata for Web Resources At Time of Dissemination" (2007) available at http://www.cs.odu.edu/~mln/pubs/jcdl07/jcdl07-best-effort-metadata.pdf (accessed 18 Nov 09).

[4] W3C, "Hypertext Transfer Protocol" (1999) available at http://www.w3.org/Protocols/rfc2616/rfc2616.html (accessed 18 Nov 09). Hypertext Transfer Protocol is an application level protocol for distributed, collaborative, hypermedia information systems.

[5] Multipurpose Internet Mail Extensions, a specification for formatting non-ASCII messages so that they can be sent over the Internet. Many email clients and browsers support MIME which enables them to send/receive graphics/audio/videos files via the Internet and display output files in that are not in HTML format.

[6] W3Schools.com, "Introduction to HTML" available at http://www.w3schools.com/HTML/html_intro.asp (accessed 18 Nov 09). Hypertext Mark-up Language is a language for describing web pages.

[7] Uniform Resource Locator is a pointer to a "resoruce" on the World Wide Web (WWW). available at http://download.java.net/jdk7/docs/api/java/net/package-use.html (accessed 18 Nov 09). URI (Uniform Resource Identifier) consists of a string of characters used to identify or name a resource on the Internet.

[8] See note 3 above.

It is normally taken for granted that a "web site" and "web resource" refer to the same thing but according to the World Wide Web Consortium (W3C)[9] a web site is a

> *collection of interlinked web pages including a host page, residing at the same network location. Interlinked is understood to mean that any of a web site's constituent web pages can be accessed by following a sequence of references beginning at the site's host page; spanning zero, one or more web pages located at the same site; and ending at the web page in question.*[10]

A web site may also be defined as a collection of resources having a common domain name, accessible via the Internet.[11] The distinct difference between a web site and a web resource is that a resource is necessarily interlinked on the same network location (the host page of the site) and can be accessed using any implemented version of HTTP, provided that each resource is distinctly identified by a URL.

3. What is preservation all about?

Preservation of web resources has become something of a buzz phrase, reflecting the importance of thinking about the future. The objective of the ReStore project is not merely to preserve a web resource as a static record, but to focus on actively sustaining selected web resources in order to extend their utility beyond the duration of the projects that led to their development. In this discussion we therefore use the phrase "actively sustaining" rather than the terms "preservation" or "preserving".

In the long term, even physical materials suffer some degradation: this is well recognised in the deterioration of archived magnetic tape or film media. The equivalent process of degradation for digital materials is typically caused by format obsolescence, due to changes in software applications technology, often a rapid process in comparison with the degradation of physical materials. Such degradation applies equally to web resources, which start to decay due to lack of maintenance or arrival of newer web tools and technologies. Digital degradation can be ameliorated by specialised technical processes such as format migration.[12]

When funding for a web project ceases and primary project investigators and developers disperse, as noted earlier, the resource may be left in the hands of a third party data storage centre. This centre provides merely for the continued existence of the resource, making sure that the number of files and folders remain the same ("enumeration"), but with no intention to attend to missing and/or broken links, new software updates or other maintenance activities ("representation"). Until recently, preservation measures - archivists would digitally label each item and create a unique record for easy retrieval – were applied only to things such as digital libraries,

---

[9] B Lavoie and H Frynstyk, "Web Characterization Terminology & Definitions Sheet" (1999) available at http://www.w3.org/1999/05/WCA-terms/ (accessed 18 Nov 09).

[10] See note 1 above.

[11] *Ibid*.

[12] S Hitchcock, T Brody, JMN Hey and L Carr, "Digital Preservation Service Provider Models for Institutional repositories" (2007) available at http://www.dlib.org/dlib/may07/hitchcock/05hitchcock.html (accessed 18 Nov 09).

research papers, student theses and academic journals. It is timely therefore to review institutional repositories (IR), digital repositories and sustainable web resource repositories.

Research publications and student theses are often preserved in an IR maintained by an educational institution. IRs are digital collections that capture and preserve the intellectual output of communities.[13] An IR thus helps academic institutions to better manage, report and promote the outputs of their research, with benefits for the researchers of both today and tomorrow.

In general terms, an IR has many features of a digital repository. A digital repository provides a setting in which digital content, including web content, is stored, and can be searched and retrieved for later use.[14] A web resource repository shares many - but not all - features of a typical digital repository, such as the storage and accessibility of content through the use of common standards, or more recently, protocols. A digital web repository is however, distinguished, from other digital repositories on the basis of the sustainability of its content. The repository itself and the sustainability of content are therefore addressed separately below.

4. Current approaches to the preservation of web resources

Preserving and sustaining web resources in a repository is now frequently discussed in workshops, seminars and conferences. When we talk about preservation of digital resources such as digital libraries, web resources, research papers etc, we associate them with digital, institutional and on-line repositories. A repository is a place, room or container where something is deposited or stored.[15] The terms "repository" and "preservation" are used interdependently, creating the impression that preservation would be incomplete without a repository.

Various initiatives such as the National Archives,[16] UKDA Store[17] and the UK Web Archive[18] are actively archiving and preserving web resources that represent topics of UK cultural, societal, religious, political and scientific significance. The preservation techniques being used are mainly snapshot-based.

Snapshot-based preservation involves archiving snapshots of a web page or set of pages to a central location where they can be accessed by users through a URL or unique identifier. Currently, these efforts involve crawling (collecting) a web resource site (or sections of it) using crawler[19] software. The software runs through each

---

[13] R Crow, "The case for institutional repositories: A SPARC Position Paper" available at http://www.arl.org/sparc/bm~doc/ir_final_release_102.pdf (accessed 18 Nov 09).

[14] JISC, "Digital Repositories, Helping Universities and Colleges" (2005) available at http://www.jisc.ac.uk/uploaded_documents/HE_repositories_briefing_paper_2005.pdf (accessed 18 Nov 09).

[15] Merriam-Webster "repository-Definition from the Merriam-Webster On-line Dictionary" (Nov 09) available at http://www.merriam-webster.com/dictionary/repository (accessed 18 Nov 09).

[16] "The National Archives" available at http://www.nationalarchives.gov.uk/ (accessed 18 Nov 09).

[17] "UKDA-Store" available at http://store.data-archive.ac.uk/store/ (accessed 18 Nov 09).

[18] "UK Web Archive" available at http://www.webarchive.org.uk/ukwa/ (accessed 18 Nov 09).

