World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

May 28, 2005

Code Number: 035-E
Meeting: 150 *SI* - ICABS (IFLA/CDNL Alliance for Bibliographic Standards)

Collaboration in digital archiving in the UK

Caroline Brazier
Head of Collection Acquisition and Description, The British Library, UK

Abstract
This paper gives a brief outline of current collaborative developments in the UK towards the establishment of a national digital published archive. It looks at recent changes in legislation in support of the creation of a digital archive, building on the concept of the National Published Archive for print publications. It discusses collaborative pilot projects designed to further our knowledge and experience and examines the challenges they face. It also considers how vital a collaborative framework is to ensure future progress, not just between the libraries which will be responsible for the national digital archive, but for the publishing industry too.

Introduction

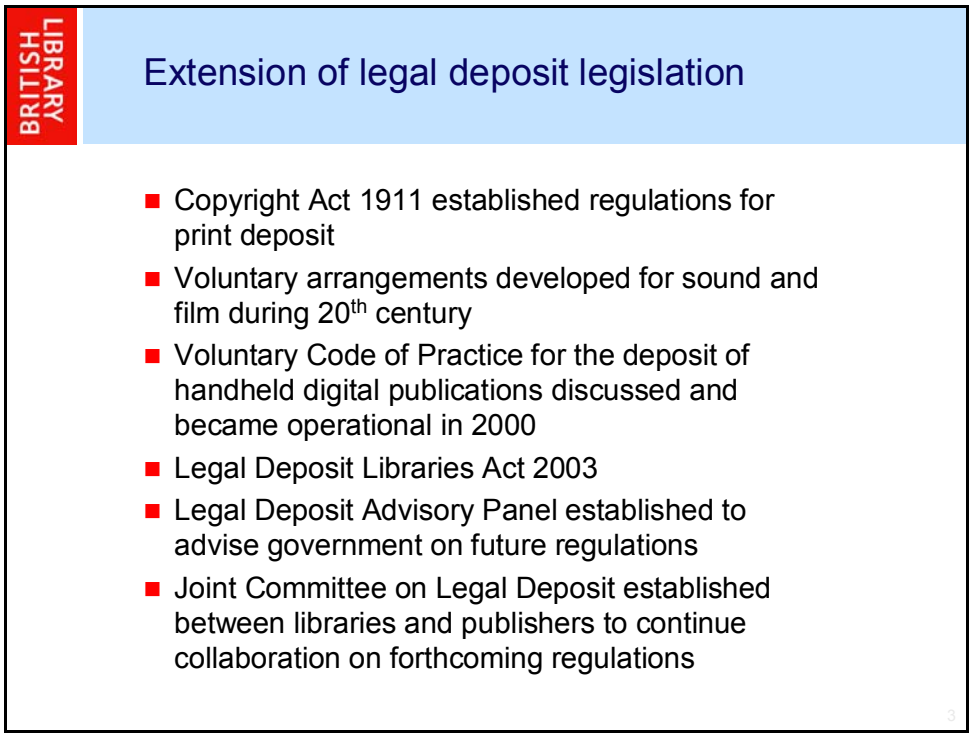
There has been a long tradition in the UK of collaboration between the six UK legal deposit libraries in the development and management of the national published archive, built around a framework of legal deposit legislation which has been relatively stable since 1911. There has traditionally been close cooperation over collection development and operational issues around collecting. There has also been more formal collaboration over the running of the "Shared Cataloguing" programme, where the six libraries – the British Library, National Library of Scotland, National Library of Wales, Bodleian Library, Oxford, Cambridge University Library and

Trinity College Dublin – have formally shared the cataloguing of the legal deposit monograph intake, all thereby contributing to the British National Bibliography.

The rapid increase and development in digital publishing, starting with offline formats such as CD, during the 1990s led to rising concern that the national published archive would increasingly fail to reflect the nature and content of UK publishing unless legal deposit legislation was developed to allow the libraries to collect, preserve and make available digital formats as well as print. A brief exploration of legislative, technical and operational developments over the last 10 years shows how far we have travelled towards meeting our objective of a comprehensive digital archive for the UK.

The extension of UK legal deposit legislation

The 1911 Copyright Act established the regulations under which print has been deposited for the last 94 years. As non print publication formats emerged throughout the 20th century, and recorded sound and visual images became publicly available, these were collected under voluntary schemes and not covered by formal extension of deposit legislation. The Libraries grew increasingly concerned during the 1990s about the rapid proliferation in digital publication formats, first in handheld, offline formats such as CD and then in online formats. Discussions were held with representatives of the UK publishing community during the late 1990s and a Voluntary Code of Practice for Deposit for Offline Publications was agreed and came into operation in 2000.

