Building an Internet Archive System for the British Broadcasting Corporation

Cathy Smith

Abstract

Since its beginnings as the British Broadcasting Corporation (BBC) Networking Club in April 1994, the BBC's Web site has grown to over two million pages. While bbc.co.uk inarguably offers a valuable source of information, entertainment, and education for its users and provides an online arena for peer-to-peer communication, it also brings into focus the challenge of digital preservation. Apart from the sheer volume of material the site represents, the nature of that material is forever changing both to reflect editorial strategy and to benefit from new technologies and improved production techniques. To support its own internal business requirements and to satisfy external legislative requirements, the BBC's Information and Archives Department is building a Legal and Historical Internet Archive System to capture a selection of content as it is published to the "live" site. This article looks at how the design and development of that system supports the preservation of heterogeneous digital material in the wider context of archiving the BBC's new media output.

The British Broadcasting Corporation's (BBC)[1] Information and Archives Department—a relatively recent amalgamation of research libraries, archives, and preservation services across the BBC's national and regional operational centers—manages much of the corporation's physical and electronic records and audio visual assets. One of its current projects is the introduction of a system for the automatic capture of the BBC's online services published to bbc.co.uk. This article describes the development of that system and its design and implementation in the context of the corporation's main business driver: the creation of distinctive programs and services.

LIBRARY TRENDS, Vol. 54, No. 1, Summer 2005 ("Digital Preservation: Finding Balance," edited by Deborah Woodyard-Robinson), pp. 16–32

Information and Archives (I&A) has always been responsible for a wide variety of material: documents, television and radio programs, and in recent years the digital and electronic equivalents. Preservation is an ongoing part of the remit; converting archived content into new formats has been an issue within the Television Archive since 1948, when nitrate was used for film production and archivists were already aware that alternative stock would need to be developed to save material in the long term. And in the BBC archives there is a vast amount of it.

Stored in multiple sites around the United Kingdom are over 600,000 hours of complete television programs, stockshots, and unedited or untransmitted material held on film, 2-inch and 1-inch videotape, and U-Matic and Beta formats. More than 300,000 hours of radio are available on wax cylinders, vinyl records, audio cassettes, CDs, and DAT or one-quarter-inch tape, with in excess of 25,000 sound effects captured on CD and vinyl. And then there are the 100 million documents held at the Written Archives Centre, 3 million photographs—hardcopy and electronic, 22 million newspaper cuttings, 1.2 million commercial music recordings, and 4 million items of sheet music. And all of that needs to be managed through processes for intake, cataloguing and indexing, research and access, and long-term preservation. So, although we are in an increasingly digital environment, the issues are not new—it is only the challenges that are different.

What Are the Drivers for Archiving and Preservation for the BBC?

The BBC is beholden to several pieces of often contradictory UK legislation. One of these is the Broadcasting Act of 1996, which dictates that "it shall be the duty of each broadcasting body to retain a recording of every television or sound programme which is broadcast by that body—

(a) where it is of a television programme, during the period of 90 days beginning with the broadcast, and

(b) where it is of a sound programme, during the period of 42 days beginning with the broadcast" (Queen's Printer of Acts of Parliament, 1996a, Section 117, Part V).

In other words, the BBC is legally required to record its TV and radio output off-air to enable the corporation to answer complaints from the listening and viewing public.

This is reiterated by the 2003 Communications Act, which pays legal lip-service to the existence of platforms other than radio and television by applying the retention periods of 90 and 42 days to "every programme service" (Queen's Printer of Acts of Parliament, 2003a, Section 334). The act also includes a requirement "to comply with any request by OFCOM[2] to produce to them for examination or reproduction a recording retained in pursuance of the conditions in the licence" (Queen's Printer of Acts of Parliament, 2003a, Section 334).

Despite being relatively new legislation, disappointingly the act does not take into consideration the technical challenges of capturing new media services direct from broadcast, preserving them in their original form, and being able to re-create them as required in answer to legal claims. While neither act specifically mentions the Internet—or in fact any other new media platform—the BBC does not want to set any legal precedents but would rather demonstrate best endeavors to meet the letter of the law. To that end, an internal agreement was made in 2000 between the Head of Online and the BBC's Programme Complaints Unit that bbc.co.uk would be defined as another broadcast channel and its content captured off-air as with TV and radio.

But there are other legislative reasons for retaining the BBC's broadcast output that shape archive policy. The period of liability under the Defamation Act of 1996 is currently one year, which suggests that the BBC should be retaining content for longer than the Broadcasting Act's ninety-day requirement. Under section 2 of the act, any "offer to make amends" is an offer to "make a suitable correction of the statement complained of" and to " publish the correction and apology" (Queen's Printer of Acts of Parliament, 1996b). Consequently, at least one function of the archive is to provide a record of both the original defamatory material and any consequent official response broadcast on air.

