

## Implementing the Draft Regulations for Non-Print Legal Deposit

This document has been produced by the British Library, with review comments from all the Legal Deposit Libraries, and provides technical information for publishers, as part of the consultation process, about how the Legal Deposit Libraries (“deposit libraries”) expect to manage online material that is harvested or deposited.

It describes the systems, processes and structures that are currently in place or which the deposit libraries plan to implement when Regulations first come into effect. In future years it is probable that some of these may be changed to improve effectiveness, to enable operating efficiencies, or for new technological environments. In all cases, any new systems, processes or structures will comply fully with the Regulations, and the deposit libraries will seek to maintain a process of dialogue with publishers and their representatives.

For the purposes of illustration, this document assumes that the final form of Regulations as enacted by Parliament will be as drafted in the Government’s consultation; if not, the deposit libraries will make adjustments as necessary in order to comply with the actual Regulations. This document also makes an assumption that the Secretary of State will be satisfied over the equivalency of laws in the Republic of Ireland pursuant to section 13 of the Act, and therefore that the Library of Trinity College is authorised to participate in the arrangements described.

### **Introduction - Shared Technical Infrastructure**

The deposit libraries have been developing a shared technical infrastructure for legal deposit. This is based upon the Digital Library System first developed by the British Library and now supported in partnership with the National Library of Scotland and National Library of Wales. The Bodleian Library Oxford, Cambridge University Library and, subject to the outcome of this consultation and the Secretary of State’s confirmation under §13 of the Legal Deposit Libraries Act 2003, the Library of Trinity College Dublin will all pay a contribution towards the costs of supporting legal deposit and will be able to access all legal deposit content in the shared technical infrastructure.

Non-print online material deposited under the regulations will be stored in this Shared Technical Infrastructure, which currently has four storage nodes located in St Pancras, Boston Spa, Aberystwyth and Edinburgh.

Each node stores a full copy of all the materials held within the system. The nodes are in constant communication with each other across a secure network, with automated routines for self-checking, replication and repair; if a digital file stored in one of the nodes becomes corrupted or lost, it is automatically restored from one of the other nodes. Furthermore, each node also uses a RAID 5 array in which files are copied and stored on two or more physical disks, with self-checking and replication between the disks. Finally, at present a separate and additional copy of the material is also stored offline, in a ‘dark’ archive, as a last resort backup in case of catastrophic failure affecting the entire system.