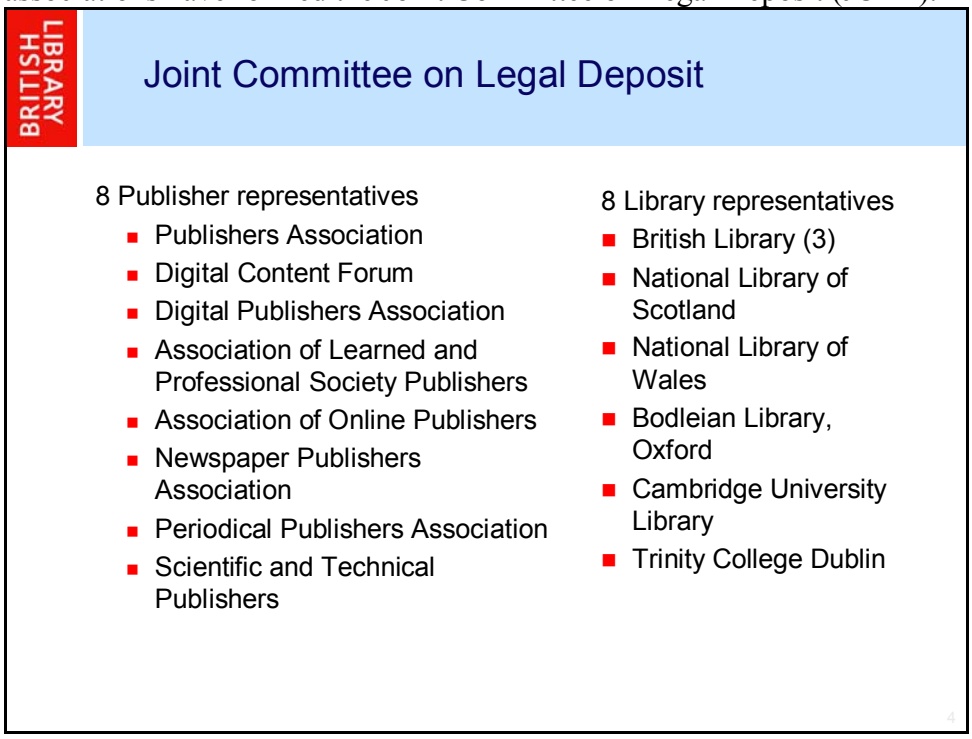


Formal representations to government continued and resulted in the passing of the Legal Deposit Libraries Act in 2003. This Act extends the concept of legal deposit to the non print world, although it confirmed the continuation of the voluntary schemes for sound and film. The 2003 Act is framework legislation and does not specify detailed arrangements for deposit of different digital formats. These will be confirmed

by a forthcoming series of Statutory Regulations to be issued under the Act. A Legal Deposit Advisory Panel is currently being established to advise government on the nature of those regulations. Both the Legal Deposit Libraries and the associations representing different aspects of the publishing industry will have representation on the Advisory Panel.

Collaboration between Libraries and Publishers

At present our non print deposit and harvesting operations are conducted under voluntary codes of practice or, in the case of websites, express granting of permission. To oversee the voluntary codes and to contribute towards the development of formal regulations under the new act, the Legal Deposit Libraries and several UK publishing associations have formed the Joint Committee on Legal Deposit (JCLD).



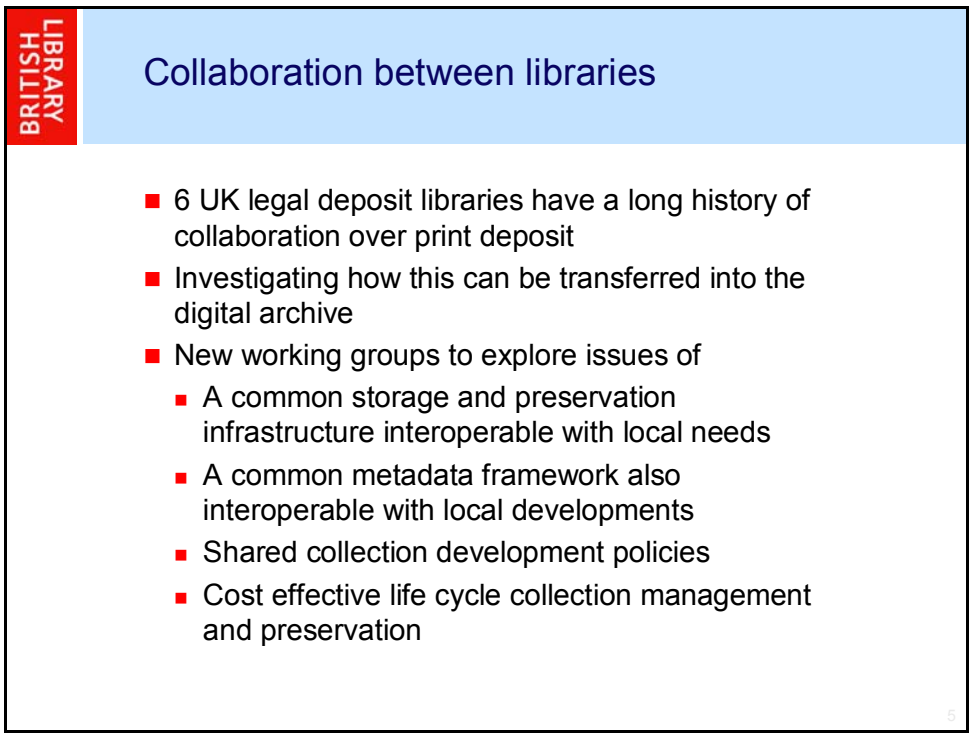
Through its bilateral discussions, the Joint Committee explores the issues for libraries and publishers surrounding the deposit of different digital formats, and seeks to reach joint recommendations for proposal to the Advisory Panel. At present its focus has been on the workings of the present voluntary scheme for offline, as regulations covering handheld formats are expected to be the first addressed by government under the new Act.