Another raft of legislation centers on information and data management. Schedule 1 of the 1998 Data Protection Act outlines the data protection principles, which include that "Personal data shall be obtained only for one or more specified and lawful purposes" and that "appropriate technical and organisational measures shall be taken against unauthorised or unlawful processing of personal data" (Queen's Printer of Acts of Parliament, 1998, Schedule 1, Part 1).

BBC online publishes an increasing amount of user-generated content. "Talk" (http://www.bbc.co.uk/communicate/) calls for feedback on the BBC's output and provides a gateway to the message boards, while sites like "iCan" (http://www.bbc.co.uk/dna/ican/) and "Collective" (http://www.bbc.co.uk/dna/collective/) provide platforms for peer-to-peer debate on local and cultural issues. The personal data shared over the BBC's network needs to be managed in accordance with the requirements of the Data Protection Act; although this is more of a concern for content producers who directly receive and manipulate the data, there could be implications for data that is ingested into a long-term archiving system.

More specifically relevant to bbc.co.uk, the United Kingdom's Legal Deposit Libraries Act of 2003 is enabling legislation intended to extend the concept of "legal deposit" to cover electronic publications, including Web pages. The act is clear about who dictates the delivery of content to the deposit libraries: "where a work is published or made available to the public in different formats, (to) provide for the format in which any copy is to be delivered to be determined in accordance with requirements specified (generally or in a particular case) by the deposit libraries or any of them" (Queen's Printer of Acts of Parliament, 2003b, Section 6). Though daunting, the act highlights the general desire to preserve the national heritage locked in electronic publications even if the technical problems have yet to be resolved. It also raises the question of who has responsibility for preservation.

Will the British Library—and any other institutions granted "legal deposit" status under the Legal Deposit Libraries Act—become the nation's "digital preserver"? And would that mean the BBC should hand over its content and relinquish such a role? Considering the business demands of the corporation, probably not, but there will be a need for developing a close strategic relationship with the deposit libraries not least because they will have the authority to dictate the means of delivery, which in turn could have a major financial impact on the corporation aside from the implications for content creation processes.

Part of the thinking behind the Legal and Historical Internet Archive System is that, if we can demonstrate that the BBC is responsibly managing its own collection of published Internet output, then perhaps we could work with the deposit libraries to agree upon the nature and regularity of deposits instead of having to meet the blanket requirements of the Legal Deposit Libraries Act. Deciding which URLs suit the needs of which national institutions based on their own selection criteria (for example, the National Library of Wales has already expressed an interest in content from BBC Cymru'r Byd [http://www.bbc.co.uk/cymru/]) would relieve pressure on the BBC's delivery mechanisms and provide focused collections specific to local needs.

I&A already enjoys long-established relationships with external agencies, including the British Film Institute and the British Library Sound Archive, the first of which receives previously agreed-upon examples of our television output, and the second provides public access to our collection. So why not establish something similar for bbc.co.uk?

The Legal Deposit Libraries Act also requires the delivery "with the copy of the work, a copy of any computer program and any information necessary in order to access the work, and a copy of any manual and other material that accompanies the work and is made available to the public" (Queen's Printer of Acts of Parliament, 2003b, Section 6). This in itself raises issues of copyright and licensing. What special dispensation needs to be agreed to allow legal deposit libraries not only the right to archive commercial software but also to use it to provide public access to stored content? Within the BBC the situation is no simpler. Software licensing models are different for each manufacturer and none encompasses the need to retain examples in the long-term for purposes not covered when the original licenses were signed.

But licensing aside, discussions with the BBC's technology support departments have focused on the possibility of creating a central store of decommissioned software that might be recommissioned for the purposes of supporting archived content at least in the short term. Clearly, however, that sounds simpler than commercial relationships currently allow.

Apart from external legal requirements, the BBC has the tenets of its own Royal Charter and Agreement to adhere to, including the corporation's responsibility for preserving the national heritage as reflected in its own broadcasting history. The agreement states that the BBC is to "establish and maintain . . . an archive or archives of films, sound recordings and other recorded and printed matter which is representative of the sound and television programmes and films broadcast or transmitted by the Corporation" (Department of National Heritage: Broadcasting, 1996a, Section 11). This is an expansion of one of the "Objects of the Corporation" in Section 3 of the charter to "establish and maintain libraries and archives containing material relevant to the objects of the Corporation, and to make available to the public such libraries and archives with or without charge" (Department of National Heritage: Broadcasting, 1996b).

