Non-Print Legal Deposit Digital Preservation Review

Final Report

Date of Completion: 15/12/2017

The Digital Preservation Coalition (DPC) exists to secure our digital legacy. We enable our members to deliver resilient long-term access to digital content and services, helping them to derive enduring value from digital assets and raising awareness of the strategic, cultural and technological challenges they face. We achieve our aims through advocacy, community engagement, workforce development, capacity building, good practice and good governance.

www.dpconline.org

1 Executive Summary

In November 2016 the UK and Ireland Legal Deposit Libraries (LDLs), acting jointly, invited the Digital Preservation Coalition (DPC) to offer external assessment of the digital preservation actions that they undertake to deliver their commitment to the UK Legal Deposit Libraries (Non-Print Works) Regulations 2013. These Regulations implement the UK Legal Deposit Libraries Act 2003 which, *inter alia*, makes provision for the 'use and preservation of material deposited' for 'Non-Print Legal Deposit' (NPLD) Collections.

Unlike other aspects of the legal deposit regulations, preservation actions for Non-Print Legal-Deposit are concentrated at the British Library (BL) so the assessment was to focus particularly on digital preservation workflows and capabilities of the BL. The DPC has therefore assessed digital preservation processes at the BL as they are applied to NPLD collections.

This report outlines processes and conclusions from that assessment. It describes the design of assessment metrics, the assessment methodology, and the outcomes of two phases of assessment and review. A number of recommendations arose from the Initial Assessment, which were reported in an interim report to the LDLs in May 2017, including 14 recommended actions in five areas: Completeness Checking Across the Lifecycle (see Glossary); Integrity Checking (see Glossary); Regulations Relating to Access and Rendering Software Deposit; Governance of Preservation Operations; Management of Ingest and Replication (see Glossary) Processes. The DPC returned to the LDLs in September and October 2017 to review the extent to which the LDLs (in particular the BL) had adapted or changed their processes in response to those recommendations. The evidence and findings of that second review are presented here.

Significant conclusions from this assessment are summarized as follows:

- Much of the BL's digital preservation practice has been found to be exemplary. The BL has been a world leader in many aspects of digital preservation. Its expertise has been wisely and consistently brought to bear on the preservation of NPLD collections in fulfilment of the 2013 Regulations.
- The scale and complexity of the challenge which the preservation of NPLD collections generates should not be underestimated. Moreover, it represents a continuously evolving challenge for the BL on behalf of the LDLs; resources, skills and technology need constant renewal.
- BL resources have been appropriately prioritized towards the larger content streams. These were addressed first. Focus is now moving to addressing the smaller, but more complex and interactive, streams which represent a greater technical challenge.
- A number of operational failings were observed in the preservation of NPLD collections at the BL. These were reported in the Initial Recommendations. Subsequent assessment verified that actions had been initiated to resolve them and in most cases were fully complete by the time of the second review.
- Attention should continue to be paid to those actions in progress that address identified weaknesses but which will take longer to mature. In particular, changes to the governance of the preservation functions should be closely monitored to ensure that the new structure remains effective.
- The LDLs have responded positively, effectively and in a timely manner to the review. The reviewers have come to an independent judgement with detailed knowledge of best practice and standards around the world, and full access to all relevant documents, processes and staff at the LDLs. In our view, the LDLs, and the BL in particular, demonstrated a firm commitment to continuous quality improvement, and their digital preservation practice is exemplary in many respects.

Table of Contents

2 Introduction

2.1 About the Non-Print Legal Deposit Digital Preservation Review

In November 2016 the UK and Ireland Legal Deposit Libraries (LDLs), acting jointly, invited the Digital Preservation Coalition (DPC) to offer external assessment of the digital preservation actions that they undertake to deliver their commitment to the UK Legal Deposit Libraries (Non-Print Works) Regulations 2013. These Regulations implement the UK Legal Deposit Libraries Act 2003 which, *inter alia*, makes provision about the 'use and preservation of material deposited'.

The purpose of the project was defined explicitly as being 'to provide independent assessment of the extent to which the Legal Deposit Libraries have provided a digital preservation service consistent with their requirements to deliver the Legal Deposit Libraries (Non-Print Works) Regulations 2013'.

The main outcome of the project was this report, an independently verified review of digital preservation practices as applied to Non-Print Legal Deposit (NPLD) Collections within the Legal Deposit Libraries. Submitted in December 2017, the report highlights current strengths and weaknesses and suggests areas of challenge likely to arise in subsequent years.

The focus of this project was digital preservation, not access. Access and security have already been subject to rigorous assessment by the LDLs. Hence, by commissioning this project the LDLs have turned their attention to the core activities, infrastructure, organization and staffing concerned with preserving NPLD. For the purposes of this assessment, access and security have been considered only in the context of their direct impact on digital preservation.

Three collections form the primary scope for this assessment: web archive (including the small voluntary deposit collection and the vast domain harvest (see Glossary)); non-print deposit of eJournals; non-print deposit of eBooks. These three constitute a majority of the non-print legal deposit collections. In addition, effort has been targeted on two smaller content types – cartographic data and digital music scores – which also fall into the scope of the Regulations, with work under way to build preservation workflows for them.

This report is the final outcome of a series of iterative assessments delivered over five phases throughout 2017:

- Phase One: design of the assessment and initiation of the project.
- Phase Two: an Initial Assessment of Preservation Capability, providing a 'snap-shot' review of current practice and making specific recommendations.
- Phase Three: an opportunity for the British Library (BL) in particular to respond to the recommendations
- Phase Four: a Review of the Response to Initial Recommendations providing further independent review to assess explicitly if and how weaknesses identified in the earlier phases had been addressed and resolved.
- Phase Five: presentation of the final and formal report on the LDL capability to preserve NPLD Collections.

The LDLs have further requested that the DPC be available to support the Department for Culture Media and Sport (DCMS) review for up to 12 months beyond the end of the project; this includes adding an appendix to the report that identifies any significant changes in capability and practice between the end of the project and the point of review. The whole process, including final report,

was completed by 15 December 2017, with the appendix review to be submitted on request thereafter.