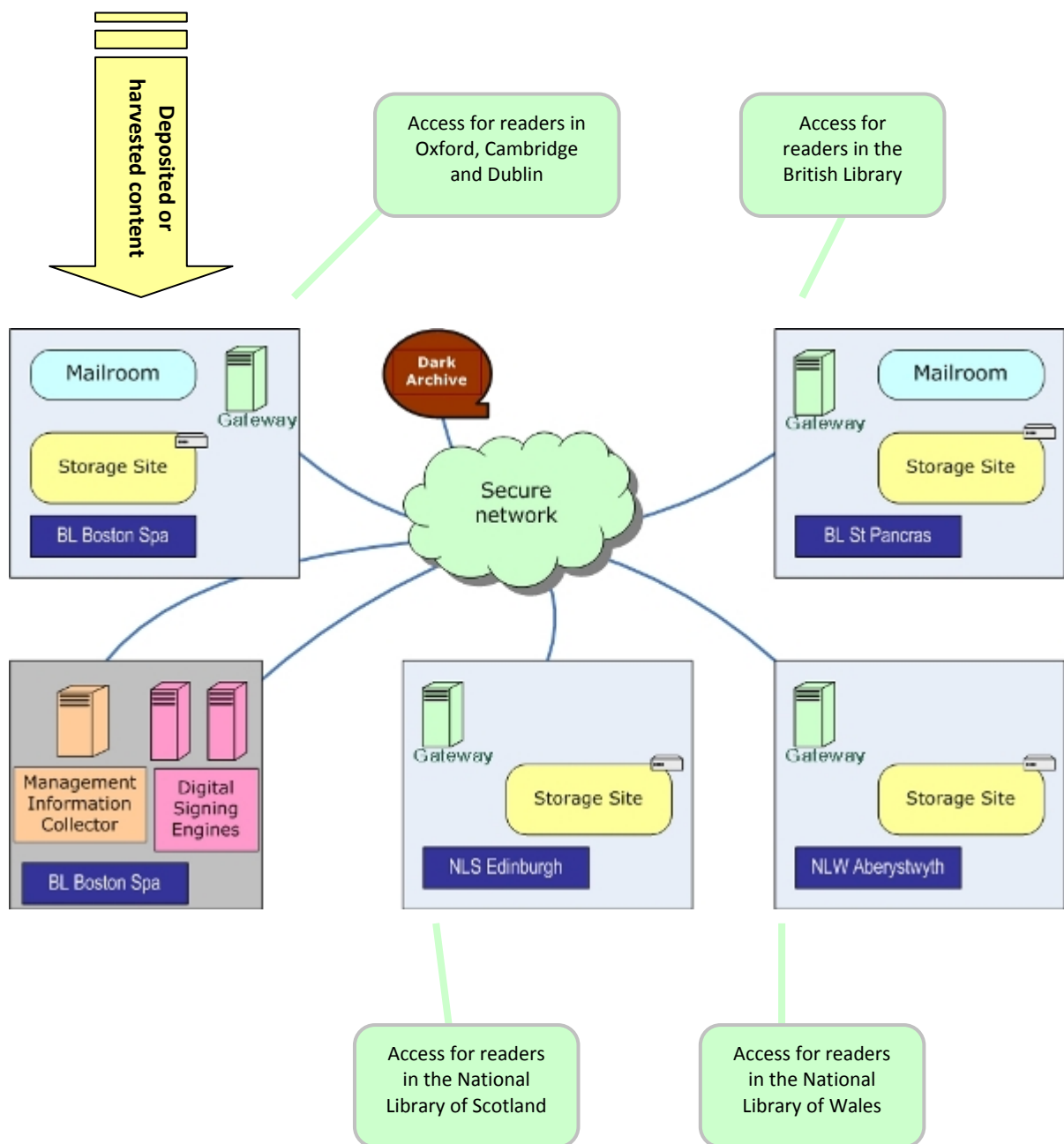
These measures are designed to ensure that the system is extremely resilient and capable of preserving content for many years. The system is also designed to be secure; a digital rights policy module in the system will allow just one reader at a time in each library (i.e. up to six in total, but only one in each library) to use a deposited item, in compliance with the Regulations.

Descriptive bibliographic information (metadata) will also be stored in the Shared Technical Infrastructure. Copies of the metadata will be exported, with periodic updates as required, to each library’s resource discovery system or catalogue. Readers in each deposit library will be able to identify legal deposit material within the library’s collections and submit a request to read it. For readers in the British Library, the National Library of Scotland and the National Library of Wales, access to the material will be via their local node; readers in Oxford, Cambridge or Dublin will

connect to one of the nodes via a secure network in order to view the material. In all circumstances, access will be managed by an overall Digital Policy module which controls all use of material in the system.

The deposit requirements and processes for non-print offline publications (see Regulations 6 to 13) have the same characteristics as those for print publications. Offline non-print publications will be stored locally by the British Library and each other deposit library that requests and receives a deposited copy. Under the terms of the Regulations, the digital files on the offline carrier may be copied and stored in the Shared Technical Infrastructure.

Fig 1. High-level design diagram



### **Processing new online legal deposit material – “freely available” websites**

The deposit libraries envisage that freely available websites will usually be deposited via a harvesting process, carried out by any of the deposit libraries. In practice it is likely that the majority of harvesting activity will be undertaken by the British Library, but other deposit libraries, particularly the National Libraries of Scotland and Wales, may also harvest publications of special importance and may supplement the work carried out by the British Library. Material harvested by each deposit library under legal deposit will normally be stored in the shared technical infrastructure and made available to readers in all of the deposit libraries, subject to access restrictions specified by the Regulations.

Deposit libraries do not intend for the harvesting process to affect the normal operation of a website, nor will it necessarily require any active intervention by the publisher or website owner; the harvesting tool will identify itself to the website’s hosting system when requesting content, so activity can be logged and so that the publisher’s server can ‘permit’ the harvesting, but the requesting, permitting and depositing process can be (and is being) constructed so as to be entirely automated.

### **Processing new online legal deposit material – other publications**

For publications that are subject to access restrictions, such as a password or a requirement to pay before viewing, and which cannot therefore be harvested directly, one of the deposit libraries will engage directly with the publisher to discuss and identify a mutually acceptable deposit mechanism. A range of options will be available, including various methods for a publisher to deliver the material or for the publisher to provide access for the relevant library to download material directly from the website. The requesting deposit library will also agree with the publisher which publications are in scope, if necessary implementing deposit in stages by type of publication. In most cases, it is likely that the requesting library will be the British Library, but for some it may be the National Library of Scotland, the National Library of Wales or another deposit library. The requesting library will store deposited content in the Shared Technical Infrastructure so that it can be made available to the other deposit libraries, thus avoiding the need for a publisher to set up separate depositing arrangements with each library.