So not only should I&A be responsible for establishing a process for capturing, storing, and preserving the BBC's online output, but it should be considering how to make that collection accessible to the public. Both are something of a challenge and both are crucial in this period of charter renewal and external review. The BBC's services are in the spotlight, and should there be any major changes to bbc.co.uk, it is imperative that I&A be ready to record them within an archive of "what was" or, at the very least, to retain descriptive metadata of the content of decommissioned pages alongside information detailing context of their removal from the live site.

In fact, the site has already begun this process of change following the publication in July 2004 of the report of the Independent Review of BBC Online (Graf, 2004) conducted on behalf of the UK's Department of Culture, Media and Sport. The report's recommendations led to the closure of a handful of sites deemed either to duplicate content available elsewhere on the Web or—though serving the public interest—to attract too little traffic to make them economically viable. But irrespective of the external drivers, archived bbc.co.uk content has potential for reuse, provides a good source of research, and enables producers to reflect on past output in the same way as any of the BBC's broadcast material held in the Radio and Television Archives.

So What Is It That Needs Capturing?

bbc.co.uk began in 1989 when it was registered as bbc.uucp ("Unix-to-Unix copy") with dialup access to the site via Brunel University; it was only available to BBC Development. It was then reregistered with the UK's academic "naming" body, the Name Registration Service (NRS), as bbc.co.uk in 1991, but it was still only available inside the BBC to the renamed BBC Research and Development Department. Three years later, on April 13, 1994, the BBC Networking Club "opened for business" at bbcnc.org.uk. Commissioned by BBC Education to support the BBC2 series *The Net*—the first program with an online forum via the "Auntie" bulletin board—it was aimed at introducing viewers to the Internet: what it was, how it worked, and what it had to offer. Membership cost £5 a month, which gave subscribers access to "early-adopter sites" including "Top Gear," "Tomorrow's World," and Radio 4's "Woman's Hour." It also provided nine-day TV and radio schedules and a "Guide to the Internet" introduced by a character called "Babbage." The BBC Networking Club was an Internet service provider, communication facilitator, information supplier, and Web publisher. In November 1997 News Online "went live" via their in-house-developed content production system, and by March 1998 the BBC Online home page was providing a portal to the BBC's Internet services.

Since then bbc.co.uk—as is now the accepted branding for the site—has grown to at least 750GB of content with daily incremental updates of 3–4GB. Content includes as wide a range of file types imaginable, providing audio, video, animation, and text, and meeting any number of editorial criteria within the BBC's overall remit to provide "information, education and entertainment" (Department of National Heritage: Broadcasting, 1996b, Section 3). The potential of the Internet as a broadcasting platform is exploited to the full with constant experimentation of potential services that technologically dictate the means of archiving sites and pages.

Currently content is produced within genre-based departments—for example, Learning and Interactive, Drama and Entertainment—and overseen by a central editorial and technical team that also has responsibility for other new media platforms, including interactive television and mobile devices. Methods of content creation range from the use of commercial tools and software—Microsoft's FrontPage, Macromedia's Dreamweaver, Adobe's GoLive—to handcrafted HTML and in-house content production and management systems.

The biggest challenge for archiving content is the degree to which it might be described as "dynamic." In other words, it should be relatively easy to collect flat HTML pages—complete pages held on Web servers—but less easy to harvest those that are created within client Web browsers when a particular URL is requested at a particular date and time and content is provided from backend databases.

Only a relatively small percentage of bbc.co.uk pages have been or are currently dynamically driven. The now decommissioned and only semi-dynamic "myBBC" site gathered content from specified sources based on user profiles. This content was held on the Web server in readiness for publication and distributed via the publication system in response to user requests. In comparison, the bbc.co.uk home page is updated at least twice daily with headlines and supporting images provided directly from the News Online content production system. And then there is content that appears dynamic: the constantly updated travel information is actually provided via data feed from a third-party supplier. It is then FTPed to the BBC to be transformed into HTML on a production server and published to what is called the Master Content Server (MCS) for Web distribution every thirty minutes. The MCS is key to content publication. Online producers publish new files to the "BORG2 queue"—a mechanism that chronologically makes updates to the site and that provides the means of content delivery to the Legal and Historical Internet Archive.

The Legal and Historical Internet Archive System

On August 26, 1999, Sir John Birt, the BBC's director general at the time, emailed a request to the then Head of Heritage to "work out what we need to do to preserve the BBC's early work on the Internet." This led to I&A being commissioned by the head of BBC Online to devise an "Online Archive Policy," which in turn led to a series of interviews questioning producers about the type of content they were creating; what they had retained of previously published pages and sites; and what they thought was logistically and technologically possible in terms of capturing output that could be accessed and re-created in the long term.