[19] Crawler software agent generally refers to software which are used to access particular web sites without the users realization and collect or harvest web pages of the site for storage purposes. UK Web

section of the site and stores the contents, appearance and layout of a web page in a remote data store using a commonly used format called ARC or more recently WARC.[20] Once collected, the pages of a particular web resource are stored in a digital repository that is accessible to users through a web interface, allowing each individual item to be accessed through a unique identifier or URL. This method of preservation does not however ensure that every page of the site has been collected, processed and archived. This enumeration problem is one of the serious flaws of harvesting web resources by crawler software. A further problem with snapshot-based preservation is that it may not be a sustainable solution to a decaying web resource. Commercial web crawlers are estimated to index only about 16% of the total surface web,[21] and a small fraction of the "deep web"[22] or "hidden web" that is estimated to be up to 550 times as large as the surface web.[23]

Another initiative, called the Internet Archive,[24] crawls the web and takes snapshots of the web pages of an institution, but cannot guarantee to capture all of its web-based assets, nor preserve its scholarly material in perpetuity. There are also problems with depth of capture that are particularly relevant to database-driven sites and dynamic content.[25] Similarly the UK Web Archive Consortium[26] preserves web resources by capturing web pages using crawling technology (Heritrix)[27] on certain dates and times.

The common factor in snapshot-based preservation approaches is that they are still unable to reach hidden pages on web resources that are either password protected, generated "on the fly" by a web server,[28] or included in another file "server-side".[29] "On the fly" creation of web pages (commonly called "dynamic web pages") involves

---

Archive available at http://www.webarchive.org.uk/ukwa/ (accessed 18 Nov 09) and Archive-IT available at http://www.archive-it.org/ (accessed 18 Nov 09) are using such technologies as part of preservation of web resources.

[20] Web Archive File Format.

[21] ML Nelson, H van de Sompel, X Liu, TL Harrison and N McFarland "mod_oai: An Apache Module for Metadata Harvesting" (2005) available at http://public.lanl.gov/herbertv/papers/ecdl-mod_oai-submitted.pdf (accessed 18 Nov 09).

[22] Deep web covers those web resources which are often hidden behind web scripts from web search engines like Google and Yahoo, MSN etc.

[23] "ARCHIVE-IT" available at http://www.archive-it.org/ (accessed 18 Nov 09).

[24] "Internet Archive" available at http://www.archive.org/ (accessed 18 Nov 09).

[25] JISC, "Preservation of Web Resources" (Mar 09) available at http://www.jisc.ac.uk/publications/documents/bpwebpreservation.aspx (accessed 18 Nov 09).

[26] UK Web Archive Consortium, "UK Web Archive" available at www.webarchive.org.uk (accessed 18 Nov 09).

[27] HERITRIX, "HERITRIX" (Oct 09) available at http://Crawler.archive.org (accessed 18 Nov 09).

[28] W3C, "Web Characterization Terminology & Definitions Sheet" (99) available at http://www.w3.org/1999/05/WCA-terms/#Server3 (accessed 18 Nov 09). A server that provides access to Web resources and which supplies Web resource manifestations to the requestor.

[29] Server-side scripting language is used almost exclusively for the web. As its name implies, its primary use is including the contents of one file into another one dynamically when the latter is served by a web server.

sending queries to and from a local or remote database, another area beyond the reach of a crawler released for archiving web resources for preservation.

5. The ReStore repository: http://www.restore.ac.uk

The ReStore project was initiated as a result of a realisation that many research-council funded resources for training and capacity-building were being lost through obsolescence and lack of maintenance after the initial funding period had ended. Research resources, particularly those funded by major research councils, generally represent much greater financial and intellectual investment than many other types of digital resource. They are also created with the specific aim of recording results or methods that are intended to be built upon and referenced by subsequent researchers. In the past, these goals were achieved primarily through conventional academic publications made available through physical libraries and archives. Our approach, that promotes actively sustaining, rather than passively preserving, web resources, distinguishes it from those reviewed so far, but it falls short of sustaining and exposing metadata for each web resource, an issue to be discussed below. It does go some way to addressing the problems relating to snapshot-based web preservation. The following sections describe the preservation process as implemented in ReStore.

A web resource site is deemed to need "ReStoration"[30] when its contents begin to become outdated, links begin to fail and when the content is not presented appropriately on users' computers. The ReStore project was launched, as part of the implementation of the idea of "Web resource ReStoration", to take care of a specific pool of on-line resources and to develop guidelines for a long term strategy. ReStore is aimed at preserving and maintaining quality on-line resources (static, dynamic, deep) created by ESRC-funded projects in research methods, and to ensure their fault free on-line presence after original funding for the project has ceased. It is primarily concerned with extending the period of maximum value rather than preserving for posterity, although the ReStoration process is also likely to make resources better suited for long-term static archiving.

ReStore, while collaborating with the original award holder and primary investigators, ensures that the ReStored resources are up to date and all links are fully functional. Fundamentally, the ReStore project aims to:

1. build a prototype service for sustaining on-line resources;
2. establish a service to sustain on-line resources in the field of research methods; and
3. lead the development of a long-term strategy for ESRC in sustaining on-line resources.

The prototype service will inevitably expose inherent problems associated with active preservation, and should help web resource creators and developers to develop a mechanism aimed at sustaining web resources from day one of their creation.

A further goal of ReStore is to raise awareness among web resource project proposers, researchers, authors, editors and contributors. Through the sustainability guidelines that are currently in preparation, we seek to promote the importance of the

[30] ReStoration is a term we use to specifically refer to restoration of web resources in ReStore repository.

sustainability of valuable web resources before creators venture to develop other ill-considered web publications.

ReStore experience so far shows that the problem of sustainability - in terms of copyright issues - is very complex as a result of the involvement of many different parties in web resource creation. The complexity is compounded, in some cases, by sub-standard technical approaches adopted at the time of resource development. Many academic resource authors who are subject experts nevertheless fail to take account of existing advice and best practice, thus impairing the future resilience of their own resources.

6. OAIS:Standard framework for web resource preservation

Having considered all of these issues, it is reasonable to ask whether there is any standard approach that could be adopted by the web preservation community to perform the task of sustainable web resource preservation. The answer potentially lies in the framework called OAIS (Open Archival Information Systems),[31] which is widely used, but does not address all the practical issues relating to web resource preservation. The OAIS model aims to facilitate a wider understanding of the requirements of long term information preservation, but does not assume or endorse any specific computing platform, system environment or database management system.