Another major area of joint activity is on the development of a voluntary scheme for deposit of e-journals, where a pilot is currently being established. In both offline and e-journal work to date the issue of access, and the danger this may pose to publishers' ability to sell their publications commercially, has been one of the major areas for discussion and compromise.

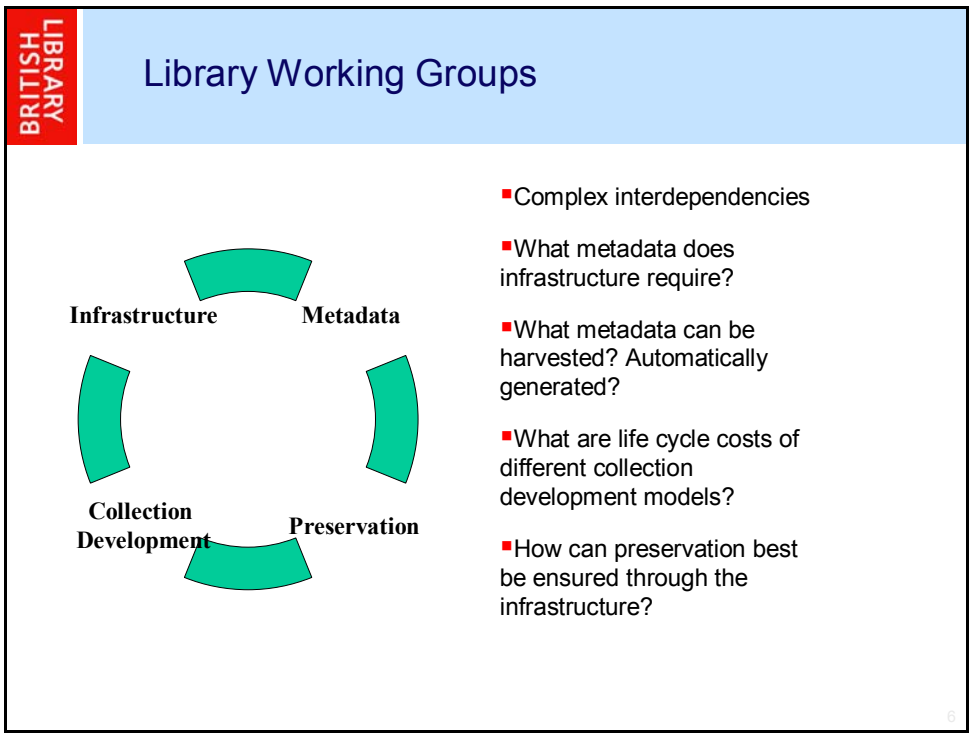
The Committee is also exploring the legal issues of territoriality, to try to establish workable definitions of "UK publishing" in a global publishing environment. Once offline and e-journals are firmly established, the Joint Committee will shift its focus to

explore the issues surrounding deposit of databases, datasets and the harvesting of UK websites.

Collaboration between the legal deposit libraries
Collaboration between the libraries is not new and is already strongly established in the development of the print National Published Archive. However, due to the shift in emphasis from print to digital which the new legislative framework has brought, the Libraries have established new working groups to develop the digital archive framework.



These working groups on Infrastructure, Metadata, Collection Development and Preservation are exploring the issues and problems involved in setting up and managing a fully operational shared digital archive. It has to be capable of ingesting, processing, storing, preserving and facilitating appropriate access to content, which is likely to be far more restrictive than normal commercial licensing arrangements.