Engaging with content producers was sometimes difficult. At the time the Internet was seen as ephemeral: content was published today and deleted or overwritten tomorrow. Why should it be archived at all? Even the legal teams argued that, if it proved too difficult a challenge, we had had few complaints about BBC Online; not having the content with which to answer claims was a defence in itself. But that was not good enough in light of the charter requirements and ignored the potential for research and reuse afforded by an Internet archive.

So what were the options? As an alternative to the existing model for television and radio with I&As managing central collections, a decentralized BBC Online collection would mean individual production departments capturing material at point-of-creation as a process integral to their content production or media asset management systems. Local collections would then be linked to a central metadata repository managed by I&A, which in turn would be linked to databases holding relevant rights information necessary for the reuse of content. It quickly became apparent that production departments did not want archiving to encroach upon content creation. Archiving was I&A's responsibility on behalf of the BBC and, while they would advise on its development, any archiving system would have to remain centralized and independent.

An initial trawl of the Internet and conversations with other broadcasters pointed to there being no similar projects or initiatives underway and certainly no off-the-shelf software solutions. It did seem, though, that the functionality required for an online archiving system was closest to that provided by systems designed for newspaper publishing and electronic records management. They could handle multifaceted compound objects—individual components such as text, images, graphics, and complete documents—and offered version control and metadata management. But it was the need to implement as cost-effective a solution as possible that led to the decision to design and develop an in-house system that would

- be dependent on the online production process for the delivery of content, but neither be integral to nor have any detrimental effect upon it;
- be as "future-proof" as possible in order to support the capture of new formats;
- support digital objects irrespective of the source or file type;
- have the ability to display each Web page as originally published to bbc.co.uk;
- provide archival storage for retaining Web pages in perpetuity.

The budget also partially determined the scope of the system. Instead of aiming to provide a source of reusable content—which would have demanded access to adequate rights information—the system would meet only the BBC's legal, charter, and historical requirements. On that basis, BBC Technology was commissioned to develop, implement, and support the Legal and Historical Internet Archive System.

To keep the first iteration of the system as simple as possible and to focus on ensuring basic delivery of content to the archive, it was decided that the system would *not* capture

- audio/video content—most of which would be stored in its original format in the Television or Sound Archives;
- material published via News Online's content production system including Sport, Weather, and the World Service, all of which—excluding home pages and indexes—remains on the live site;
- dynamic database-driven content, which only exists in response to user requests and cannot be captured at the point of publication.

To ensure that the BBC retained some kind of record of published output, it was also decided that the system would store in perpetuity all metadata associated with content whether in- or out-of-scope.

The aim, however, is for future iterations of the system to include currently excluded output, examples of which have obvious historical and legal value:

- Most audio/video material available on bbc.co.uk is created for television or radio transmission and reformatted for the Web, but there is a growing trend for publishing full online versions of interviews, concerts, and festivals originally edited for broadcast. Without an inclusive archive strategy, there is a risk that these will be lost along with any examples of audio/video content being commissioned specifically for the Web—another current trend.

- News Online's home pages and indexes most clearly demonstrate any design or style changes and, because of the frequency and number of changes made to those pages as they keep pace with developing news stories, they could be the most litigious of all the BBC's output.
- Being dynamically created, message boards also fall out of scope but are historically valuable as a demonstration of the Web providing a platform for peer-to-peer communication on topics important in today's society.

That said, there is the possibility of investment in other methods of content capture to complement the system: using stand-alone personal computers to store snapshots of databases feeding dynamic sites, for example. In effect this is no different from I&A retaining and maintaining the machines capable of playing archived film or vinyl records, but it would still represent something of a "first" for digital archiving and preservation.

The working title for the system—the Legal and Historical Internet Archive—deliberately emphasizes that it is not intended as a source of reusable material. One of the main reasons for this is the lack of necessary rights information. The system will collect URLs at the point where they are published and made publicly available, which means that no copyright details will be harvested by the system since such sensitive information is deliberately not included in HTML metatags. Unfortunately, this also means that system users are technically able both to download and reuse stored files. There is currently no way of preventing this from happening, though users will be reminded at login that they have no authorization for downloads and warned that, in the event of a legal challenge over republished material, the system will have retained a record of all transactions.

The system will also be able to quarantine potentially libelous or defamatory content with access restricted to I&A system administrators and New Media's legal team via the use of a password. This eliminates the risk of content reuse during any period of investigation.