The OAIS model is based on an abstract approach to long term preservation, and is less effective regarding implementation of a specific design for repository architecture. Its importance lies in its commitment to a dual role of preservation and the provision of long term access to information, with a view to addressing the issues of technological obsolescence and future media failure. ReStore will also have to give consideration at some point how to ensure fault free access to ReStored web resources following changes in technological infrastructure, through the media, protocols or other Internet standards. Without providing all the details of OAIS, we will compare and contrast the ReStore repository with the OAIS model in order to demonstrate how it conforms to established standards aimed at long term sustainable web resource preservation.

7. Assessing ReStore's conformance to the OAIS reference model

The purpose of this review is not to assert the conformity of ReStore with the OAIS model but rather to identify loopholes and highlight limitations of the OAIS model that arise during the implementation of web resource preservation. As shown in Figure A, data (in this case web content) are ingested from a producer and handled by four parallel processes.

The "Data management" process provides services and functions for populating and maintaining descriptive information about content stored in the archive. The "Archival storage" provides service and functions for maintenance and retrieval of information packages. Both of these layers are supported in parallel by "Preservation planning" and "Administration" layers, which are largely responsible for high level

[31] See note 2 above.

preservation planning and administrative strategies focusing on processes involving IPR agreements, hardware and software platforms for a repository.

Since the OAIS reference model merely presents a framework for long term preservation and does not specify any implementation details, individual repository development groups could break out functionalities differently as per their budget, requirements and technical environment during the formation of their repository platform.

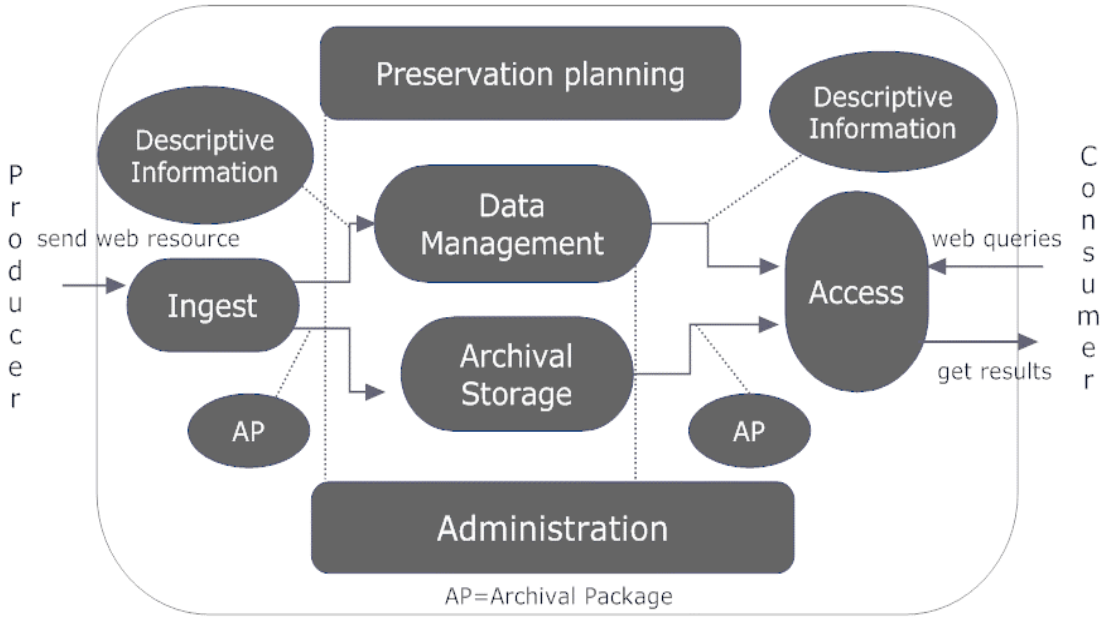

Figure A[32]

Figure B shows the various entities of the OAIS model modified somewhat to reflect the abstract design of the ReStore architecture. The ReStore repository model is aimed at sustainable web resource preservation, and thus incorporates most of the OAIS functional entities, but not all are implemented as in the OAIS reference model. ReStore does not, for example, package descriptive information in order to provide access to actual content in the sustainable storage, nor does it offer a service aimed at recording digital preservation metadata for long-term accessibility such as file format, content authenticity, fixity, integrity or platform related details. These would be required in the future for new technological environments and systems. Figure A primarily relates to digital object repositories that have incorporated software such as EPrints, [33] Dspace, [34] or Fedora [35] and in which the primary processes are handled by

---

[32] Digital Preservation Europe, "DPE:Digital Preservation Europe" (Feb 09) available at http://www.digitalpreservationeurope.eu/video-training/prague-2008/?media=3 (accessed 18 Nov 09). Figure A is based on one of the diagrams presented in a training session "Preserving Digital Objects - Principles and Practice DPE, Planets CASPAR and nestor joint training event Prague, Czech Republic" in October 2008.

[33] EPrints, "Open Access and Institutional Repositories with EPrints" (09) available at http://www.eprints.org/ (accessed 18 Nov 09). EPrints is the most flexible platform for building high quality, high value repositories of research literature, scientific data, student theses, project reports,

the repository software under the control of a repository manager In Figure B, by contrast, we are talking about an on-line resources repository. Here, it is necessary to consider the architecture of the entire resource, which may comprise multiple digital objects of different types (images, video, audio, documents, etc.). The web resource received from the author is the Submission Information Package (SIP), one of the functions of the "Ingest" process in OAIS. There is much greater human input required, including interaction with authors, checking and updating content and attending to matters such as the transfer of IPR.

The dashed entities i.e. "Metadata Information" and "Access" in Figure B are not fully represented in the ReStoration model. The ideal scenario would be to further enhance the overall flow of processing so as to store metadata and content in separate locations, as suggested by the OAIS model and discussed further below, exposing content and metadata at the time of dissemination (when a web resource is being accessed). This would be possible only when the OAIS model is developed to address a more detailed implementation level.