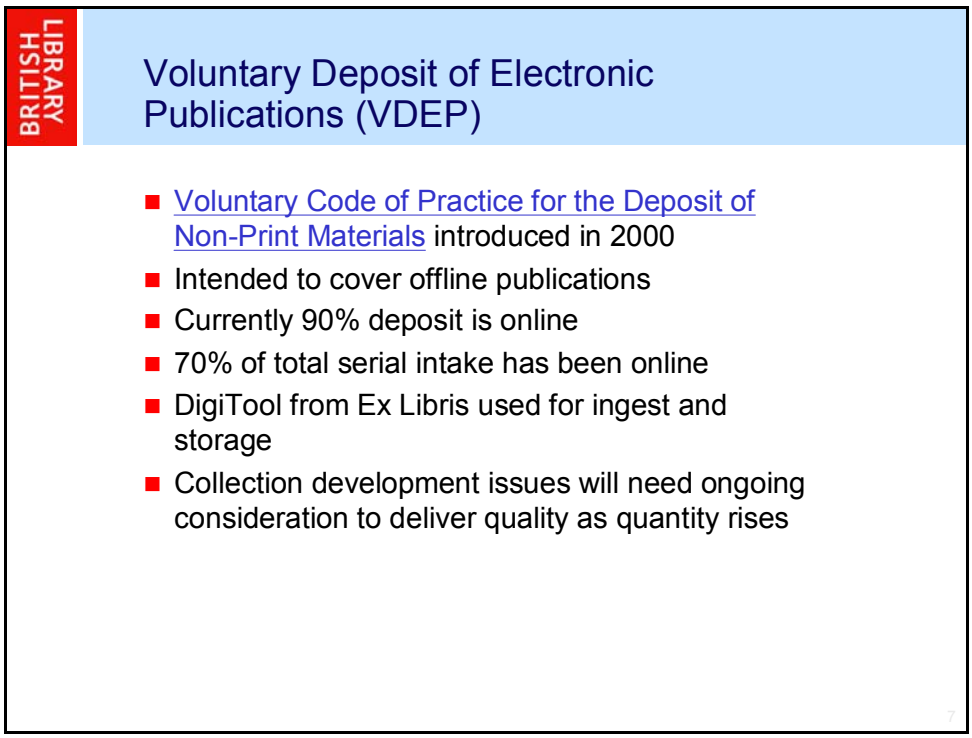


The discussions so far have centred on the technical infrastructure and the requirements of a shared infrastructure to support different institutional aspirations and local technical requirements in the 6 libraries. Discussions on an appropriate metadata framework, covering not only descriptive metadata, but also technical, administrative, structural, preservation and digital rights are also highlighting the complexity the new system will need to be capable of addressing. The metadata framework must not only underpin resource interoperability in resource discovery and collection management between the 6 libraries in digital archive management, but also support interoperability between a central digital legal deposit collection and the 6 separate, local digital archive collections which each institution is developing in parallel.

Collection Development questions are also of paramount importance in defining the requirements for a future shared infrastructure. The Libraries must define their intentions with regard to the nature and scale of collection building they envisage for the national digital published archive. The sheer volume and dynamic nature of digital publishing, particularly on the web, will drive the Libraries to a far more selective approach than has operated for print.

In pragmatic terms, the actual degree of selectivity must be informed by the availability of resources to handle the full life-cycle cost implications of non-print material, not only for initial ingest and metadata creation but also for longer term preservation. So the work of the four committees is very much interconnected.

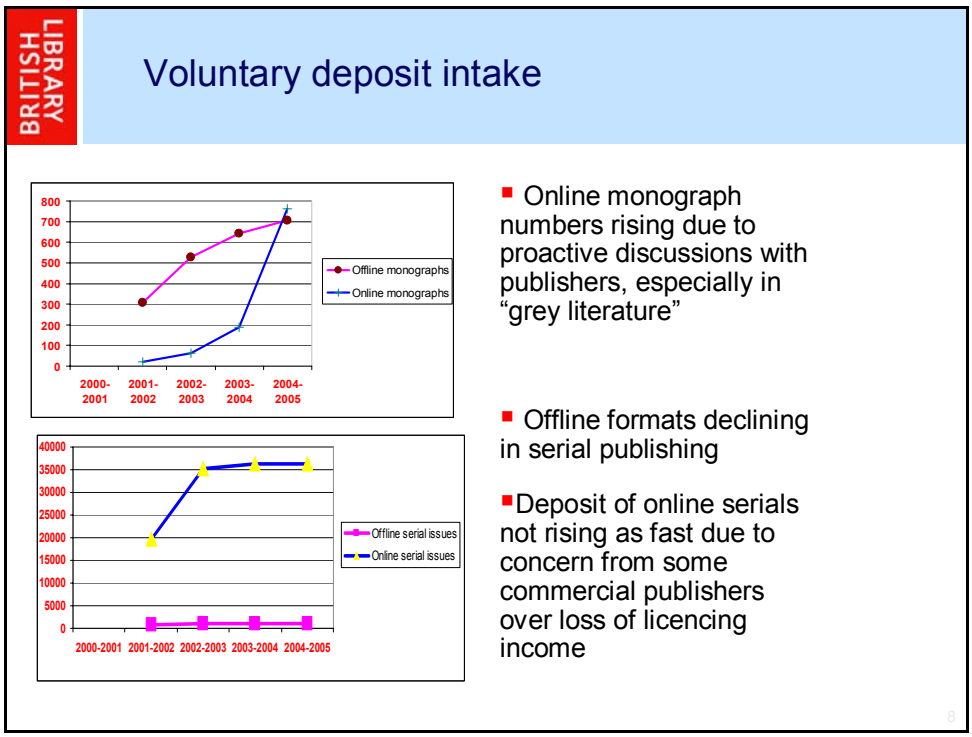
Voluntary Deposit of Electronic Publications