The assessment has included not just processes and data but staffing, skills, planning and policy matters. The majority of digital preservation effort with respect to NPLD collections is delivered by the BL, though others (especially the National Library of Scotland and the National Library of Wales) have been involved in, and informed about, the project. In addition, the project runs in parallel with an independent assessment of digital preservation policy and capability being undertaken at Cambridge University Library and the Bodleian Libraries with funding from the Polonsky Foundation. Therefore, lessons learned from the project have been further shared with relevant staff in those institutions. Moreover, the assessment has been structured in such a way that it could be repeated, whether by the Legal Deposit Libraries at some later date, or by other institutions seeking validation of their digital preservation activities.

The DPC intends to use the experience gained from this assessment to benefit the rest of its membership, and will attempt to derive case studies, briefing days, *Technology Watch Reports* and other supporting materials from the process as appropriate, and subject to the agreement of the LDLs.

2.2 About this report

This document is the final report for this Review. It reports on both the Initial Assessment and the follow up Review of Response to Initial Recommendations that considered actions undertaken after matters were reported to the LDLs in May 2017. The metrics for the assessment are presented below, and findings are based on a series of interviews and desk-based assessment of written policy, procedures and structures within the British Library. This final report summarizes and synthesizes a significant amount of gathered evidence, capturing identified strengths and weaknesses, noting how these have been addressed and the implications for the ongoing capacity of the LDLs to discharge preservation requirements that arise from the Legal Deposit Regulations. Where appropriate, the report also makes specific recommendations on perceived weaknesses that persist and on how good practice could be replicated. It is important to emphasize that the function of these recommendations, the report and the assessment from which it derives, is to enable and to frame continuous quality improvement so that quality management within the Legal Deposit Libraries, and especially at the BL, is independently validated.

This document was submitted formally to the LDLs as evidence of competence and exemplary practice but has three other important audiences. Firstly, the evidence assembled and conclusions drawn here can be shared with those involved in reviewing the Legal Deposit Regulations, including DCMS and other interested parties. Additionally, by addressing and reviewing practice, especially at the British Library, it will interest the operational staff of the LDLs involved in the acquisition, preservation and management of non-print legal deposit collections, as well as their line managers. Lastly, the DPC, as an independent broker for the worldwide digital preservation community, is itself also an audience for this document, as it provides a case study in the independent assessment of digital preservation practice.

Appendix Three: Glossary provides definitions of technical and digital preservation-related terminology, as well as expansions of commonly used acronyms.

2.3 About the assessment: process and metrics

An array of standards and audit frameworks supports independent quality assurance and self-assessment for digital preservation;[1] indeed, the proliferation of standards has been cited as a barrier to participation. Consequently, the assessment process began with a review of the available frameworks to see which was most suited to the purpose.

The development of audit standards has given particular prominence to the concept of the 'Trusted Digital Repository', an idea initially codified within the 'Trusted Repository Audit Checklist' and expressed more formally within ISO16363:2012,[2] a standard which is in turn embedded within the European Framework for Audit and Certification of Digital Repositories (see Glossary for more about ISO16363). The European Framework offers a staged model for core trust certification, extended certification and formal certification.

Each of these frameworks assesses not only technical competencies, but also to some extent the appropriateness of the mission and the commercial resilience of the institution offering to undertake preservation. Such assessment is important in the context of digital preservation since the offer to preserve content over an extended period, especially on behalf of others, is an expensive but critical function upon which some sectors may in due course come to depend. Unless preservation is properly mandated within an organization, and unless the financial and reputational costs are fully understood by senior management, commitment can weaken, investment stop and, ultimately, preservation fails.

The function of the current review, however, is quite precise: to ensure that the requirements of the Legal Deposit Regulations are being discharged appropriately. In this context, questions of mandate do not arise, and many aspects of security, staffing and sustainability are out of scope. Consequently, the scope of the project implies that many of the organizational metrics of standard digital preservation audit tools are redundant, creating room for a deeper analysis of operational and governance matters as they relate to NPLD collections.

The project team therefore explicitly adopted existing audit tools for digital preservation, adapting them specifically to the current project. The Unified Requirements for Core Certification of Trustworthy Digital Repository,[3] formally known as the 'Data Seal of Approval' (DSA), provide the underlying toolkit. The internationally recognized 16 Core Trustworthy Data Repository Requirements[4] form the basis of this review, recognizing that the LDLs have a distinct mission and context. The scope of enquiry has been narrowed to focus almost exclusively on metrics directly relating to long-term preservation. The language and focus have also been adapted to ensure that the metrics are relevant and meaningful within a library, and specifically a Legal Deposit Library, context.

As well as adapting metrics from the Unified Requirements for Core Certification of Trustworthy Digital Repository to offer greater clarity, the approach taken here has been designed to ensure greater independence and transparency. Core Certification is normally achieved through a relatively lightweight form of peer review of documents supplied to an assessment board. In this project, however, the DPC acted as an independent reviewer. All of the LDLS' were consulted in the process and the project team drilled down to ensure consistency and to probe strengths by requesting

[1] For a comprehensive introduction see DPC (2016) 'Audit and Certification' in DPC *Digital Preservation Handbook*, online at: http://dpconline.org/handbook/institutional-strategies/audit-and-certification
[2] https://www.iso.org/standard/56510.html
[3] https://www.datasealofapproval.org/en/
[4] http://www.datasealofapproval.org/en/information/requirements/

information and interviewing staff where appropriate. This process is more exacting since it enables a more forensic and detailed analysis of strengths and weakness, drilling into areas of uncertainty and ensuring that there is no inadvertent misunderstanding. It is also more transparent than peer review, since there is no conflict of interests between assessor and candidate: the DPC will not in turn be subjected to the same analysis and therefore gains no benefit from eroding or amplifying the metrics in its favour.

Consideration has been given to each of the Legal Deposit Library (LDL) content streams where it was appropriate to do so. The review has focused on the current state of the preservation capability, but has also considered how this capability has changed since the introduction of the Legal Deposit Regulations (2013) and where current plans may take it by 2023. In this way the DPC has assessed the LDLs responsiveness to a 'moving target'.

2.4 Methodology for this Assessment

As stated in section 2.2, this assessment began with DPC reviewers collecting information by desk-based assessment of written policy and procedures, and through interviews with members of staff with follow-up questions for clarification. The preliminary list of relevant staff members was provided by British Library staff, who also co-ordinated and scheduled the interviews; the reviewers then identified additional staff to speak to after these initial interviews. In total, 18 people were interviewed on site at the British Library at Boston Spa or at St Pancras, or by phone. The first set of interviews were carried out over two months, in March and April of 2017.