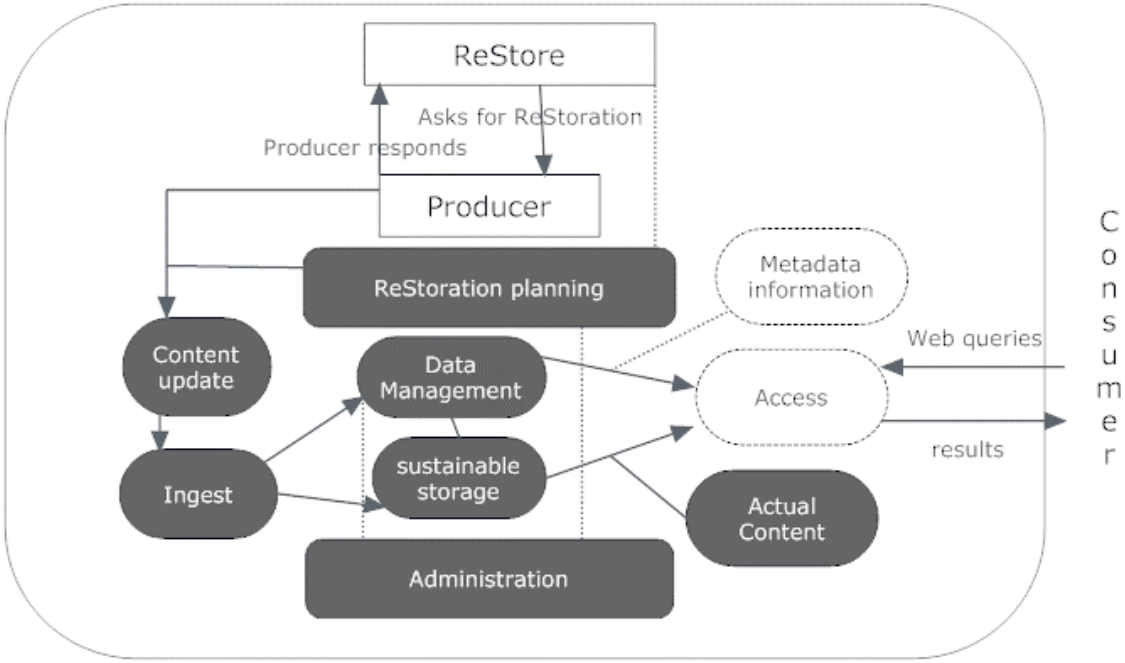

Figure B[36]

---

multimedia artefacts, teaching materials, scholarly collections, digitised records, exhibitions and performances.

[34] DSpace, "Introducing" available at http://www.dspace.org/about-dspace/introducing/ (accessed 18 Nov 09). DSpace is the software of choice for academic, non-profit, and commercial organisation s building open digital repositories.

[35] Fedora Commons, "Fedora Repository" available at http://www.fedora-commons.org/ (accessed 18 Nov 09). Fedora is a general purpose, open source digital object repository system (http://www.fedora.info) widely in use by digital library and repository manager and archivists for the purpose of preserving digital contents including web resources along with their metadata which could be accessed and shared on the web and through web API (Application programming interface) like REST/SOAP.

[36] Figure B is the modified version of Figure A. The modification has been made for the sake of establishing relationship between ReStoration and OAIS reference model.

8. What and when to sustain in the ReStore repository?

Generally at the start of a council-funded research project, there is no on-line presence and user awareness of the research is low. As the project team present their work at conferences and create an initial website, user awareness increases but the utility of the actual on-line resources is not realised until the end of the project, when the content of the website is complete and the resource is widely publicised. The resources may be valuable to researchers in all sectors and at all career stages.

The on-line resources reach their peak utility at around the time that the funding ends but user awareness continues to increase as the materials are cited in publications and presentations and also spreads by word of mouth. Since the web resource is highly likely to be indexed (commonly called "lazy preservation")[37] by search engines, bookmarked in users' browsers, and shared on social networking sites, it is of great importance to ensure that all URLs lead users to correct web pages with no dead links or outdated software. This is the time at which greatest effort needs to be devoted to sustaining the resource.

The term "ReStoration" refers to our approach to actively sustaining on-line resources rather than snapshot-based preservation of an individual page or pages in a remote data store. Collection of a particular web resource marks the beginning of preservation efforts that ultimately lead to archiving and, in the ReStore case, sustaining those resources for a specific period of time. The collection of web resources is very different in the case of ReStore as we neither crawl web resources nor harvest metadata and contents. Our intention is to identify and rectify all missing links, and server-related errors such as "page not found", Error 404 and internal server errors. We recognise, however, the significance of standard repository software, harvesting protocols such as OAI-PMH (Open Archive Protocol for Metadata Harvesting),[38] and more recently web server-based harvesting and metadata generation and exposure techniques. We will discuss these possibilities along with their limitations below.

Because of its bespoke nature, our approach could reasonably be characterised as resource-intensive, but the idea is to work closely with the primary resource author by meeting with them in order to understand the web resource in depth before restoring it into the repository. The intention is to specifically select resources for ReStoration, employing peer review of academic content worthy of such intensive effort. This approach seeks to ensure that the resource being restored is of significant value and that restoring it would maximise the return on the initial research council investment. This type of restoration not only preserves the on-line resource but maintains and regularly monitors the resource as well. We will discuss the "how" part of ReStoration in the following section.

---

[37] F McCown, JA Smith and ML Nelson, "Lazy Preservation: Reconstructing Websites by Crawling" (2006) available at http://www.cs.odu.edu/~fmccown/pubs/lazyp-widm06.pdf (accessed 18 Nov 09).

[38] OAI-PMH is a low barrier, HTTP-based protocol designed to allow incremental harvesting of XML metadata. An OAI-PMH repository is a network accessible server that can process the six OAI-PMH protocol requests and respond to them as specified by the protocol document. (See note 1 above.) OAI-PMH is also termed as metadata transfer protocol based on HTTP and XML.

9. Selection of web resources for ReStoration

Every user is familiar with the frustration of attempting to follow a broken link or opening web content that can no longer be read, played or viewed. Clearly it is possible to continuously maintain websites, but arguably most content is not worth such effort. An important question is therefore how to discern the quality of web content when there is no straightforward test. In particular, how should we determine what needs to be preserved for future use? Associated with this question would be issues relating to content ownership and copyright legislation,[39] which aims to protect the rights of the creator or owner when content is moved or copied from one place to another.

Since ReStoration involves the movement of a web resource from its original hosting location to ReStore repository, issues related to IPR, third party contributions and licensing are of paramount importance. Later sections review our experiences concerning copyright and IPR. [40] It is clear that to sustain a web resource that is of no practical value to researchers is a waste of time and effort. Bearing in mind the requirements of research resources outlined above, factors to be considered before sustaining a resource in ReStore may include the following:

- Does the resource have an active user base?
- Are the contents of the web resource being used and referenced by researchers and students as part of their academic activities?
- Are the contents of the resource of high quality and up to date?
- Have the developers and investigators taken enough care to avoid copyright infringement while uploading content, research papers, software tools and datasets?