The current Voluntary Deposit arrangements have been operating since 2000. There are as yet no shared operations, and slightly different arrangements are in place in each of the 6 libraries. This paper focuses on current practices in the British Library.

The Voluntary Code of Practice for the Deposit of Non-Print Materials[1] , which became operational in 2000, was intended as an arrangement to cover offline publications. However, we have found in practice that a lot of online materials have also been deposited voluntarily. The proportion of online to offline continues to grow and in the average monthly intake in 2005 over 90% of material is in an online format. Over 70% of the serial content has been deposited in online formats, though not from the large scale academic and research publishers. Much of the operation over the past 5 years has gone into establishing secure storage and management processes for this online material.

[1] http://www.bl.uk/about/policies/codeprac.html



Handheld publications, such as microforms or CDs, are processed in line with our existing print based processes. Online materials have gone through several phases since 2000. Initially, much online material was received via email and, while we had concerns over the research value of some of the content, the initial policy was to accept everything under this pilot phase to gain experience in the issues of handling online deposit. Everything was therefore accepted into the pilot scheme, but only catalogued if it was felt to have content likely to be of long term research value.

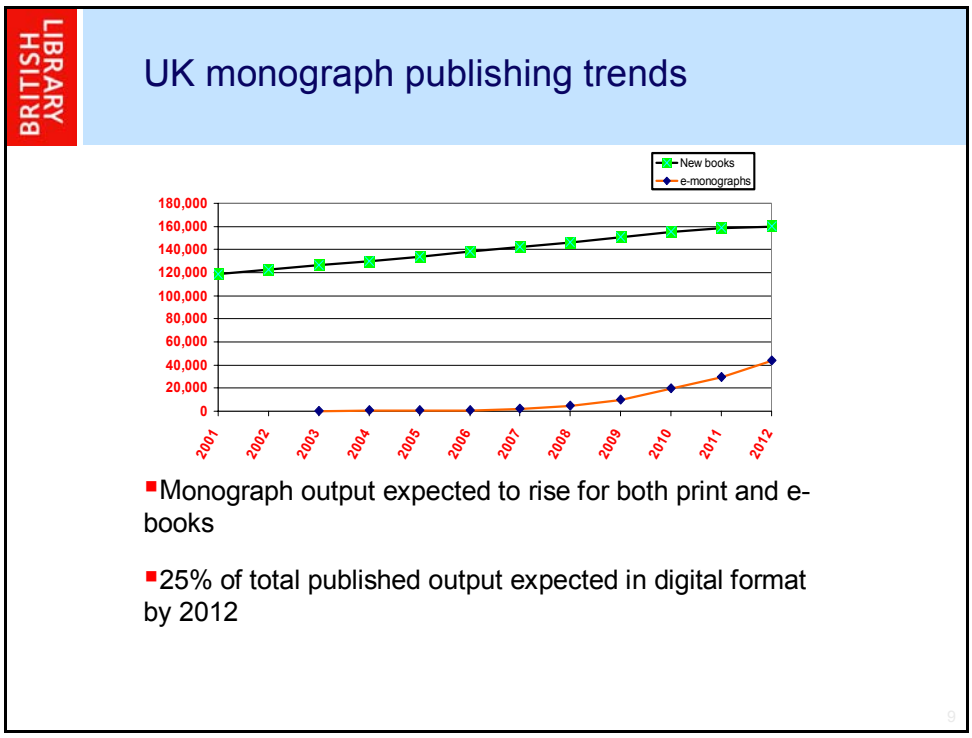
Storage needs quickly outgrew the capacity of the original email server, so in the next phase the content was burned off onto CD-ROMs as a temporary means of storage, while the Library investigated long term digital object management and storage solutions. In 2002/2003, when the British Library was selecting an integrated library management system for its day to day collection management operations, the opportunity was also taken to investigate the digital storage solutions available from the leading library software vendors, in parallel with the main procurement. The decision was taken to use the DigiTool product from Ex Libris to handle the next phase of ingest and storage of the voluntary deposit material. British Library staff worked with Ex Libris developers to refine the DigiTool product to semi-automate some of the routine ingest processing and to enable it to work as effectively and share metadata with the Aleph integrated system, used for print collection management.