Content quarantined to a separate area of the system would then be reviewed and reinstated in the main system, marked for permanent exclusion, or simply deleted dependent on the outcome of the legal proceedings. Certain content—typically that provided by a third-party supplier or independent production company—may very well not be archived because of copyright issues. Though technically in-scope, the URLs will be flagged to prevent their being captured by the system.

Because the archive needs to serve different business needs, user permissions will be dictated by relevant business processes. At the top of the access hierarchy sits the "Approver" with control over all aspects of system management, including, most significantly, the ability to delete and permanently quarantine archived content. The "Administrator" will deal with routine administrative tasks and run business and management reports, while the "Legal User" can view quarantined content and the "Restricted" or "General User" is allowed basic access to content. In order to satisfy the requirements of the "Legal Users"—New Media's team of lawyers who would have to answer any public complaints about content on bbc.co.uk—system logins will be granted to a maximum of fifty users. Compare that to the pan-BBC access to all other archive collections with the workforce now in the region of 27,000 employees. There are benefits of such a "soft launch" beyond minimizing the risk of misuse of the system with illegal reuse of downloaded content. Any problems at implementation will affect relatively few people, and management reports will provide clearly focused information on usage patterns to support decisions on the system's further development.

Continuing the theme of "simplicity," the search mechanism will be restricted to queries based on URL and/or date and/or time conducted via a Web front end integrated with the standard internal BBC desktop. To that end, captured files will be stored as "Filename.FileExtension.TimeDateStamp" and in directories mirroring those of the original Web sites. It is important that the system's file structures are also designed to support future integration with federated search engines providing simultaneous access to I&A's multiple online resources. Query responses will return the nearest version to the requested URL based on date and time and encourage chronological navigation through to the "next" and "previous" versions of that URL. Where users specify only a date and time but no URL, the bbc.co.uk home page will be returned, and where only a URL is cited, all versions of that URL will be retrieved. All of this will be achieved within three seconds in at least 80 percent of cases and no longer than five seconds for the remaining 20 percent. Users will also be alerted to the status of quarantined or excluded content in response to search queries and reminded of the scope of the collection where requested URLs have not been captured. It is important to remember that Web pages will be returned "as published," and users will have the option of navigating away from the requested URL to any linked content within the archive.

At conception there were two possible options for the build of the system. The first was based on UNIX architecture, and the plan was to add disk arrays directly to the BBC's MCS, which contains all uploaded content ready for publication to the live Web servers. The proposal was for content to be copied directly to disk, stored in perpetuity and retrievable via a Web page hosted on the MCS. The second—and agreed—option used an Intel processor and Linux operating system not directly integrated with the MCS. The content would be received into a dedicated archive server delivered from the MCS as it would be to any Web server and retrievable via a Web page hosted on that archive server.

The attraction of the second option was that it would allow both the archive system and New Media's content production and publishing system to develop in isolation despite the dependency of the first on the second for the delivery of content. It would also allay New Media's fears of archiving encroaching on the creative environment and in any way dictating the means by which content is produced.

The only other issue for the system build was the need for the underlying data model to be compliant with the BBC-developed Standard Media Exchange Framework (SMEF).[3] The framework defines terms used in the content production, distribution, and broadcast chain across all media platforms including television, radio, and the Web and was initially designed to support the sharing of data across BBC networks and via BBC systems.

Project Management

The implementation of the system was directly managed by a project team comprising one project manager from BBC Technology (the system designers, builders, and supporters), another from I&A (representing the needs of the business), and the New Media Archivist (the customer). The project team was then accountable to a project board, which convened monthly for updates on progress and whose role it was to make decisions and resolve conflicts or issues. The board included the project team and was augmented by representatives from New Media (both editorial and legal) and the BBC's Business Technology Analyst responsible for the pan-BBC storage strategy.

At the outset the project team identified the major risks as follows:

- If the rate of content creation—and consequently data volumes—dramatically increases, then the solution will run out of storage capacity, originally estimated at 4TB over the initial three years of the system's implementation.
- The BBC could be in breach of legal restrictions if quarantined content was inadvertently made generally available.
- Stored content could be corrupted or lost in the event of catastrophic technical failure.

Only the last of the three was considered unavoidable, whereas ongoing monitoring of data-ingestion rates would allow for timely increase of available storage and, with rigorous data and system management, access to quarantined content should be preventable.

It had already been accepted that the agreed storage strategy—dictated by budgetary constraints—meant that, should two disks become corrupted, then data on all disks, and consequently the whole archive, would be lost. Whereas RAID 1 Mirroring was the preferred—and more secure—option, it would have doubled storage costs compared with the alternative: three disk-arrays set up using RAID 5 Strip Set with Parity, which could cope with the loss of one disk (but not more) with no effect on system performance or stored data.