The answers to these and related questions will determine to a large extent whether the benefits outweigh the costs of restoring and sustaining the resource for future researchers. Unlike other preservation initiatives, ReStore starts by contacting the original web resource authors in order to get more information before beginning the evaluation process. The evaluation process may include:

- meeting with the web resource creator/author and legal adviser of the institution who initially hosted the resource and with whom all copyright vests;
- determining whether or not to start the ReStoration process;
- initiating a review process – author, academic and technical reviews - of the resource;
- fixing links, updating data, generating/updating metadata, standardising the look and feel of the web resource by the author and/or developer;

---

[39] Office of Public Sector Information (OPSI), "Copyright, Designs and Patents Act 1988" (2008) available at http://www.opsi.gov.uk/acts/acts1988/Ukpga_19880048_en_1 (accessed 18 Nov 09). Copyright legislation in the UK is governed by the *Copyright, Designs and Patents Act 1988*.

[40] JISC Legal, "Intellectual Property Rights" (2008) available at http://www.jisclegal.ac.uk/LegalAreas/CopyrightIPR/ (accessed 18 Nov 09). Intellectual property rights (IPR), very broadly, are rights granted to creators and owners of works that are the result of human intellectual creativity.

- sorting out issues relating to copyright and third party contributions, and negotiating a deposition license agreement between the host institutions of the author and ReStore;
- transferring the web resource to the ReStore repository;
- deployment and promotion of the web resource within the ReStore repository; and
- a 6 month post-ReStoration review to determine whether or not ReStore will continue to host the web resource.

It is important to remember the context in which this activity is being undertaken. Authors of the resources in question have already successfully obtained national research council funding to create the resources in question. These resources have been the subject of significant academic effort, may be frequently used by an extended research community, and have already been targeted by the research council as candidates for ongoing support. The alternative to active preservation is to accept that the research community will suffer frustration and delay while attempting to use a slowly-decaying resource. Resource authors are also often keen to obtain further research council funding, and are thus amenable to taking some further action to enhance the impact of what they have already done. The financial investment required to sustain a resource for a further period of (for example) three years is generally a very small proportion of the investment already made in its initial creation.

10. IPR issues

Sustaining a web resource without taking heed of the IPR issues, including copyright and third party contributions, would seriously undermine the overall concept of sustainability. All efforts aimed at web preservation and/or sustainability involve some form of transfer or export of the content from the original web host to another that will preserve and sustain the resource in the future. The movement of content from one web domain to another is subject to copyright clearance under the UK Copyright and Design and Patent Act.[41] A web resource is thus truly sustainable only when, apart from fulfilling technical criteria for sustainability, all of its contents are original, trustworthy and free from copyright infringement.

For materials in the scope of Restore, the most relevant IPR consideration is generally copyright. As a general rule, copyright in a web resource will be owned by the author of the content unless the work was created in the course of employment of the author, in which case the ownership will usually vest in the employer. Dealing with issues such as different authors' collaboration in a piece of work, assessment of third party contributions, identifying copyright infringement (such as posting materials on a web site without consent of the original copyright holder, hosting and/or embedding unlicensed software in a repository site, copying over logos and trademarks without the express consent of licensor etc.) form the bedrock of an IPR strategy aimed at sustainable web resource preservation. We will discuss these and other issues in the following sections. These problems are not usually difficult to solve, but are frequently inadvertently overlooked during academic research-driven web resource creation and can be much harder to resolve retrospectively.

---

[41]Office of Public Sector Information (OPSI), "Copyright, Designs and Patents Act 1988" (2008) available http://www.opsi.gov.uk/acts/acts1988/Ukpga_19880048_en_1 (accessed 18 Nov 09).

10.1 Nature of web content and copyright

In a typical web resource, content comprises mainly text on a web page, images, videos, logos, trade mark and, optionally, programmable script that also generates text and/or images in response to user inputs.

A web resource has three fundamental components or attributes, namely: (a) appearance, (b) content and (c) navigational behaviour amongst various web pages. The risk of copyright infringement needs to be properly assessed in each of these areas by addressing the following questions:

- who is the architect/designer of the web site templates (Appearance);
- who modelled the basic navigational flow and behaviour of web pages through buttons, tabs and links on various web pages of the site (Navigational flow); and
- who supplied, managed and published the content of the resource site (Content)?

It is the web resource creator who is primarily responsible for sorting out issues pertinent to copyright and third party contributions[42] within their web resource. Now that publishing content on the web and sharing it with others is so easy, appropriate IPR management poses a real challenge to the web preservation community.

For a web resource to be fully ReStored and sustained for a particular period of time in the ReStore repository, the following conditions must be satisfied:

- identification of any third party content contribution during design and development;
- identification of formally published work, such as journal papers, that may have been included within the web resource;
- identification of any third party software, either hosted on the site or embedded in any form on its web pages;
- identification of any content in the web resource that has been produced using third party software (licensed or unlicensed);
- all relevant consents and permissions must be obtained, so as not to infringe the rights of any third party whose material is included in the resource;
- the relevant authority or party identified as the Licensor of the resource has signed the terms and conditions of the ReStore Deposition License agreement,[43] which is the last step before full ReStoration of the web resource.

These and other criteria will form the basis of an assessment by the ReStore team of each of the web resources in the scope of the project. The issues are addressed largely by a questionnaire addressed to the author, which allows any problem areas to be rapidly identified. Such arrangements not only ensure the smooth transfer of the actual web resource into the ReStore repository but also make it incumbent upon the licensor to transfer ownership rights to the ReStore team for handling future issues such as updating, adapting the resource at regular intervals. Importantly, none of these

---

[42] A third party contributor is someone who contributed content or aided web resource development or design, either directly or indirectly.

[43] The agreement which sets out all the terms and conditions relating to license and licensor must be signed before the web resource goes LIVE on the ReStore site available at www.restore.ac.uk (accessed 18 Nov 09).

issues should come as a surprise to a web resource author who has paid due regard to IPR in their work. Unfortunately there are often one or more areas that have been overlooked, thus requiring some specific attention at the ReStoration stage.

Our review shows that none of the major web resource preservation groups engage in individual licensing of every web resource potentially at risk of obsolescence. Most of these initiatives use a single step blanket copyright clearance agreement that in various ways could result in serious violation of IPR and copyright infringement. The complexity stems from the genuinely complex, diverse and heterogeneous nature of each web resource.

In order to develop a completely sustainable web resource preservation model, formulation of an IPR strategy is needed now more than ever before[44]. ReStore has initiated the development of a set of guidelines that aims to educate all those involved in web resource creation so that resources might be sustained with less effort in the future. These guidelines are currently under review, and will be published in the autumn of 2009 and widely promoted – particularly to the ESRC research community.