### **Controlling access and copying**

The legal deposit libraries are firmly committed to protecting the rights of copyright-holders whose works are held in the legal deposit digital archive and are investing in a stringent and secure system. The Shared Technical Infrastructure has a sophisticated layered security model that uses proprietary dedicated gateways to limit the footprint, transaction scope and accessibility of the underlying content in the store. This will ensure that legal restrictions are enforced and that no illegal copying takes place.

## Technical Questions and Answers

1. ***Will the deposit libraries require specific content formats? Will requirements extend to a specific numbered version of the format, or to specific tag sets or data schemes such as DTDs? What will publishers be expected to do if they don't publish in a required format? What will happen as formats evolve and new ones are created?***

The deposit libraries cannot impose any requirement to deposit in specific formats; our expectation is that content will normally be deposited in the format used for publication and we aim to be capable of processing all commonly used ones.

However, in cases where the libraries are not yet capable of processing content in the format used for publication, or where content is published in more than one non-print format thus allowing a choice, we envisage that the requesting library will contact the publisher to discuss and choose a format that is acceptable to both parties. We envisage that we will generally be able to reach agreement, however the Regulations set out a policy around format choices in the unlikely event that agreement cannot be reached.

Naturally the deposit libraries will have certain preferences where a choice is available and, wherever possible, we intend to publish information about these preferences to assist publishers. For example, a joint group of representatives from the deposit libraries and certain publishing organisations recently prepared guidance notes for publishers who wish to deposit their scholarly electronic journals under an existing voluntary deposit scheme, and these notes are publicly available on the BL's website. See:

<http://www.bl.uk/aboutus/stratpolprog/legaldep/depositingelectronicjournals/depositing.html>

2. ***What if publishers have the required or preferred format as part of a pre-publication process – would that be sufficient?***

The deposit libraries did recommend to the Legal Deposit Advisory Panel that this should be permitted in appropriate circumstances, as it was recognised that an "as published" format may not always be ideal for preservation. However the Regulations as currently drafted, with their important legal protections (against infringing copyright or database rights, and in the case of a defamation claim), only cover 'copies of published material'. Therefore the deposit obligation and the legal protections offered by both the 2003 Act and the Regulations do not cover material deposited in a format that is only used as part of a pre-publication process. Publishers, in their responses to the consultation, may wish to suggest a reconsideration of this point if they feel that it would be helpful, but should note that the draft Regulations necessarily reflect the position of the primary legislation in respect of this point.

3. ***What will be the metadata requirements? What if publishers don't routinely produce metadata in the required format or level of granularity?***

The Government's consultation document indicates that publishers will be expected to deposit metadata, and describes the principles on which the Regulations are based (see Section 12 of the consultation document):

- "Publishers are not expected to generate metadata solely for the purpose of Non-print Legal Deposit;
- "Where metadata forms part of the works it should be deposited at the same time as the remainder of the non-print works;

- “The creation, adaptation, enhancement and use of metadata do not form part of the Legal Deposit draft Regulations;
- “Metadata that is collected by the Deposit Libraries through the legal deposit of works cannot be sold to any third party.”

The deposit libraries will comply with the Regulations once they become law and will only request descriptive metadata that a publisher routinely produces in order to access the work. A publisher’s metadata will be used by the deposit libraries for catalogue entries, resource discovery tools, national bibliographies and other such purposes. However the requesting library will not require a publisher to reformat metadata, nor to produce new or additional metadata, specifically for legal deposit. The deposit libraries expect to carry the burden of reformatting or adapting metadata where necessary and, where a publisher’s metadata is not of sufficient granularity or where the publisher does not routinely produce any metadata at all, the deposit libraries will arrange to extract or generate metadata from the deposited publication separately and at their own cost.

**4. *What will be the transport mechanism, e.g. “push” or “pull”?***

Both options will be available.

For “paid for” and restricted access publications (see Regulations 17 to 19), the requesting library will contact the publisher to discuss and identify a mutually acceptable solution; options might include the publisher providing access so that the library may harvest directly from the publishers website, the publisher uploading content to a secure facility from where the library may download it, or the publisher delivering the content direct to the library.