Unfortunately, there were a number of initially unidentifiable risks that had an impact on the projected eight-month time scale:

- The sale of BBC Technology—a BBC subsidiary—led to industrial action and resulted in the project manager's resigning prior to implementation.
- The unique system-build made unusual demands of a standard application: the modified version of Apache Web server unexpectedly required the development of a secondary layer of code in order to provide the necessary functionality. This meant eight weeks of redevelopment work and the drafting in of a second developer.
- The hardware failed to perform as required, and it took a second storage box and several weeks of investigation before the problem was identified: the mechanism for managing the ingestion of content could not process the 17,000 hourly transactions and so caused an unacceptable backlog in the queue for updates to be published to bbc.co.uk.

The positive consequence of the unforeseen delays was the decision to move from Network Addressed Storage (NAS) to the BBC's newly established Storage Area Network (SAN), designed to rationalize storage management across the corporation and thereby reduce costs. The benefits for the Legal and Historical Internet Archive System were twofold:

1. Content would be ingested directly into the archive via a fiber card, thereby reducing the number of transactions for processing, speeding up data delivery, and avoiding congestion in the bbc.co.uk publishing queue.
2. Backup would be to the more resilient tape rather than disk, and the risk of data corruption would be reduced.

But there was one caveat. While the corporate advantages of central storage were inarguable, it was important to establish the difference between the administration of "active storage" to meet day-to-day business demands and that required for the archive management of content intended for retention in perpetuity. At the very least this meant ensuring data integrity throughout its lifespan and enabling possible future emulation or migration for the purposes of digital preservation.

The Service Level Agreement with BBC Technology—which became Siemens following the sell-off in October 2004—was designed to cover the level of system support provided and thereby went some way toward safeguarding the collection through the promise of monitoring and managing firewalls, general system security, anti-virus and other protective measures, and general hardware maintenance. It was also agreed that any requirement for increasing the system's functionality over the three-year agreement would be treated as an official "Change Request."

Post-Implementation and Future Development

Implementation marked the start of an ongoing review process: Did the system fully meet the requirements for a legal and historical archive? Should the collection be subject to a "Selection and Retention Policy" or remain as comprehensive as is technologically possible? How useful is the system to content producers and how can it be made more useful? Should it be integrated with content production systems? Can it support the reuse of content and how? How far does it fit with I&A's digital archiving and preservation strategy and with the BBC's broader technology roadmap? The data from regular technical and management reports were designed to help answer some of those questions.

The system's basic performance—its degree of success at capturing specified content and its search response times—would be monitored by Siemens' support team as part of the Service Level Agreement and judged against agreed benchmarks. More useful for the ongoing development of the system were the patterns of usage. On a weekly basis reports include the following:

- For individual logins: number and length of search sessions; total number of searches both successful and unsuccessful; and number of and details of downloaded files
- For legally quarantined or excluded content: number of URLs placed into quarantine; length of time URLs held in quarantine; number of URLs to be quarantined in perpetuity; number of URLs reinstated into the main collection; and number of URLs excluded from the capture process and/or deleted
- For general system usage: number of concurrent users

Aside from the development plans made based on the use and performance of the system, there are several things that need to happen if it is to fully support the archiving and preservation of bbc.co.uk. Digital collections need preservation strategies, but there are as yet no answers to the emulation versus migration question. Arguably, neither option preserves the original integrity of archived data, but each at least assures the recording of a digital memory. Emulation retains content to be replayed, rebroadcast, or retransmitted in its original format but via an imitation of its original technical context, while migration transfers that content from one format to another or from one storage device to another, essentially altering its original characteristics. And for that reason it was made clear throughout the project that the ingestion of individual files, the design of file structures, and the adopted storage strategy must support any future digital preservation process by allowing either blanket access to large volumes of data or focused access to specific file types.

Whether as a "Change Request" during the first three years of its implementation or beyond that initial period, the following requirements have already been highlighted for the next phase of the development of the Legal and Historical Internet Archive System:

- Extending search capabilities to include multilingual free-text searching across all stored content
- Integration with I&A's intranet—research.gateway—to support federated searching across all online archives and information sources
- Retrospective ingestion of existing bbc.co.uk content from a range of sources and on a range of formats: screenshots and complete pages held on servers and/or solid media and tape archives kept as backups to bbc.co.uk
- Extended archive coverage to include content published to mobile or other Internet-based platforms, including Wireless Application Protocol (WAP) services; home pages and indexes from any sites published via the News Online content production system and including the World Service, Sport, and Weather; dynamic- and user-generated content including message boards; rich media where content is unique to and/or commissioned by bbc.co.uk

The other main area for development would be the standardization, inputting, and management of metadata. The system is dependent on the delivery of content via the external publishing mechanism. This means that pages carry no metadata beyond that required by search engines for indexing purposes: it has little value for internal business use beyond the legal and historical remit of the first version of the system.