10.2 Formulating IPR strategy

In the case of a web resource, where numerous people are directly or indirectly involved and where the level of understanding of participants varies, it is a daunting task to formulate a complete policy and set of standards. Adopting an appropriate framework at the outset of a web resource project would however have great advantages when it comes to sustaining the resource in the future. Three major types of stakeholder must be considered before formulating an IPR strategy:

1. primary resource authors and/or creators;
2. web resource project funders; and
3. third party contributors.

It is possible that each of these will have different policy frameworks or none at all[45]. To appropriately combine all these interests so as to establish a unique IPR policy may not be practical. However, even simple approaches to record-keeping and IPR management may greatly assist. The following steps could prove to be a valuable starting point in the direction of web resource sustainability. These also form the core of the resource review process adopted by the ReStore team:

- Is there anybody contributing to the site who is not part of the project team?
- Are records being kept of current project staff or team members and of their contractual employment arrangements with the organisation that owns the resource?
- Is it possible that any third party content is being incorporated into the site e.g. software tools, temporarily hired developers, content contributors, etc?
- Is all third party content in the site being properly tracked and proper permission obtained and recorded for its use?

---

[44] JISC, "Managing and Sharing e-Learning Resources: How repositories can help" (2008) available at http://www.jisc.ac.uk/publications/documents/elearningrepositoriesbpv1.aspx (accessed 18 Nov 09).

[45] A Charlesworth, N Ferguson, S Schmoller, N Smith and R Tice, "Sharing eLearning Content: a synthesis and commentary" (2007) available at http://ie-repository.jisc.ac.uk/46/1/selc-final-report-3.2.pdf (accessed 18 Nov 09).

- Are contributors aware that they need to seek owners' permission to upload journal papers or other outputs to the web resource site?
- If there is user-generated content as part of the resource, has a non-exclusive irrevocable license been obtained for publishing, adapting and repurposing this content?
- If third party content (technology, services, software, etc.) has inadvertently been used during the development or subsequently, has consideration been given to the level of risk involved?
- Has consideration been given to developing a take down policy and appropriate notice?
- Is it clear who deals with issues relating to copyright, IPR in the institutions involved?

The above list does not exhaustively cover every aspect of IPR strategy relevant to web resource development but it illustrates where major issues that can arise and reflects the range of situations we have encountered on the ReStore project.

11. ReStore and digital repository networks

As per our earlier definition, a repository is a container for keeping things for future use. A number of standard repository software tools have been introduced in the recent past to help institutions and organisation s start preserving their web resources. EPrints,[46] Fedora,[47] DSpace,[48] are examples of open source repository software being used to digitally preserve resources including web resources. As these software platforms are, however, certain to evolve over the next 4-5 years, the emphasis should be on a "repository service" rather than a particular software platform.[49] The repository software helps the archivist to improve the visibility of hidden knowledge in the web resources, share knowledge through metadata with other repository platforms, enhance long-term preservation of digital assets and improve cross-searching facilities across digital repositories.

Similarly, even though digital libraries are accessed as web sites, anyone involved with digital libraries will be able to point out many differences between everyday websites and a true Digital Library (DL).[50] The web is an amalgamation of digital pages with little metadata and unpredictable additions, deletions and modifications, which is quite different from a DL that has rich metadata and well-organised content.[51] Further, unlike a web resource repository such as ReStore, which only supports HTTP request response events, a DL also supports other protocols such as OAI-PMH, which is the most widely used protocol for metadata preservation,

---

[46] See note 33 above.

[47] See note 35 above.

[48] See note 34 above.

[49] JISC, "JISC:Digital repositories roadmap: looking forward" (2006) available at http://www.jisc.ac.uk/media/documents/programmes/reppres/reproadmap.pdf (accessed 18 Nov 09).

[50] Joan A Smith, Michael L.Nelson "Creating preservation-ready web resources" (2008) available at http://www.dlib.org/dlib/january08/smith/01smith.html (accessed 18 Nov 09).

[51] See note 50 above.

deployment and harvesting, and is compatible with almost all repository software such as Eprints, Fedora, Dspace and recently mod_oai.[52]

11.1 Web repositories and OAI-PMH

According to a report published by JISC,[53] two current standards underpin much current repository activity. Firstly, OAI-PMH is used to support the regular gathering of metadata records from repositories by other service components in the information environment. Secondly, metadata records exchanged using the protocols are typically based on the Dublin Core metadata and standard.[54] All these standards have, however, been evolving since the publication of the report, and such protocols are now capable of processing metadata in other formats as well. We will now turn to a discussion of metadata generation, deployment and collection.

A web resource repository that supports processing of OAI-PMH requests from other digital libraries and repositories is part of a bigger knowledge network in which the emphasis is on information sharing, data and metadata preservation, and not merely content storage. Since the ReStore repository is still a prototype service, the focus is not support for such protocols, but is currently on sustaining web resources for which metadata already exists. Long term preservation will however be considered, and may eventually result in exposure of preservation metadata for all contents of the ReStore repository.

A good model for sustainable preservation needs as much metadata as possible including keyword lists, content summaries, subject, structural details, copyright, authorship, application version, etc.[55] Although currently relying on the user-provided metadata in each page of a web resource, automatic metadata generation could, in the future, make all ReStored web resources in the ReStore repository available along with a rich set of metadata (potentially in multiple formats) for the benefit of dissemination and sustainable preservation.

12. Metadata generation, deployment and harvesting

Metadata is an integral aspect of web resources preservation. Metadata are structured data that describe the characteristics of a resource. Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information.[56] Almost all digital repositories require metadata

---

[52] Mod_oai is an Apache 2.0 web server optional module like mod_perl which exposes a web server as an OAI-PMH repository.

[53] See note 49 above.

[54] *Ibid*.

[55] JA Smith and ML Nelson, "Creating Best effort Preservation Metadata for Web Resources At Time of Dissemination" (2007) available at http://www.cs.odu.edu/~mln/pubs/jcdl07/jcdl07-best-effort-metadata.pdf (accessed 18 Nov 09).

[56] National Information Standards Organization (NISO), "Understanding Metadata" (2001) available at http://www.niso.org/publications/press/UnderstandingMetadata.pdf (accessed 18 Nov 09).

along with the original web resource before the resource is hosted on the repository site.