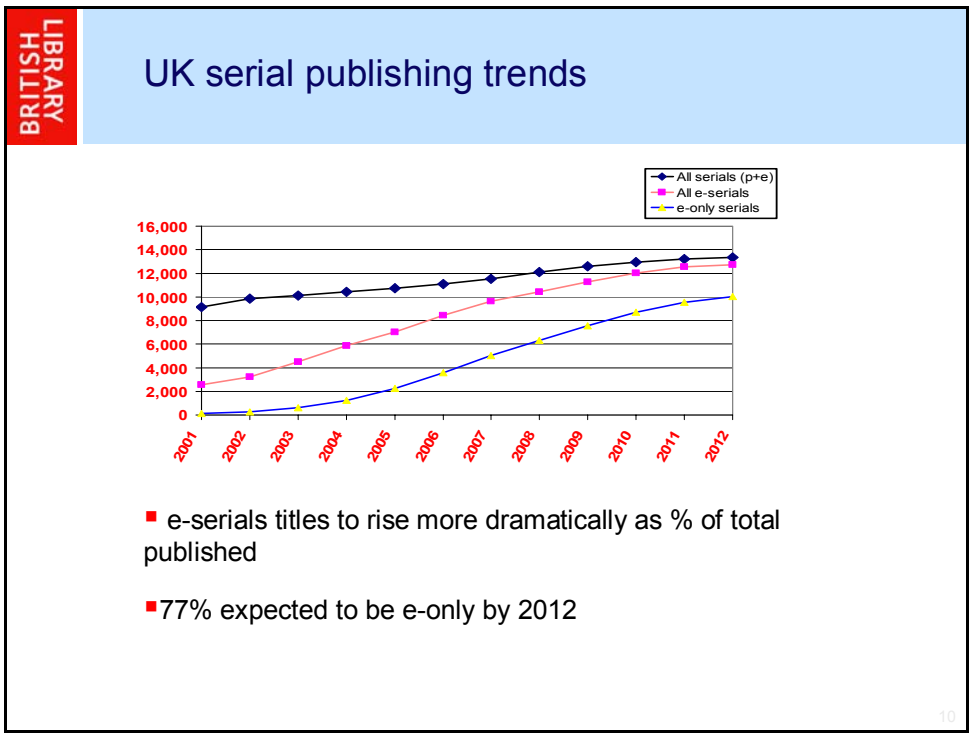
The present arrangements will be reviewed in line with the ongoing development of the British Library's main Digital Object Management (DOM) system and the overall development of the infrastructure for the 6 legal deposit libraries.

The question of collection development is also one the 6 libraries will have to focus on. The British Library has already had to change its internal collection development policy for online voluntary deposit as much material was not felt to be of sufficient long term research value to collect on a large scale in perpetuity (e.g. online birthday cards, digital "vanity" publishing, advertising materials), while other materials clearly

did not originate in the UK or were intended for internal audiences only (e.g. internal company communications).

The predicted scale of digital publishing also suggests the libraries will have to be selective. Research into the future scale of online publishing, carried out for the legal deposit libraries in support of the legislative process in 2003, shows a considerable increase in online serial and monograph publishing. And this does not cover the growth in freely available web publishing formats.





Electronic Journal pilot

The majority of content received under VDEP has been serial in nature. However, the mainstream academic and research journal publishers have not been willing to deposit voluntarily due to concerns over access and digital rights management. The Joint Committee has therefore made these issues a priority for immediate review between publishers and libraries to explore the issues and concerns on both sides.

A voluntary pilot is being set up with c.20 academic journal publishers offering an initial total of c.75 – 100 titles. The intention is to explore the issues involved in ingesting, storing and preserving a variety of formats and content types from a range of service providers, to inform future work on regulations to control e-journal deposit. The participating publishers and their publishing associations are acting as proxies for the broader publishing industry and will be disseminating results back to the rest of their community. It is planned to hold a workshop once the results of the initial pilot work on ingesting are complete. Once again, a major issue to be resolved is over appropriate ingest models and metadata transfer with and/or extraction from deposited content to ensure cost effective archive management between the legal deposit libraries.