**5. *Will there be a “digital manifest” (aka “electronic bill of lading”) requirement?***

Including a digital manifest in the deposit package may be encouraged as good practice, for content that is uploaded or delivered directly by a publisher. However the requesting library will not insist upon digital manifests as a specific requirement.

**6. *To what extent will the specifications differentiate between publication types, e.g. journals, magazines, newsletters, newspapers, books, etc.?***

The number and variety of formats, publishing models and digital specifications that publishers use is enormous and they continue to evolve. In the Government’s Impact Assessment<sup>1</sup>, seven generic high-level models were identified; but it is recognised that, at a more granular and practical level, the range of differences between publication types and between the practices of individual publishers is much wider.

In order that the deposit libraries’ operational and cost burdens remain manageable, the deposit libraries will regularly publish information for publishers about their preferred specifications and deposit practices, and will encourage publishers to use one of these preferred formats and methods wherever possible. However, no specification will be imposed as a requirement, nor will the deposit libraries expect a publisher to implement processes or specifications which create an unreasonable or unnecessary additional burden.

---

<sup>1</sup> See: <http://www.culture.gov.uk/consultations/7449.aspx> : Impact Assessments, pages 50-52

7. ***How will the BL deal with content that requires licensed software to manipulate it, and is there any ongoing obligation to publishers?***

Regulation 22 states that a copy of any computer program and any information (including any tools and data) necessary to access the work, including any information required to allow a reader to read the work, must be delivered together with the publication.

In practice this is only likely to be necessary in a small number of cases where the software required is bespoke or highly specialised as deposit libraries will normally have separate licences for most browsers, most office productivity packages, e-book reading software and suchlike. The requesting library would expect any potential requirements to be identified, discussed and mutually agreed as part of their initial dialogue with a publisher.

Deposit libraries will normally request that publishers remove, before depositing the material, any Technical Protection Measures that are designed to enforce digital rights management policies. This is because such measures and software typically obstruct the long-term preservation of a publication and may well conflict with the provision of access in the deposit libraries' premises on the terms required by the proposed regulations.

8. ***To what extent will publishers be expected to adapt what they deposit to meet the requirements of the BL and its need to 'future-proof' its archive, or will they only be required to deposit what they have and the BL will do the rest?***

The deposit libraries will publish, and update as necessary, information about the range of formats, specifications and deposit methodologies that they can support as a matter of routine, indicating any that are "preferred" and any that cannot be supported. An early example of this being put into practice, for publishers voluntarily depositing scholarly e-Journals, may be seen at:

<http://www.bl.uk/aboutus/stratpolprog/legaldep/depositingelectronicjournals/depositing.html>

Publishers would be encouraged to use one of these preferred formats and methods wherever possible, so that the burden for deposit libraries of implementing adaptations for each individual publisher is kept to a minimum. However no specification will be imposed as a requirement, nor will libraries expect a publisher to implement processes or specifications that represent an unreasonable or unnecessary additional burden.

9. ***What technical solution is being used for harvesting?***

The British Library currently uses Web Curator Tool (WCT) for selective web archiving. The development of WCT is a collaborative project between the British Library and the National Library of New Zealand. WCT manages selective archiving workflows and gathers copies of web resources using embedded crawler software called Heritrix. WCT is open-source, freely available under the terms of the Apache Public License. The British Library plans to use Heritrix for domain crawls in the future.

10. ***What standards are being used [for harvesting]?***

The British Library uses a set of software tools, designed or extended specifically for web archiving, serving different purposes. For example, Heritrix is used to collect copies of websites, the Open Source Wayback Machine (OSWM) is used to render archived websites and Nutchwax is used to create indexes for search. These tools are commonly used by national libraries and archives around the world. Archived websites are stored in the WARC

format, a file format specifically designed to support harvesting, access and preservation of archived web content. WARC is an ISO standard.

**11. *What happens to external links?***

External links from an in-scope website will be downloaded by the crawler if the linked content is also within scope of UK legal deposit as defined by the Regulations. Content outside the scope of UK legal deposit will not be harvested; deposit library readers, when using the archived website and following the link, will see a message informing them that the relevant page or resource has not been archived.

**12. *What happens to live feeds within my website?***

If live feeds referred to here are RSS or Atom feeds, then these are straight XML documents and will be downloaded like any other (in-scope) object.

**13. *Will I have to do anything to my website to allow the harvesting bot to access it?***

No.