One of the core requirements for incorporation of a web resource into the ReStore repository is that every web page must have its own metadata, ideally added manually by the resource authors. The ReStore guidelines emphasise the importance of this to our long term sustainable web resources strategy, because the vast majority of web resource developers are not IT professionals and may be unaware of the importance of metadata. Three steps to understanding metadata are equally important to ReStore and other repository networks:

- metadata creation;
- metadata deployment; and
- metadata harvesting.

12.1 Metadata creation

Adding metadata to a web resource at the time of web page creation is the ideal. By embedding metadata in every web item in a web resource, the author and/or developer are able to enhance the sustainability of their resource from day one. Promoting this metadata generation at source is an element of ReStore's work in providing guidance for ESRC-funded resource creators.

12.2 Metadata deployment

Various standards for embedding metadata in a web page are currently in use. An HTML document could be linked to the web page, which is not necessarily held on the same server. Another approach would be to link a database to the web resource and populate each web page with the metadata from the database. Harvesting metadata and committing it to the database is another process by which metadata could be collected from either a harvester,[57] or directly from web pages.

The deployment can also be performed by placing the metadata interactively into web resources by using scripting languages that can import metadata stored in XML, RDF or other formats into a web page. Such an approach has the disadvantage, however, of burdening the web server with another function besides processing client requests.

12.3 Metadata harvesting

A web resource repository such as ReStore hosts only web resources that are not being actively maintained because their original funding has ended. Its web resources repository runs on a platform that is configured for HTML, XHTML,[58] PHP,[59]

---

[57] Metadata harvester is a metadata indexing system (an application) that issues OAI-PMH protocol requests in order to harvest XML formatted metadata (created by web server or metadata tools).

[58] The Extensible Hypertext Mark-up Language, or XHTML, is a mark-up language that is almost identical to HTML but cleaner and stricter, and also conforms to XML syntax.

[59] PHP is a scripting language originally designed for producing dynamic web pages.

MySQL,[60] Server and Apache [61] web server.[62] The web server is not configured for honouring OAI-PMH requests made by OAI-PMH repository.

Typically an archivist will crawl (not harvest) a target web site, such as UK Web archive or Internet Archive, then process each resource (contrary to just in time processing by a mod_oai-enabled Apache web server) with various metadata utilities (discussed below) to extract technical information (largely Dublin core supported metadata and HTTP header information).[63] This pre-processing of a web resource for preservation takes place at the location of the archivist.

From the point of view of the web resource developer or web master, the ideal way of doing this would be to install on the web server a tool that manages itself, and which automatically provides the necessary "extra information" (ie metadata) for the archiving site to prepare the web site for preservation, and which does not impact on the normal operation of the web server (ie processing HTTP request and providing response to client browsers).[64]

The ReStore repository only ensures that the metadata embedded in HTML META tags is sufficient to describe the contents of a web page, rather than all of the the attributes and actions of the digital assets (images, videos, PDF, Word document etc) represented by a web page. All OAI-PMH enabled archives, and search engines such as Google, MSN and others, currently index all web resources residing in the ReStore repository through HTTP header information.[65] This information is not, however, sufficient in itself for sustainable web resource preservation.

13. Automatic metadata generation

The role of metadata in sustainable preservation of web resources is crucial, and descriptive metadata - generated either by the resource developer or by tools - add substantive meaning to the resources being sustained. Although the preceding sections have highlighted the importance of manual metadata creation, the fact remains that the majority of web resource creators do not have the necessary skills for or spend enough time on metadata creation. There is therefore a role for software tools and technologies that automatically create metadata for web resources and package them into various formats for easy portability and dissemination.

---

[60] MySQL AB, "MySQL: MySQL 5.0 Reference Manual" available at http://dev.mysql.com/doc/refman/5.0/en/what-is-mysql.html (accessed 18 Nov 09). MySQL is a relational database management system which acts as a database component in a web application or web resource environment.

[61] Apache, "The Apache HTTP Server Project" available at http://httpd.apache.org/ (accessed 18 Nov 09).

[62] A computer program that is responsible for accepting HTTP requests from clients (such as web browsers i-e Internet explorer, Firefox etc) and serving them HTTP response along with optional data contents which usually are web pages such as HTML documents and linked objects such as images, flash video etc.

[63] See note 50 above.

[64] *Ibid*.

[65] HTTP header contains information for a web server (e-g Apache web server) and web client (browser) such as content type (e-g html, php, xml, jpg, gif, mpeg, doc etc) content size (e-g bits, bytes, gigs), date of content creation, date of content modification, upgradation etc.

As a result of the unstructured and complex nature of web resources, the reliability of utilities and tools for automatic metadata creation is still questionable. Web research groups in the UK and abroad are still looking for ways to fine-tune these tools and utilities for better analysis of web resources, and higher precision in the generation of metadata. In addition to third party utilities for metadata generation, a web server also generates metadata, but it is not enough to describe individual web objects in a web resource. Web servers are optimised for the "here and now" and support of digital preservation is not a functional design requirement. [66]

There are various utilities that could be used for metadata generation, either at the end of web resource collection or on the fly, while a web page is being requested by a client browser. The ReStore web resource repository is not currently implementing any of these utilities, but the task of automatic metadata generation would be best performed by the current server if mod oai software was added,[67] so this type of automatic metadata generation and deployment may be considered in the future.

A variety of open source and command tools such as Jhove, Open Text Summariser, KEA, ExifTool, etc are available for analysing files and generating preservation metadata. These are not usually integrated into the Apache web server setup, which is focused on analysing each file and rendering it on a user's browser. To incorporate one of the above tools for equipping the server to carry out the extra task of generating OAI-PMH compliant metadata, mod_oai can be added into the server configuration file. This would enable the server to generate not only an HTTP response but an XML[68] formatted document containing HTTP-header metadata (file type, modification date, etc.) as well as the resource file itself, which is the core of sustainable web resource preservation. mod_oai, which presents a processed form of web page containing both data and metadata, offers a sustainable Archival Information Package (AIP) and implements the OAIS design discussed in relation to Figure B above. The third and last step in implementing OAIS is the Dissemination Information Package (DIP) representing the version of a web page on the repository site that has to be disseminated to a service provider such as OAISter http://www.oclc.org/oaister/ for long term digital preservation. This last step is possible only once the repository has been registered with such a provider.

One of the most important advantages of using such software modules on an Apache web server is that the server packages the resource and associated metadata in a format that is long lasting and does not frequently change, unlike traditional preservation approaches where the hazard of format change is very high.

13.1. Limitations of a mod_oai compliant web server

---

[66] See note 50 above.

[67] Mod_oai is an Apache 2.0 web server optional module like mod_perl which exposes a web server as an OAI-PMH repository.