Post-production cataloguing of over two million constantly updated Web pages is clearly not feasible, but perhaps neither is mandating producers to complete metadata records at the point of content creation. An alternative might be to network the archive with content production and/or media asset management systems, which already contain metadata gathered during the creation and publication process. Is it not better to link to sources of information than to input that information a second time?

Though it does not solve the problem of standardizing metadata—I&A is far from imposing even the blandest set of required fields across quite disparate content production areas—it at least begins the process. But I&A will have control over the retrospective addition of metadata to archived content and is considering the use of Dublin Core (http://dublincore.org) as the basis for developing a standard not only for bbc.co.uk but for implementation across all media. In particular, the recently approved "Provenance" field, providing a "statement of any changes in ownership and custody of the resource since its creation," will help ensure data integrity throughout its life cycle and, it is hoped, in perpetuity.

Adopting an internationally recognized standard would not only support the internal and external sharing of content but the provision of future public access to the system where technologically possible without compromising the BBC's own network infrastructure.

New Media Archiving

Archiving and preserving bbc.co.uk is no different from archiving other New Media services: each is interactive and each comprises content from different sources in different formats with different drivers for short- and/ or long-term retention. And they each require appropriate storage and preservation strategies. Archiving interactive TV—essentially the result of a collision between broadcast and computing technologies—means capturing audio and video streams, text, graphics, presentational templates, and software applications. Mobile services—including phones and personal digital assistants (PDAs)—could mean the retention of physical devices or, if not, the acquisition of appropriate content emulators.

With such rapid change in the development of platforms for content distribution, it is essential to work as closely as possible with the creators of that content. Digital archiving and preservation are not just about what to do with material when you have it but also about facing the challenge of how to acquire it and, as far as possible, attempting to influence its creation and delivery. That said, it is unlikely that archive requirements could ever directly influence production methods, though the Internet Archive's Brewster Kahle (Rein, 2004) believes that is exactly what should happen. That does not mean, however, that content producers should not have an understanding of those requirements and an agreement that they need to be addressed as early in the production process as possible.

The situation is not as bleak as it might sound. The BBC's New Media Department has recently initiated two archive-related projects. The "Creative Archive" will allow users to download clips from archive programs, provide tools for editing those clips, and encourage peer-to-peer sharing of content. According to a recent press release (British Broadcasting Corporation, 2004), one of the main objectives is to "pioneer a new approach to public access rights in the digital age," adopting the Creative Commons (http://creativecommons.org) model already prevalent in the United States. And there are internal benefits for the BBC: the greater the demand for access to archive material, the faster the growth of its digital collections. This certainly proved true for News Online's "On This Day" (http://news.bbc.co.uk/onthisday/), which not only provides public access to otherwise inaccessible news footage but, in its demand for content, supports digital preservation through the migration of analogue material to digital formats for inclusion on the site. Both projects—as well as the planned Interactive Media Player designed as a portal to the previous seven days of BBC television and radio programing via the Web—will feed directly into the Legal and Historical Internet Archive System. Or perhaps not directly. It might be that audio/visual content of this kind could be archived prior to distribution rather than via the external Web publishing mechanism. This would relieve pressure on the ingestion process and ensure the simultaneous delivery of all associated metadata.

The Creative Archive is a good example of what Cory Doctorow (2004) described as the "remix culture" and, to support that culture, archiving and preservation needs to provide access to decontextualized content to feed the growing demand for reuse of material in any number of ways on any number of platforms. While that is not the objective of the first iteration of the Legal and Historical Internet Archive, it is certainly on the agenda for further phases of development.

There is much that could be said in conclusion and not least about the lessons learned so far. For instance, it was crucial to continually remind the project managers of the business drivers for the system's implementation. While a lack of detailed technical knowledge meant occasional "scope creep" and infeasible demands being made on the developers, it was important to maintain a focus on the unique requirements of a server-based archive—not least the need for long-term data integrity and opportunities for as yet undefined digital preservation strategies. That said, it is apparent that traditional archive models developed for a linear, analogue, and physical world are not always appropriate in a technologically advancing one and that there are benefits to be had from embracing that technology. Resources can be directed away from more mundane collection management tasks to tackle the issues of digital curation and preservation.