At the start of each interview, the DPC reviewers provided a brief explanation of the review and answered any questions about the process or purpose of the interview. The interview questions had been developed to capture information about the defined metrics (see Appendix Two: Metrics). When relevant issues arose in the conversation, the reviewers encouraged staff to expound on these topics or their experiences. By exploring these organic conversational leads, the reviewers captured greater context and broader understanding of digital preservation (DP) across the British Library. In one case, the reviewers requested the interviewees show them processes or tasks referred to in the interview, and in many other cases requested access to documentation. Thorough notes were taken during each interview, with follow-up questions for clarification or for further information asked via email. The findings in this report are based on those interview notes, observations, and follow-up clarifications. All requested necessary documentation was readily supplied and no requests for interviews were denied, meaning that the reviewers were able to act freely, and this report is written on the basis of full disclosure.

The report and recommendations from this Initial Assessment of Preservation Capability were shared with the LDLs in May 2017. This includes 14 recommended actions in five areas (see Section Five).

In September and October 2017, the DPC returned to the BL with a series of follow-up interviews and desk-based assessments for the fourth phase of this review. These were designed to establish the extent to which the recommendations had been acted upon: whether recommended actions had been adopted; or new working methodologies established that exceeded the recommended action or in some other appropriate way made redundant the issue that had been identified. All follow-up documentation requested was submitted in a timely manner. The process followed was similar to that of the second phase, with the addition of requests for live demonstrations of key systems and workflows. The DPC also undertook interviews with staff at the National Library of Scotland to understand their role in the preservation and management of NPLD collections as well as the governance of the preservation functions. In total, 23 people were interviewed during the fourth phase.

3 State of Play and Future Challenge: Preserving NPLD

3.1 The NPLD Collections

Five collections currently form the main focus of NPLD digital preservation activities. Three of those, eJournals, eBooks and the web archive, are operating under business-as-usual ingest to the Digital Library System (DLS). The following table provides some basic statistics indicating the scale of the collections to date.

	Volume (Terabytes)	Number of objects (to nearest 1000)
eJournals	13	12,949,000
eBooks	2	533,000
web archive	123[5]	316,000[6]

The two remaining collections – cartographic and digital music scores – are yet to be ingested into the Digital Library System (DLS), but projects are in progress to develop workflows for them. These collections are significantly smaller than the first three described above, but come with their own challenges. In particular, the cartographic data consists of an array of data types that must be processed, combined and styled within appropriate geospatial software to enable it to be viewed.[7]

As noted above, all of these collections are within scope for this review. The following table indicates the total volume and total numbers of DLS objects, now and five years ago:

	Volume (Terabytes)	Number of objects (to nearest 1000)
Current (as of Sept 2017)	139	13,798,000
5 years ago (as of Dec 2012)	27	365,000

These statistics provide some indication of the maturation of the LDLs' digital preservation capability, as ingest rates have dramatically increased, throughput scaled up, and outstanding issues with data processing (typically relating to metadata quality) addressed.

3.2 Future Challenge: Emerging Formats

This section provides some context to the specific recommendations of this review by illustrating the considerable challenge faced by the LDLs in preserving digital content both now and in the future. Many factors can have a direct and indirect impact on the implementation of digital preservation actions, not least of which are the scale, complexity and rapid evolution of published digital content. This section describes how this context places the BL and other LDLs at the forefront of an evolving, and therefore challenging, digital preservation task.

3.2.1 Digital Preservation Sector Leader

The context for this report is the rapidly evolving but still relatively novel field of digital preservation. The BL has long been recognized as a sector leader in this field and in many aspects of this field it has led the world. It began addressing digital preservation challenges before many other organizations

[5] Note that the 2016 and 2017 domain crawls are yet to be ingested to the DLS. Including these progressively larger crawls brings the total of NPLD web archive content to 470TB.
[6] Note that these are WARC files (see Glossary) ingested into the DLS. Each ingested WARC file contains many http responses so typically represents large aggregations of web pages and websites.
[7] 'Map collecting in the digital age' by cartographic curators from the BL and NLS provides a detailed background to their work in this area http://journals.sagepub.com/eprint/7EWh3pC4ksGxmCKk6uJE/full

and, as an 'early adopter', it has had to resolve many of the problems from first principles. It has done so openly and collaboratively, meaning that later entrants have had a simpler journey towards implementation. For example, the BL developed its own digital repository software long before the emergence of the digital repository products that exist today. Whilst this review has identified some shortcomings in the technology developed by the BL and the processes it has implemented to manage and preserve digital content, the broader achievements of the BL in preserving NPLD content should not be underestimated. The BL has been, and remains, a sector-leading institution. Many aspects of the digital preservation expertise applied to the NPLD are genuinely world class.

3.2.2 The challenges of preserving 'simple' content

The evolving digital publishing medium presents a challenging moving target for preservation. On the surface, the eBook and eJournal content streams appear relatively straightforward digital entities to preserve. On closer inspection however, the challenges quickly become clear, and they are myriad. Although the scope of collecting policy has not changed, the sheer volumes of material and their complexity have, and this has a profound effect on digital preservation. Greater scale increases pressure on fixed resources; greater complexity means a continuous challenge to the effectiveness of previously established solutions. So, as currently constituted and for the foreseeable future, digital preservation for NPLD requires constant investment in capacity and sophistication, even as the scope of the collections remain well defined and stable. Significant challenges revealed during the review included:

- Expanding collection: as publishers move their already-published holdings from print to digital, new submissions to the BL not only include current publications, but also back copies originally published in print, possibly several years ago. This significant increase in volume makes predicting and planning acquisitions very challenging. Analysis of the material deposited under NPLD shows that 42% more content is received from the publishers who have transitioned to electronic deposit compared to when they deposited in print.
- Revisions: changes to content can lead to revisions, from minor correction of punctuation through to significant edits. Some of these changes will be of historical/curatorial interest; in other cases, only the most recent version will be key.
- Changes over time: publishing organizations and their outputs change and evolve over time. This might include closing down their operations, merging with other publishers or corporate entities, or closing or changing the name of their journal titles. These unpredictable changes result in further complication to established content workflows.
- Repackaging: material previously published in one form can be repackaged and published (otherwise identically) in different granular form; for example, journal articles re-published as an eBook.
- New media richness and interactivity: digital provides many new forms of informational communication. Sometimes these forms match well with emerging standard data formats, in other cases content becomes tied to proprietary software technology, presenting obstacles to preservation, management and access.
- Unpredictable deposits: the simple act of depositing the data can result in challenges itself. A publisher established as a submitter of content to the BL for preservation may supply its data in a typically gradual fashion (often referred to as a drip feed). However, on successful completion of lengthy negotiations with a new publisher to deposit its published material

digitally, an entire back catalogue may suddenly appear on the BL's FTP site, the mechanism used for submission. This unpredictability of the frequency and volume of deposits makes forward planning and the day-to-day management of ingest workflows a continual challenge.