[68] XML (Extensible Mark-up Language) which is a general purpose specification for creating custom mark-up languages. It is called extensible because it allows user to define the mark-up elements unlike HTML.

A web server differentiates between a static web page and a dynamic one and generates HTTP header information accordingly. Adding mod_oai can substantially increase the role of a typical Apache web server in preserving web pages but this is not a panacea for all preservation-related problems, especially when it comes to sustaining dynamic web resources or sections of web resources. The mod_oai compliant web server, which - it has been suggested - could be a replacement for DSpace, [69] Fedora,[70] EPrints, [71] and other digital repositories, still needs to be fine-tuned for dynamic web resources that are rendered on the fly, based on diverse scripting and programming languages, or are populated with content from a remote database server.

An Apache web server, like the one used for the ReStore repository, will, depending on the type of file, often transform files before serving them to the client eg .cgi, .php, .shtml, .jsp etc).[72] Passing a secured web page – one that requires security credentials from users - to a crawling and harvesting repository would be a serious breach of copyright terms and conditions.

The ReStore approach offers some solutions to the issue of file counting (everything or less), serving both static and dynamic files with locally configured commonly used web servers such as Apache, IIS,[73] etc. A further limitation of the need to configure a web server to accept OAI-PMH requests is the high burden that this would place on the Apache web server to process both standard HTTP requests from client browsers and those from OAI-PMH repository harvesters.

14. Sustain forever?

There is a genuine question, especially important when the cost of maintenance is high, regarding how long an on-line resource should be maintained. In the case of ReStore, the working assumption is that each resource will be sustained initially for three years, subject to review. The review is based largely on resource usage statistics obtained from several sources such as Google analytics, the repository's own web stats server and counter software. Other factors that can influence continued active sustaining of a web resource are the (externally reviewed) quality and utility of its contents and the uniqueness of its research findings, tools, software, etc.

Continual digital curation of a web resource on any repository platform comes at a cost, and requires meticulous planning to ensure continued public accessibility of contents. Before committing monetary resources and expertise to sustaining a web resource, a cost-benefit analysis should be undertaken to assess its true value. Ideally, users and usage should be the key drivers of the decision as to whether to continue sustaining the resource or remove it, perhaps with the intention of static web archiving thereafter.

---

[69] See note 34 above.

[70] See note 35 above.

[71] See note 33 above.

[72] See note 1 above.

[73] Internet Information Services (IIS) created by Microsoft is a set of Internet-based services for servers used in the similar capacity as Apache web server but processing different scripting (programmable) web pages like asp, aspx etc.

15. Future work

Live websites gradually implement software upgrades, change hardware platforms and perhaps even adopt new protocols.[74] Consider gopher,[75] ftp,[76] and telnet,[77] which have mostly been replaced by http/https, scp, and ssh[78]. HTML 1.0 has evolved to SHTML and XHTML, and a number of early HTML tags have been deprecated.[79] These all indicate the challenge facing sustainable preservation efforts. Copying, backing up, and storing web resources in a database and accessing them through these protocols can guarantee smooth access to every web asset in every web resource but it falls short of sustaining the way we access them today. The case of resource metadata is similar: today's metadata may be insufficient in format or coverage for tomorrow's search engine harvesters and access protocols.

The present ReStore work plan does not include automatic metadata generation or deployment and serving of metadata harvest requests. Rather, the emphasis is on training and motivating web resource creators in the realm of research method resources to help them create standard web resources with due regard to appearance, information representation aspects, metadata and IPR issues.

Considering the efficacy of OAI-PMH [80]enabled repository services, we are inclined to enable our web server to process OAI-PMH requests in the future. Preferably the web server would apply metadata analysis tools at the time of dissemination request by an OAI-PMH harvester, unlike crawler software that crawls, collects and stores resources and applies metadata analysis tools later. In the case of OAI-PMH, which is more focused on resource metadata such as authorship, copyrights, creation and modification dates, adaptation of the ReStore repository to process OAI-PMH protocol requests may be considered.

16. Conclusion

Having reviewed current digital repository initiatives, it is clear that both crawling and harvesting approaches to web resource preservation have some serious limitations. None of the approaches ensures that web resources are preserved in their entirety. The grey areas that cause problems stem generally from the dynamic nature of web resources and in particular from the dynamic web pages within web resources.

---

[74] See note 1 above.

[75] The Internet Engineering Task Force (IETF), "The Internet Gopher Protocol" (93) available at http://www.ietf.org/rfc/rfc1436.txt (accessed 18 Nov 09). Gopher protocol is a TCP/IP application layer protocol designed for distributed document search and retrieval over the Internet and used to be an alternative to the World Wide Web. It was popular for campus-wide information systems.

[76] File Transfer Protocol (FTP) is a standard network protocol used to promote sharing of files and transfer data reliably and efficiently over a computer network.

77 The basic purpose of Telnet protocol is simply providing a facility for remote logins to computer via the Internet.

78 Webopedia, "What is SSH" available at http://www.webopedia.com/TERM/S/SSH.html (accessed 18 Nov 09). Secure Shell is a program to log into another computer over a network, to execute commands in a remote machine, and to move files from one machine to another.

[79] See note 1 above.

[80] See note 38 above.

After affirming the need for sustainable web resource preservation and discussing current approaches, we are able to conclude the following.

- Approaches to both short-term and long-term web preservation are at an experimental stage, and the web preservation community needs to move toward a consensus on standards, concepts, terminologies and the technological environment aimed at sustainable web preservation.
- Web resource preservation should not be just an individual or group activity, but rather be embedded within organisational strategies in order to ensure accessibility to valuable knowledge that is accumulated slowly but can vanish very fast. In the case of ReStore, ESRC - as the original funder of the resources - is taking proactive measures to preserve and enhance the impact of its research investment.
- No single technology platform, hardware or software tools will produce the desired result of preserving everything available on the web. All content producers need to be made aware, trained and educated on how to produce web resources that last longer, regardless of how far in the future they may be accessed by a community of users.
- Not everything on the web could or should be preserved or sustained, and therefore a well planned selection strategy must be at the centre of any sustainable preservation policy.

We have drawn on our experience with the ReStore initiative to date, and compared it with the current state of the art in traditional web preservation models, including the ISO standard OAIS reference model. We do not imply or assert that the ReStore approach is either better or worse than another, but do identify our approach as a unique way of sustaining on-line research method resources for future access. One of its most important contributions may be that of encouraging website creators to plan from the outset for the future sustainability of their resources.