Electronic journals pilot

- Pilot to explore technical issues involved in cost effective ingest and storage
- Participant publishers sought as volunteers with the help of the relevant publishers associations
- 20 academic and research publishers offering 75-100 titles involved in phase 1
- Results to be disseminated back to the industry via the participating publishers and workshops

11

UK Web Archiving Consortium (UKWAC)

A major limitation on the current voluntary deposit operation is that if we are alerted by a web publisher to a new website or changes to an existing site, we do not have the technical infrastructure to harvest automatically. While we can manually harvest individual documents or files, entire websites with dynamic links have had to be excluded for pragmatic reasons. This gap is now being closed with the establishment of the UK Web Archiving Consortium (UKWAC).

UK Web Archiving Consortium

Members

- The British Library
- National Library of Scotland
- National Library of Wales
- The National Archives
- Joint Information Systems Committee
- The Wellcome Library

Aims to harvest 6,000 UK sites over 2 years

Uses PANDAS software from National Library of Australia as basic harvesting tool

UKWAC site launched in March 2005 (Website: www.webarchive.org.uk)

12

The UKWAC Consortium members are The British Library, National Library of Scotland, National Library of Wales, The National Archives, Joint Information Systems Committee (funded to support developments in Higher and Further education across the UK) and the Wellcome Library. The objective is to harvest 6,000 UK websites between 2005 and 2007, in support of individual institutional objectives, but also the whole contributing to the development of a UK web archive.

The present technical infrastructure to support UKWAC operations is based on the PANDAS software from National Library of Australia as the basic harvesting tool. The UKWAC site was launched in March 2005 (Website: www.webarchive.org.uk). The partners will share the costs relating to the establishment of the infrastructure between them, including hosting the service and providing technical support.

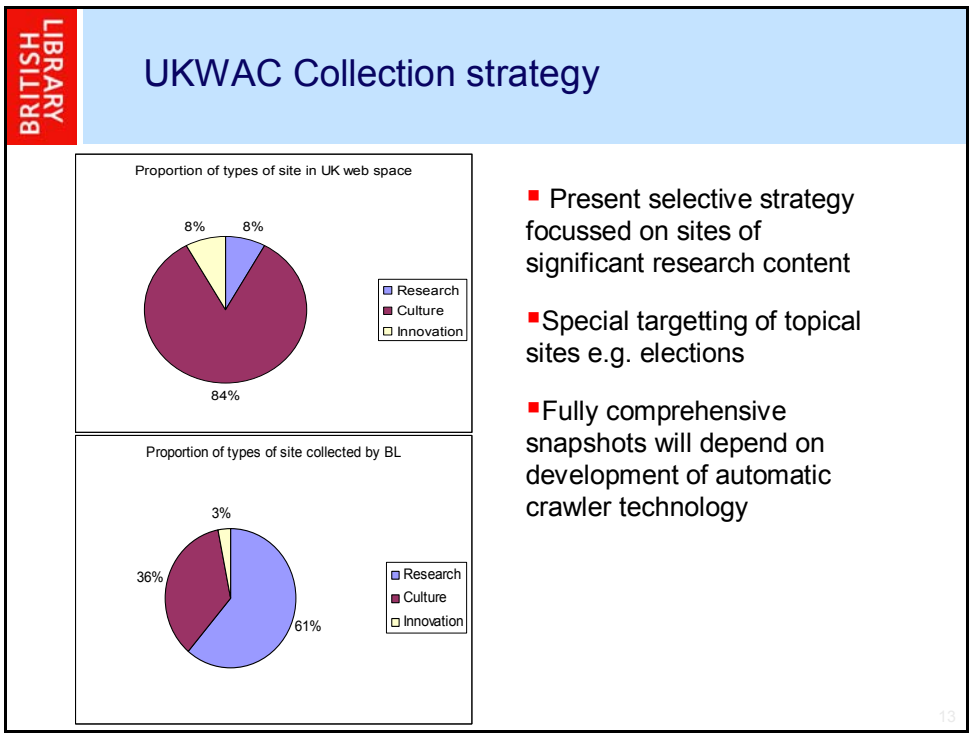
UK websites can be defined in several ways, but the relevant definitions for UKWAC are sites

- with a .uk domain name
- hosted in the UK (.com, .org etc)
- owned by UK organisations
- containing significant UK intellectual content
- with intellectual content of particular interest to the UK

These may or may not coincide with the legal definitions of territoriality which will underpin the future development of legal deposit regulations. However, as all harvesting under UKWAC is being done on the basis of express permissions granted by site owners, this is not seen as a major issue under the present project.