Increasing volume, complexity and unpredictability of content place considerable strain on digital preservation workflows in a variety of ways. Greater volumes (both in numbers of items and sizes of component files) place strains on the workflows that must process them, requiring more resilient software processes and greater workflow automation to enable issues to be resolved without backlogs arising. Evolving complexity requires evaluation and research into new file formats, and new types of digital content. In some cases this may require new preservation techniques and, most likely, evaluation and implementation of new software applications to handle them. Unpredictability requires greater flexibility to react to changes in content and its supply. Deposited data that doesn't conform to previously encountered norms must be detected and workflows adapted to process it. The accuracy and completeness of digital preservation activities will be impacted with adaptation to meet these challenges.

3.2.3 Future Challenges

The LDLs have been facing many of the challenges described above for a number of years. Discussions held throughout this review have revealed a maturing preservation and data processing capability that has developed over time to meet these challenges. The BL's digital preservation capability is no longer the first response to a new problem, but a more nuanced solution that has benefited from real experience. This is illustrated particularly well by the latest developments in the eJournal processing workflows (discussed in section 6.1). However, further complexities are rapidly approaching, and the LDLs are working to understand them and prepare for their impact. Chris Fleet, Map Curator at the National Library of Scotland (NLS), notes a number of issues being faced as the LDLs work in partnership to develop a workflow to ingest and preserve NPLD cartographic data. These include:

- the limited suitable metadata published with geospatial datasets, so metadata often needs to be created by the LDLs;
- the wide variety and complexity of geospatial data formats and structures (vector, raster and spatial databases (see Glossary) with the usual need to style and process the data before it can be displayed;
- the widespread use of proprietary formats, and the lack of a widely supported non-proprietary archival format for vector map data, which creates more work in reviewing and loading datasets, and the potential need for software to be deposited along with the data;
- relatively rapid change in geospatial formats over time (such as the Ordnance Survey's transition from NTF to GML (see Glossary));
- the need to display and manage the datasets within appropriate geospatial software;
- limited staffing resources, as there are relatively few map curators within the LDLs, and fewer still with digital skills.

Taking on the preservation of cartographic data marks the progression for the LDLs from addressing content that is relatively easy to use and manipulate to content that presents a real obstacle to user access. By carefully processing, normalizing (see Glossary) and ingesting the metadata that

accompanies an eJournal article, it is possible to preserve the article over time. By preserving metadata this way, the eJournal can also be made discoverable to users, and enable them to access it and open the respective eJournal PDF itself – PDF being a format that presents few current challenges to user access. However, with cartographic data, the act of understanding the content and rendering it or presenting it to the user in a useful manner is, relatively speaking, far more difficult. The LDLs are actively working to prepare for these challenges.[8]

As well as the development of ingest processes for cartographic data and digital sheet music, a project is underway to assess where the preservation of NPLD will go next, and how the challenges of the inevitably more complex data will be addressed. Current content types under examination include:

- books as apps
- interactive narratives
- structured data (databases) (see Glossary)

This content further combines data with interactivity and software, blurring the conventional line between the object of preservation and the mechanism of access. Inevitably this raises the preservation challenge by an order of magnitude, as the complexity of the data goes way beyond that which has been previously deposited for preservation.

3.3 Illustrations of Practice

This section provides examples of the activities that the BL has undertaken to address the evolving digital preservation challenges described above. It does not seek to be exhaustive, but to indicate the range and depth of activity required to address new challenges faced, as well as to demonstrate how existing preservation capabilities can be evaluated and refined.

3.3.1 Web Archive Quality Assurance

The Web Archives provide a particularly difficult challenge to quality assurance due to the scale of content. The BL first developed web archiving on a permissions basis – crawling (see Glossary) only the content for which curators had received permission from the website owner. These permissions-based crawls took several days and could be manually checked for quality and completeness. Permissions-based quality checking uses automated methods, such as the checks carried out by the Web Curator Tool (see Glossary). Quality checking for permissions-based content, however, still requires a great deal of manual work by the curators, who perform visual checks. By comparison, the domain-level Legal Deposit harvests for which the BL are now responsible take a couple of months. This leap in duration of crawl represents a massive increase in the size of web archives. While automated completeness checks are performed at the point of capture, unfortunately, due to the complexities of web content and enormous scale of the domain crawls, errors and problems do occur. It is impractical for curators to manually check the domain-level content. As a solution, the web archive team have begun capturing screenshots for every web page harvested. As part of the harvesting process, the team opens a seed URL (see Glossary) in a browser and captures a static image. They use this to determine where hyperlinks are located and to generate an image map of the homepage. Even if at some point the html will no longer render, the static screen shot will be

[8] A presentation from Caylin Smith and Ian Cooke of the BL at the PASIG conference, provides further detail on this ongoing work https://figshare.com/articles/Emerging_Formats/5414983

available, albeit as a lower fidelity representation of an interactive web page. It will also provide a reference point revealing how a preserved web page originally appeared in a web browser of the time. These screenshots will provide a reference for curators to check the quality of the domain-level crawled content when necessary. Some content can be prioritized for visual quality check, such as news sites that cover national events. However, for the vast majority of the other content, quality assurance is about ensuring that the parameters set for the crawlers capture what they are meant to capture. In the meantime, the web archive team continues to try and improve automated reporting so that curators receive the information they need as soon as possible by developing quality metrics.

3.3.2 Preservation of Ingested Collections: Assessments, Sampling, and Action plans (PICASA) project

The PICASA project developed a process for assessing the condition of preserved digital content by manually sampling digital files from across the collections stored in the DLS. This provided some validation of the effectiveness of the automated checking applied to content at the point of ingest. Issues identified by this evaluation could then be examined and the respective software tools that apply the checks re-configured, replaced or enhanced.