But decisions need to be made and policies formulated without having all the answers: Which are the best archive formats? Should content be selected for retention, or is that too labor intensive in a server-based environment where Moore's Law dictates that storage costs will keep falling? How will issues be resolved in relation to the archiving of software and hardware? Anyone engaged in digital preservation must accept that data cannot always be saved or accessed in its original format; must be prepared to develop emulated environments or processes for frequent data migration; must work as closely as possible with information technologists and content creators; must openly share experiences and knowledge through national and international bodies and organizations; and must appreciate their own business context—for the Legal and Historical Internet Archive System this includes the challenges of a fast-changing broadcast environment—while remaining focused on what does not change: the need to capture digital material and archive, preserve, and make it accessible.

Notes

1. The British Broadcasting Corporation is a public service broadcaster operating on the basis of a Royal Charter and Agreement and funded by a licence fee. It is consequently

accountable to the British people to whom it delivers local and national television and radio output, including digital and interactive services. It also produces more than two million Web pages and via the BBC World Service provides radio programming in forty-three languages (http://www.bbc.co.uk/info/).

2. OFCOM is the regulator for the UK communications industries, with responsibilities across television, radio, telecommunications, and wireless communications services (http://www.ofcom.org.uk/).
3. The SMEF data model is being marketed by the BBC as an industry standard available without charge but governed by a no-signature license (http://www.bbc.co.uk/guidelines/smef/).

References

British Broadcasting Corporation. (2004). *BBC Creative Archive pioneers new approach to public access rights in digital age* [Press release]. Retrieved February 28, 2005, from http://www.bbc.co.uk/pressoffice/pressreleases/stories/2004/05_may/26/creative_archive.shtml.

Department of National Heritage: Broadcasting. (1996a). *Copy of the Agreement Dated the 25th Day of January 1996 Between Her Majesty's Secretary of State for National Heritage and the British Broadcasting Corporation* [Electronic version]. Retrieved February 28, 2005, from http://www.bbc.co.uk/info/policies/charter/pdf/agreement_text.shtml.

———. (1996b). *Copy of Royal Charter for the Continuance of The British Broadcasting Corporation* [Electronic version]. Retrieved February 28, 2005, from http://www.bbc.co.uk/info/policies/charter/pdf/charter_text.shtml.

Doctorow, C. (2004). *Written testimony to Select Committee on Culture, Media and Sport* [Electronic version]. Retrieved February 28, 2005, from http://www.eff.org/IP/BBC_CMSC_testimony.php.

Graf, P. (2004). *Report of the Independent Review of BBC Online* [Electronic version]. Retrieved February 28, 2005, from http://www.culture.gov.uk/NR/rdonlyres/395685B7-0373-49E2-BD1B-CE9F2E581E99/0/BBConlinereview.doc.

Queen's Printer of Acts of Parliament. (1996a). *Broadcasting Act 1996, Chapter 55* [Electronic version]. February 28, 2005, from http://www.hmso.gov.uk/acts/acts1996/1996055.htm.

———. (1996b). *Defamation Act 1996, Chapter 31* [Electronic version]. Retrieved February 28, 2005, from http://www.hmso.gov.uk/acts/acts1996/1996031.htm.

———. (1998). *Data Protection Act 1998, Chapter 29* [Electronic version]. Retrieved February 28, 2005, from http://www.hmso.gov.uk/acts/acts1998/19980029.htm.

———. (2003a). *Communications Act 2003, Chapter 21* [Electronic version]. Retrieved February 28, 2005, from http://www.legislation.hmso.gov.uk/acts/acts2003/20030021.htm.

———. (2003b). *Legal Deposit Libraries Act 2003, Chapter 28* [Electronic version]. Retrieved February 28, 2005, from http://www.legislation.hmso.gov.uk/acts/acts2003/20030028.htm.

Rein, L. (2004). *Brewster Kahle on the Internet Archive and People's Technology.* Retrieved February 28, 2005, from http://www.openp2p.com/pub/a/p2p/2004/01/22/kahle.html.

Cathy Smith, New Media Archivist, Room BC3 D5, White City: Broadcast Centre,Wood Lane, London W12 7TP, cathy.smith@bbc.uk. Cathy Smith is New Media Archivist at the British Broadcasting Corporation in London, working in the Information and Archives Department. She holds an M.S. in Information Science from City University, London, and has been in her current position since 2000. Responsible for establishing archiving practices and procedures for the BBC's New Media output, one of her main projects is the implementation of an Internet archive system designed to capture content published to bbc.co.uk.