There are estimated to be over 4 million "UK" sites, so it is essential that selection criteria are clearly defined to ensure the long term needs of researchers are met. The British Library's collecting policy for UKWAC is based on building an appropriate balance between sites that are concerned with

- research
  e.g. sites hosted by universities; charities; campaigning organisations; government bodies
- culture
  e.g. sites representing British cultural diversity/ significance; key events of national life; topicality
- innovation
  e.g. award winning sites; sites illustrating web information, communication and training strengths



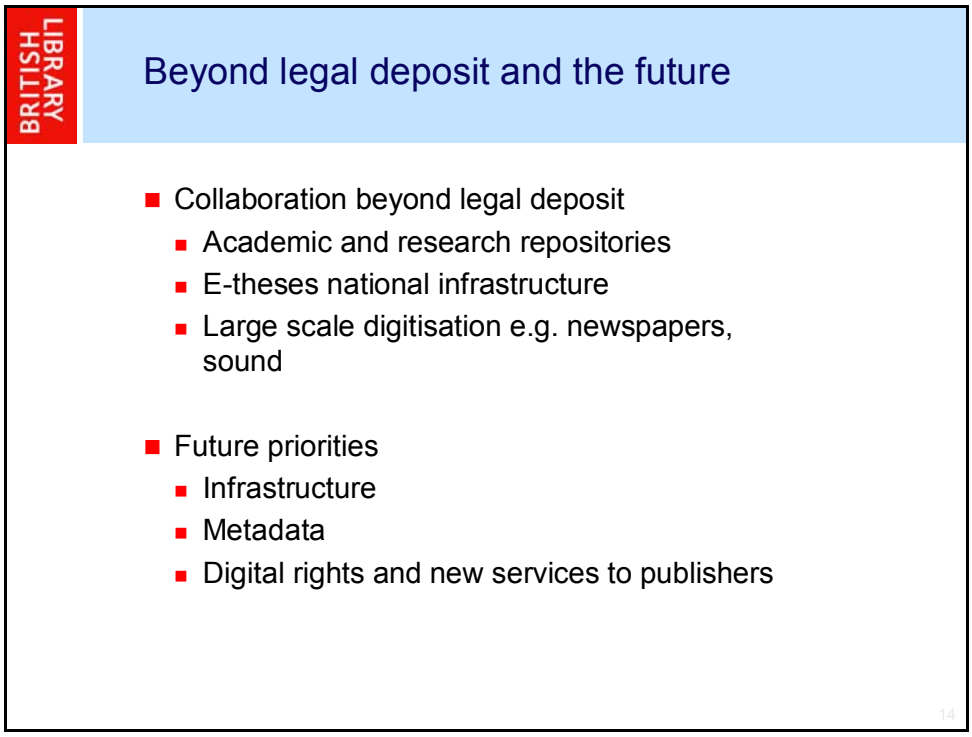
In addition to sites selected for regular harvesting, the UKWAC project also targets sites covering high profile topical or newsworthy aspects of UK life, but with a limited lifespan (e.g. Tsunami sites, 2005 general election sites).

This work is in early stages of development and both technical and operational lessons are being learned. Work on getting permissions can be slower in some cases than originally envisaged. The question of how far permissions can extend (e.g. in harvesting linked content) is also an issue. The frequency at which sites can be harvested at present also creates problems with sufficient coverage of the deep web, including databases, datasets and dynamic material.

This model of highly selective harvesting and archiving will not be able to scale up sufficiently to enable the creation of a comprehensive web archive. We are also collaborating with the International Internet Preservation Consortium (Website: http://netpreserve.org) to work on the development of a smart crawler, with the capacity to harvest periodic snapshots of the entire UK web presence. To cover content which has already gone (and the life span of the average website has been calculated at 44 days) we will also consider purchasing back files of relevant UK webspace from the Internet Archive (Website: http://www.archive.org), once permissions and regulations in the UK are firmly established.

UKWAC is intended to run for 2 years, when it will be evaluated and recommendations made on how to proceed with web archiving on a long-term, scalable basis. But in the meantime we hope to have developed an archive of web based material, which will form a significant future component of a national digital archive and to have shown that collaboration is a successful way to share the task of building an archive of web-based materials.

Future developments



There is a lot of activity and development under the various aspects of digital archiving in the UK, which I've covered. I have focused on the work in support of digital legal deposit developments, but there is also ongoing work on development of academic and research institutional repositories, the development of a national infrastructure for e-theses and large scale digitisation of newspapers and sound resources which will also contribute to the UK's national map of digital archiving.

There is clearly still a long way to go. But in the meantime we will continue to focus on 3 main areas. These are firstly the development of the tools and infrastructure for ingest, storage, preservation and resource discovery of legal deposit digital content.

Secondly, we will continue our work on cost effective creation, extraction or generation of metadata which will underpin our ability to manage the content of the national digital archive in the future.

And last, but by no means least, is the development of policies on digital rights and access to archival content. Ongoing collaboration with publishers and other content producers is vital to underpin future operations. We need to explore ways to ensure a national digital archive does not pose a threat to publisher interests, but offers them the opportunity to develop vital new services for the research community.

16 April 2005