PICASA identified a preservation issue with some eJournal and eBook content that required further attention. Incoming PDFs were validated (see Glossary) with JHOVE[9] to assess whether they were constructed in accordance with the specification for the PDF file format. If validation errors were reported, modified versions of the PDFs were created. Subsequent user access was provided via these modified files. Investigation by the Digital Preservation Team (DPT), however, revealed that this process was unnecessary, and furthermore sometimes introduced errors to the PDFs. The modified PDF files were therefore carefully disposed of. Access and preservation reverted to the original (unmodified) files which had sensibly been retained. New processes and policies were then introduced to reflect this change and ensure newly ingested content was no longer unnecessarily modified.

Rigorous manual checking of files is too labour-intensive to perform at scale. Sampling projects such as PICASA, therefore, provide a valuable approach for identifying problematic content and the respective deficiency or gap in automated checking through which said content passed at ingest into the DLS. Such issues can then be rectified by fine tuning existing automated checks that are applied to all content on ingest.

3.3.3 Improving Digital Preservation Performance: the ISO16363 Project

In 2015 the DPT began the ISO16363 project to review and refine repository management and associated digital preservation processes. The project performed a Library-wide ISO 16363 based self-assessment, and then sought to address a number of findings. To address issues raised by this self-assessment, the project team examined current processes for checking content stored in the DLS and the performance of the Object Authenticity Checker (OAC), a tool developed by the Library to check that digital objects have not been damaged or altered. As a result, the team have worked to build reliable integrity checking and to expand oversight of relevant processes and operations. The project has also developed new resources to support digital preservation (DP) across the Library, such as new documentation. Subsequent improvements made to the OAC by Application Support were informed by the ISO16363 project, most notably the creation of automated reports in spreadsheet format (or other human readable formats) to facilitate analysis by members of the DPT

[9] http://jhove.openpreservation.org/ see “Validation”, Appendix Three: Glossary

and other staff. The project has developed a Policy Documentation Map to make it easier for staff to identify key documentation and to understand the relationship between different policies at a glance. A Designated Community (see Glossary) Document was produced, which better articulates and describes the consumers of digital collections (including NPLD content). It provides guidance for setting up projects or for when the DPT provides quality assurance on a project elsewhere in the BL

4 Initial Assessment of Preservation Capability

This section of the report describes the detailed assessment of preservation capability supplied to the LDLs in May 2017.

Each lettered sub-part of the numbered metrics has a list of points of evidence gathered by the review along with a finding that summarizes the assessment of the preservation capability against an explanatory statement and a simple rating as follows:

- Satisfactory [S]: The preservation capability is of a good quality and meets requirements
- Action Required [A]: The preservation capability does not meet requirements with one or more issues that should be addressed

4.1 Content Preservation

Requirement: Preservation risks are mitigated by identifying, assessing and taking action to ensure content is understandable, sustainable and accessible to users *(adapted from DSA R3, DSA R10 and DSA R11)*.
[1a] Appropriate checks are applied to ensure quality and completeness of content on deposit/acquisition and subsequently ingest, where possible.

Key Evidence	Finding
• Thorough quality checking (e.g. web archives stream), cross-checking and item completeness checking (e.g. eJournals stream) is employed across the LDL streams, despite challenges. • Concern raised over missing eBook content estimated at 25000 titles. • Concern raised by completeness check of web archive content (2015 domain crawl only) against contents of the Digital Library System (DLS) showed 44000 missing WARC files (see Glossary). • Beyond small-scale manual sampling and the example noted above, there is an absence of automated completeness checking to tally submitted/acquired items with the contents of the DLS and report the results.	The checking of content received on deposit or acquisition represents a considerable challenge. Missing or partially missing content is not always evident, metadata schemes can change without warning, publishers merge, titles change and a number of other issues all conspire to create a moving target for preservation (see section 3.3). Commendable effort is employed by the BL in monitoring, analysis, quality assurance and cross checking to ensure content is complete at the item/title level where possible. The absence of any rigorous completeness checking leaves open the possibility that acquired content goes missing, is held up, or otherwise deleted at some point in the preservation lifecycle. Anecdotal evidence suggests this may have happened. A lack of management information reporting significantly hampers investigation of any processing issues [A].

[1b] Appropriate format identification, validation and other content characterization processes are applied as appropriate.

• Files are characterized on ingest, including file format identification and, where appropriate, file format validation. • Engagement with JHOVE Steering Group via the Open Preservation Foundation[10] to enhance validation performance. • Investigation of appropriate validation tools and approaches for emerging content (e.g. eBooks).	Applied format identification and validation meets with current expectations of best practice. Furthermore, the BL has been engaged in pushing forward understanding and best practice in this area and sharing it with the wider community [S].

[1c] Technology watch and other appropriate monitoring activities are implemented.

• Clear understanding, scoping and description of content streams (Content Profiles, key staff engaged with collections – Digital Preservation Team (DPT), Curators, IT). • Awareness and understanding of current and future content issues explored, researched and documented in File Format Assessments and tool evaluation work. • Collaboration and knowledge sharing with DPC and other international organizations.	The BL's work here has provided a good foundation of knowledge on which to base its preservation approach, and to enable preparation for ingesting new content streams. Attention has been given to upcoming issues and emerging formats [S].

[10] http://openpreservation.org/

[1d] Content preservation risks are assessed and identified.

Key Evidence	Finding
• Collection Profiles (documents that contain descriptive and other relevant information about the different collections preserved by the Library) and File Format Assessments (analyses of particular file formats to identify preservation risks, and how the respective file format should be managed and preserved effectively) demonstrate thorough risk assessment of content streams and detail actions to mitigate these risks. • The internal Digital Preservation (DP) training programme ensures content specialists across the BL can identify preservation risks as they arise. • The DP Helpdesk, which allows staff from across the BL to report issues and ask questions related to DP captures unanticipated risks identified by content specialists and enables DPT staff to engage with colleagues across the BL. • The PICASA project (see section 3.3), applied manual sampling and assessment of NPLD content in order to identify preservation risks. • The DPT carried out an adapted ISO16363 internal review to measure issues related to long-term preservation, an activity that has also provided oversight for potential risks across the BL and between workflows.	Assessment and identification of content preservation risks meets with current expectations of best practice. Ongoing programmes and policies have been successfully launched to ensure risk assessment and identification activities are continuous and embedded across the BL. The DP training programme and DP Helpdesk in particular support institution-wide awareness of the preservation risks associated with particular types of content. Activities such as the PICASA project allow for the early identification of unanticipated risks and effectively prevent the introduction of errors and problems into BL workflows [S].