For web pages and publications which are "paid for" and/or subject to public access restrictions (see Regulations 17 to 19), the requesting library will contact the publisher to discuss and identify a mutually acceptable method of delivery, options for which might include the publisher providing a login or other means of access for the deposit library to harvest the material. However, in no case will the requesting library ask the publisher to make changes to their website that could affect its normal operation or presentation for other users.

**14. *Do I have to disable my robots.txt files?***

No. In order to harvest complete websites the WCT harvester will archive all relevant resources within the scope permitted by Regulations, where necessary disregarding general robots.txt instructions that are intended to exclude indexing by search engines. This is in order to avoid imposing any administrative burden for website owners for legal deposit. However the deposit libraries' harvester will also use the appropriate protocols (a "user-agent string") to identify itself and provide a link for further information, and to inform the website owner about which resources have been harvested. For example the British Library's current user-agent string being used for the UK Web Archive, when harvesting websites for which the owners have volunteered permission, is "Mozilla/5.0 (compatible; heritrix/1.14.1 +http://www.webarchive.org.uk/)" with the IP address 194.66.232.85.

**15. *Will the harvesting tool affect the running of my website?***

No. The crawler can be configured to impose a delay between fetching URLs (Uniform Resource Locators) from the same host which is a multiple of the amount of time it took to fetch the last URL downloaded from that host. For example, if it took 800 milliseconds to fetch the last URL from a host, and the delay factor is set to 5, the crawler will then wait 4000 milliseconds (4 seconds) before processing another URL from the same host. We will not request multiple URLs from the same host concurrently.

In addition, a limit can be set on the number of URLs and bytes downloaded: exceeding the limit will automatically stop the crawler. This ensures that, if the crawler falls into some kind of "trap" or "infinite loop", it will stop automatically when either of these limits is reached.

**16. *How will my content be held by the deposit libraries?***

Offline non-print publications, such as a CD Rom or microfilm, will be stored by the British Library and each other deposit library that requests a copy. A copy of the digital files on the offline medium may also be made and stored in the Shared Technical Infrastructure for preservation.

Online non-print publications (whether freely available or behind authentication and subscription barriers) will be stored in the Shared Technical Infrastructure, on self-checking and self-repairing nodes currently in Boston Spa, St Pancras, Aberystwyth and Edinburgh.

**17. *Will my content be made available on the internet?***

No. For non-print regulations, the Act defines a reader as "...a person who, for the purposes of research or study and with the permission of a deposit library, is on library premises controlled by it". The deposit libraries will implement Digital Rights Management security to ensure that deposited content cannot be accessed by users outside the premises of the deposit libraries (unless of course a publisher gives separate permission for wider access to their content, whether on a voluntary basis or via an additional purchased licence).

**18. *What security measures are in place to stop people hacking in to the Deposit Libraries systems and taking my content?***

We have developed a sophisticated layered security model that uses proprietary dedicated gateways to limit the footprint, transaction scope and accessibility of the underlying content in the store. The deposit libraries supporting the shared technical infrastructure already subject themselves to independent, routine IT security audits and follow up on their recommendations. As this store is the deposit libraries' strategic long-term digital content repository, significant efforts have been made to ensure its security, integrity, resilience to hardware and software failures, scalability, and the detection of, and recovery from, content corruption.

**19. *My site uses video clips extensively - will this affect my need to deposit?***

If the site consists solely of video clips, or if any other content is merely incidental to the video clips, the Regulations<sup>2</sup> indicate that it would not be subject to legal deposit. However, the site or publication should be deposited if the video clips or sound are being used in conjunction with other (written or pictorial) content, for example, to illustrate ideas, as supplementary material or as part of a multimedia offering.

---

<sup>2</sup> See draft regulation 2(3)(c)

**20. *How will you control access to my content to ensure only one person at any one time is able to view it in each of the Deposit Libraries?***

The deposit libraries will comply fully with the Regulations in this regard. Digital material that is received under legal deposit will be stored in the Shared Technical Infrastructure. As described elsewhere in this document, the system has a layered security model which prevents access to items stored in it other than via an access gateway. When the access gateway receives a request for any item from a client device it first checks that the client device has the correct rights to access the content.

In our first implementation of access controls, the system will be configured to allow delivery of content only to specific, designated display devices in the deposit libraries; the first delivery will cause all subsequent requests for that content from the same library to be denied for the rest of the day. Subsequent future implementations of access controls will allow an additional view of the item once the previous view has ended and may potentially allow other display options for reading the item. However in every case controls will ensure that only one person in each deposit library is allowed access to an item at any one time.