• Plans are in place for a new project to develop an Integrated Preservation Suite. It aims to provide a workbench capability to combine a number of advanced preservation applications, including addressing the management and preservation of software critical for access to deposited data, the development of preservation planning at scale, and a registry of Representation Information (see Glossary).	

[1e] Action is taken to mitigate preservation risks and ensure users can access and exploit content.

Key Evidence	Finding
• Collection profiles and format assessments capture context and approaches to mitigate risks associated with specific content streams. The 'Preservation Intent' of web archive material, for example, focuses on preserving WARC files, associated metadata, and links between files. • The Designated Community document which derives from the ISO16363 project (see section 3.3) ensures user needs are understood and addressed. Because this document will provide guidance when setting up projects and for the basis of DP quality control, designated community requirements will be embedded in day-to-day practice. • Curators are responsible for representing the needs of the designated community identified for their content stream. • Capturing representation information and ensuring files are meaningful to a designated community are both requirements for the procurement and development of the new repository system (DAMPS Programme).	Action to mitigate preservation risks and ensure users can access and exploit content fulfils best practice and demonstrates sector-leading initiative. Collection profiles and format assessments address key issues for the LDL content streams with strategies based on cutting-edge sector knowledge. Detailed analysis of designated communities and their requirements underpins actions and policies that drive preservation across the BL, particularly through the ISO16363 project's Designated Community document and the role of the curators. The BL continues to innovate in this area [S].

[1f] Content preservation processes and risk mitigation actions are carefully managed and documented. Issues are documented, reported and escalated as appropriate.

Key Evidence	Finding
• Digital Preservation Policy and Digital Preservation Strategy. • Key governing boards – the Collections Policy Management Board and the Repository Management Service Group – provide oversight for the management of digital content and issues, including DP. • DP Helpdesk allows staff from across the BL to report issues and ask questions related to DP within their own remit.	The BL's Digital Preservation Policy clearly outlines the foundations and strategies for DP across the BL. Two different governing groups address DP within their remit. The Collections Policy Management Board, as a high-level steering committee, is positioned to identify gaps in policy. Furthermore, the DPT and other practitioners are able to raise issues or problems to this group when required. The Repository Management Service Group, including staff from IT, DP, and Licensing and Continuity, exists to prioritize requests for application support. This group also oversees capacity monitoring and accepts new functionality or documentation developed in project work. Low-level issues are collated by the DP Helpdesk and DPT staff [S].

[1g] Have formal documentation of policies and procedures for implementing digital preservation across the organization.

Key Evidence	Finding
• Digital Preservation Policy covers the actions for DP across the BL. • Collection Profiles provide guidance to the particular preservation issues associated with individual content streams. • A Policy Documentation Map was developed as part of the ISO16363 project to track documentation relevant to DP across the BL.	The BL has strategically implemented a range of policies to address actions and decision-making that affect DP. These policies are highly visible and maintained and updated on an ongoing basis. In an effort to co-ordinate policies related to DP, the ISO16363 project developed a Policy Documentation Map, which makes policies easier to find. This Map also allows staff to understand how different policies work at a glance [S].

4.2 Integrity and authenticity

Requirement: Bit-level integrity and authenticity (see Glossary) is ensured via replication and integrity checking processes *(adapted from DSA R7 and DSA R9)*.

[2a] Content is replicated to minimize impact of any bit loss.

Key Evidence	Finding
• DLS design is based on content replication across four geographically separated nodes. • eBooks, eJournals and web archive content is held in DLS. • Replication backlogs (particularly to Wales and Scotland nodes) have been experienced previously. Action has been taken to reduce this and monitor backlogs (although see 2b, below). • Digital sheet music and cartographic content streams are backed up. Projects are underway to ingest this content to DLS and deliver access solutions.	The DLS provides strong replication capability for eBooks, eJournals and web archive content. Projects are underway to ingest the LDL streams to the DLS. These projects appear well scoped and specified. Backlogs in the replication of data to other DLS nodes is a cause for concern, particularly in the context of some of the other findings listed in this report relating to bit integrity and completeness checking. Action has been taken to address the replication backlog over the last year [A].

[2b] Integrity checking ensures any bit loss can be identified.

Key Evidence	Finding
• The DLS integrity checking process (the Operational Authenticity Checker – OAC) is not fit for purpose. Various bugs have been identified that prevent it running continuously, and prevent recovery after a failure. • Evidence only of partial authenticity check of DLS nodes over the past two years. Previous logs not available (apparently deleted). There is no evidence that any DLS node has ever completed an integrity check. • OAC (and related automated repair function) has been re-written from scratch, is being deployed to the live system, and monitored to evaluate progress and roll out across the four nodes. • The cartographic content stream is not integrity checked.	A systematic failure to verify the bit integrity of the preserved data represents a significant preservation failing. This has placed the LDL collections at risk of loss. Until a complete integrity check has been performed the scale of any loss will remain unknown. The problem was identified following the ISO16363-based self-assessment (see section 3.3). Technical work to address this failing has to date (including re-writing and testing the code) required around two person months. This relatively rapid mitigation in the context of an issue apparently present for approximately a decade suggests this is primarily a failing of reporting and governance. Why dysfunctional code was not tested and fixed when first developed remains unclear. The lack of integrity checking of the LDL streams not held in DLS is a cause for concern but will be addressed when planned ingest projects are implemented [A].

[2c] Appropriate mechanisms are in place to minimize risk of accidental deletion or malicious damage.

Key Evidence	Finding
• Separate log-ins for each node minimizes risk of mass accidental deletion. • Layered security model restricts unauthorized access. • Access to all nodes restricted to a minimum of key staff.	Combination of repository design and procedure minimizes risk of accidental or malicious damage. Stronger role-based features likely to be provided by DAMPS should enhance this risk mitigation [S].

[2d] Technology choices and storage architecture minimize the risk of loss due to (common) hardware failure. Storage refreshment is managed conservatively.

Key Evidence	Finding
• Mix of technology across the four storage nodes (hardware and software). • Storage is managed and refreshed appropriately.	Satisfactory mitigation of storage risks [S].

[2e] Appropriate encryption, signing and related processes ensure authenticity of content and metadata

Key Evidence	Finding
• DLS design incorporates appropriate signing and identification of repository objects and metadata. • PREMIS-compliant event metadata is recorded (see PREMIS in Glossary). • Design of DLS requires metadata is committed to the store in addition to the repository database.	DLS design, operation and related processes ensure requirements are met [S].

[2f] Bit-level preservation processes and risk mitigation actions are carefully managed. Issues are documented, reported and escalated as appropriate.

Key Evidence	Finding
• Failure of integrity checking (see metric 2b). • Integrity checks not included in DLS Management Information reporting mechanism. OAC logs difficult to manually analyse and interpret. • Recent remedial work to OAC includes management reporting information of OAC activity. • OAC operational failure failed to trigger remedial actions,	Failure to retain OAC logs is a cause for concern given the importance of this evidence in terms of the record of preservation activity and the role of integrity checking in demonstrating the continued authenticity of the LDL data [A]. The absence of automated reporting of OAC operation appears to have contributed to governance issues relating to bit preservation within DLS. The addition of automated reporting (accessible by the Repository Manager role) is expected to remedy current issues relating to this metric (see also metric 4b) [A].

4.3 Intellectual Property Rights (IPR) and Regulatory Constraints

Requirement: IPR and LDL Regulatory constraints are monitored and managed to minimize impact on preservation of the collections *(adapted from DSA R2)*.

[3a] LDL regulatory constraints do not place undue restriction on effectiveness of preservation or a burden on resourcing of preservation.

Key Evidence	Finding
• The regulations have enabled the BL and the other LDLs to build a significant and valuable collection of digital data. • The onus is on the BL to adapt to changing formats, metadata and content which causes considerable challenges (see section 3.2). • Software necessary for rendering data deposited by publishers is very rarely provided by those publishers. • Legal Deposit Regulations restrict usage to those users that are onsite at an LDL.	The failure of publishers to deposit necessary rendering software (accompanying data) is a potentially significant obstacle to preservation. User feedback from the experience of accessing preserved content typically provides a valuable source of quality assurance around the ultimate success or otherwise of digital preservation. By restricting user access, this important feedback loop is severely impacted [A].

[3b] IPR constraints are well understood and managed.

Key Evidence	Finding
• Clear approach to recording IPR constraints in an easily interpretable form for subsequent access provision. • Challenges well understood by staff. • No significant IPR issues were identified.	Knowledge, understanding and implementation around IPR-related issues are strong. Staff are aware of upcoming challenges [S].

4.4 Organizational Infrastructure

Requirement: The repository has sufficient resources managed through a clear system of governance to carry out effective digital preservation *(adapted from DSA R5)*.

[4a] Staffing and funding are sufficient to sustain the repository, enable effective preservation and ensure permanent access to the collections.

Key Evidence	Finding

• Identifying total staff numbers working on digital preservation across the BL over time is challenging (most staff are not dedicated solely to digital preservation activities, and remits change over time). However, these numbers of staff based in the DPT give a useful impression:
  ○ DPT in 2013: 2.8FTE, DPT in 2017: 8FTE. DPT is expected to grow in the future given expanding remit of the team to more operational focus.
  ○ Database management team in IT is smaller in size now than in 2013, but is working more efficiently across its remit due to restructuring.
  ○ Digital processing staffing has remained roughly stable.
	Despite resourcing pressures and reductions in staff numbers at the BL, digital preservation appears to have been prioritized and protected. A growing staff quotient focused on operational preservation activities is reassuring, given both the scale of archived content and plans to address new LDL content streams. Renewed attention to documentation is important given issues related to staff turnover [S].

• Turnover of staff (particularly in IT) highlighted by a number of interviewees as challenge to continuity and effective governance of DLS. Many of the staff who designed and developed the DLS are no longer working at the BL. • Reorganization of IT, particularly around application support, has provided more flexibility in allocating resource to the right areas.	

[4b] A clear and effective system of governance is in place to manage and develop the repository over time.

Key Evidence	Finding
• Failures relating to core preservation processes appear to have been caused largely by governance issues. More specifically, a lack of management information reporting, a lack of communication and escalation of issues, and insufficient priority given to long-term preservation (also see Metrics 1 and 2, above). • A number of key actions have been taken to address governance issues, centred around the creation of a Repository Manager role and forums for escalating and planning mitigating actions for identified issues (such as the Repository Management Service Group). • DLS shortcomings with regard to areas such as management information reporting, role-based access and oversight for key staff across the lifecycle and storage architecture have been identified and are being captured in	Changes to the governance of DLS and related preservation processes have resulted in the identification of preservation issues and the instigation of mitigation work. Awareness of digital preservation and the prioritization of preservation issues have improved. Lessons learned from the DLS era of digital preservation at the BL are clearly being fed into the forthcoming DAMPS Programme. Responsibilities for essential preservation checks (such as completeness checking) could be clarified. There is the potential for the LDL structure to contribute greater preservation oversight [A].

requirements for the DAMPS Programme. • Some staff reported making use of Management Information reporting on the progress and status of content through the ingest process. Others described DLS as a black box, and were frustrated by the complete absence of reporting in this area. Some curators reported no access to DLS (other than public user access). • Core business-as-usual processes have previously suffered due to prioritization of resource to projects. A recent change has been made to maintain more IT resource on business-as-usual activities. This appears to be having an impact on addressing issues such as replication backlogs (see Metric 2a). Anecdotal evidence suggests preservation issues have not been discussed with any great frequency or in great detail at the Legal Deposit Implementation Group (LDIG).	

4.5 Expert Guidance

Requirement: Mechanisms are in place to ensure skills and expertise of relevant staff are up to date *(adapted from DSA R6)*.

[5a] Staff development/training mechanisms ensure staff have appropriate and up-to-date digital preservation expertise.