The system records all requests for access to items that it holds. This access record will allow us to verify that no concurrent access has occurred.

**21. *How will you stop people copying my content?***

As stated above, the system has a layered security model which prevents access to items stored in it other than via an access gateway. Initially, in the first implementation of access controls, the system will be configured to allow delivery of content only to specific designated display devices in the deposit libraries which will be 'locked-down' and will not provide connectivity to any external device or network including USB drives, printers, internet etc. Subsequent future implementations of access controls will permit printing in compliance with the terms of the draft regulations.

**22. *What quality assurance measures will be in place?***

The architecture of the Shared Technical Infrastructure has been designed from the ground up to ensure the security and the authenticity of the content held in it. Any changes to the architecture must be verified by a Technical Advisory Board before being implemented. As it is the strategic long-term store, special care is taken over changes to the system, with a rigorous testing and quality assurance regime.

In addition, the British Library has an IT Security Team which sets policies and standards for IT security matters across all of the Library's IT systems. This team mandates the level of security audit for the system, including an annual IT security audit. The recommendations from this audit are captured and tracked through to completion.

**23. *How will the 'partnering' nature of "non open web" deposit with individual publishers work on implementing ingest?***

The deposit libraries routinely liaise with individual publishers on operational issues concerning legal deposit of print publications and expect to work with digital publishers in a similar way, to the benefit of all parties. The Regulations set out the basis on which deposit is expected to work and the deposit libraries will fully comply with them.

As described earlier, for “paid for” and restricted access publications (see Regulations 17 to 19), the requesting library will contact the publisher to discuss and identify a mutually acceptable deposit solution; options might include the publisher providing access so that the library may harvest directly from the publishers website, the publisher uploading content to a secure facility from where the library may download it, or the publisher delivering the content direct to the library.

The deposit libraries welcome the principles set out in the Government’s consultation document (see Section 14) relating to the management of legal deposit arrangements and the governance of relationships between the deposit libraries and publishers. The deposit libraries intend to work in close partnership and collaboration with each other and with publishers, in order to implement the Regulations and to establish an appropriate management structure and governance arrangements.

**24. *Will digital content be stored outside the UK?***

No. The nodes in which online content is stored are currently in Boston Spa, St Pancras, Edinburgh and Aberystwyth. Subject to the Republic of Ireland satisfying the Secretary of State over the equivalency of their laws pursuant to section 13 of the Act, readers in the Library of Trinity College will be able to access and read the material stored in these nodes, but no online content will be stored in the Republic of Ireland.

**25. *What guarantees will the Libraries give publishers that the information provided will not be misused?***

As stated in the introduction to this document, the deposit libraries are firmly committed to protecting the rights of copyright-holders whose works are held in the legal deposit digital archive and have invested in a stringent and secure system. As outlined above, this includes a sophisticated layered security model that uses proprietary dedicated gateways to limit the footprint, transaction scope and accessibility of the underlying content in the store. This will make sure that legal restrictions are enforced.

It should be noted that individuals also have their own responsibility not to infringe copyright under UK law, and the deposit libraries make users aware of this when they are issued with a reader’s pass.

**26. *My website changes regularly, do I have to update with you regularly?***

If a website can be harvested by the deposit libraries, there is no need for a publisher to contact the harvesting library about updates. To avoid placing a burden on website owners, the deposit libraries will harvest periodic “snapshots” of eligible websites via automated requests. Rather than archiving every iteration, the intention is to take a representative harvest of the UK domain on a timely basis, likely to be once or twice a year. However in some cases, where a website or content is of particular relevance to the priorities identified in the libraries’ collection development policies, more frequent “snapshots” may be appropriate.

**27. *How will I know my content in part or whole has been harvested?***

The deposit libraries’ harvester will use an automated process (a “user-agent string”) to identify itself and inform the website owner about which resources it is requesting; it will

also provide a link for further information. For example the British Library's current user-agent string being used for the UK Web Archive, when harvesting websites for which the owners have volunteered permission, is "Mozilla/5.0 (compatible; heritrix/1.14.1 +http://www.webarchive.org.uk/)" with the IP address 194.66.232.85.

**28. *How will differences on issues such as commercial impact and embargos be resolved?***

The deposit libraries accept and are ready to satisfy the expectations set out in Section 14 of the Guidance Document (see Section 14), for an appropriate management structure with an informal appeals process; the deposit libraries will shortly be publishing the principles of a management process they will put in place.

**The British Library**

**30 November 2010**