Key Evidence	Finding
• New members of DPT attend DP training. • Project Managers receive project management training. • Technical staff receive technical training in order to stay up to date, e.g. data analysis.	The DPT maintain up-to-date training in methods of DP through external and internal courses. New staff who have not already attended an industry-standard DP course (e.g. DPTP at Co-Sector or the online Dundee University module) attend one as soon as possible after starting. Similarly,

• DPT provide training for other staff, such as an Introduction to Digital Preservation within BL, and an OAIS (see Glossary) course.	project managers and technical staff attend periodic training to keep skills and knowledge up to date. DPT also deliver training to other staff across the BL to ensure DP can be applied where relevant. Overall, the staff development and training mechanisms at the BL demonstrate a high level of dedication to maintaining and growing the skillsets of staff in DP and other relevant areas [S].

[5b] Appropriate support is provided by sources of external guidance, support and review.

Key Evidence	Finding
• DPT hold memberships of IIPC[11], OPF, and DPC. • The DPT has a dedicated training budget. • Funds for are available for conference attendance.	In addition to the BL's dedication to ongoing training, appropriate funding and support also ensure the retention and growth of knowledge in DP-related areas. Because of the quickly developing nature of methods and technology in DP, the BL has allotted a dedicated budget to training for DP. The BL also provides funds for attending conferences and for memberships of community organizations, which ensure engagement with external developments in DP and access to shared solutions. The external guidance, support, and review provided by the BL demonstrates a very mature DP programme [S].

[5c] Good awareness of state-of-the-art in technology and techniques for digital preservation.

Key Evidence	Finding
• DPT keep up to date with sector relevant listservs and other resources. • DPT contribute to Digital Scholarship Programme through speaking on DP issues.	The DPT demonstrate a high level of engagement with resources in the overarching DP community in the UK and abroad, such as listservs and wikis. This engagement ensures the DPT stays aware of new and emerging technology and techniques relevant to DP. Often, because

[11] http://netpreserve.org/

• DPT regularly participate in external programme committees deciding what will happen at sector-relevant conferences.	the BL has a relatively mature and robust DP programme, the DPT is itself at the forefront of these new developments. The team is very active in sharing this knowledge internally through the Digital Scholarship Programme, and externally through acting on programme committees for conferences [S].

4.6 Technical Infrastructure

Requirement: The repository functions on well-supported operating systems and software and is well managed with the application of appropriate processes and standards *(adapted from DSA R15)*.

[6a] Repository application is well managed and fit for purpose.

Key Evidence	Finding
• Large numbers of items are flagged and held in various ingest queues, as part of everyday operation of the DLS, requiring manual intervention to resolve. It was reported that considerable numbers were held for months or years (although this situation was reported to have significantly improved more recently). • Staff painted a mixed picture of the legacy documentation available that describes the DLS and its processes. With the recent example of investigating and addressing issues with the OAC process, documentation on the OAC was reported to be difficult to find and incomplete. • Software development processes and associated source code management have seen gradual improvements since 2013, noting in particular a review of the Agile development and project management process, documentation of work in progress	Improvements to the software development and project management processes within IT are positive and demonstrate a commitment to continued process improvement, despite resourcing pressures. Holding content for long periods in ingest queues without resolution is cause for concern, particularly given an uncertain picture around completeness checking across the lifecycle (see metric: 1a). This adds to the uncertainty around effective preservation for any item ingested to the DLS. Changes to resourcing and prioritization of this work suggest significant improvement however [A].

allowing work to be put on hold and picked up again at a later date if necessary, and the establishment of a dedicated test team. • Note increased resourcing focus on business as usual activities (see metric 4b).	

[6b] Metadata profiles are fit for purpose and meet functional needs of preservation and access.

Key Evidence	Finding
• Metadata profiles are standards led (see metric 6b). • No specific issues reported or identified around LDL collection metadata not being fit for purpose.	The metadata profiles were not examined by this review, but appear to be satisfactory. The access and preservation functions appear to be adequately supported (although note impact of LDL user access restrictions, Metric: 3a, above) [S].

[6c] Appropriate standards are applied to preservation processes (e.g. architecture, metadata).

Key Evidence	Finding
• Metadata profiles utilize METS, MODS and PREMIS (see Glossary). • OAIS underpins DLS design and provides the blueprint for planned preservation workbench development. • ISO16363 provided basis for self-assessment and digital preservation process improvement across the BL.	The BL's digital preservation activities are sufficiently founded on appropriate standards [S].

4.7 Infrastructure Security

Requirement: The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users *(adapted from DSA R16)*.

[7a] An appropriate disaster recovery plan is in place to address a major fail of services.

Key Evidence	Finding
• Disaster recovery plan in place, but identified as in need of a refresh.	Appropriate effort has been directed to disaster recovery. Preservation and

• Revised disaster recovery plan under development (target completion mid-summer 2017). This aims to consolidate several different elements to ensure no issue falls between the cracks. • Revised disaster plan will be tested. • 4-node DLS design focuses on ensuring content survives disaster.	survivability have been identified as the priority in a recovery situation. The current review of the recovery plan is deemed prudent by the BL given the significant growth in the DLS (from terabytes to petabytes) since the establishment of the current plan [S].

[7b] Appropriate mechanisms are in place to ensure the cyber-security of the repository and these are independently verified.

Key Evidence	Finding
• Layered security model is cornerstone of DLS design. • Regular independent cyber-security review, with evidence of identification of issues and their mitigation. • Recent production of documentation map and other documentation improvements. • Access monitoring in place to identify unusual patterns of usage. • LDL content held in intermediate storage is not subject to same security precautions as DLS content.	Security issues are carefully monitored with a variety of systems and processes in place to sufficiently address obvious cyber-security concerns. Content not held in DLS remains a cause for some concern, but ingest projects are already in process for these streams as noted above [S].

[7c] An appropriate succession plan, contingency plans, and/or escrow arrangements in place in case the repository/organization ceases to operate.

Key Evidence	Finding
• The nature of the BL's funding and legislation-led mandate to preserve LDL collections makes the closure of the repository and/or a cease in funding at short notice highly unlikely.	This is not considered to be a concern given the nature of the BL and its LDL preservation activities [S].

5 Initial Recommendations

This section of the report provides specific actions based on the assessment of preservation capability against the criteria (see section 4). It was provided to the LDLs in May 2017.

5.1 Completeness checking across the lifecycle (see Metric 1a)

A completeness check that compares acquired/deposited content with the contents of the DLS is necessary to verify that no NPLD content is missing. Anecdotal evidence suggests eBook content may have been lost, and this should be verified. The copy of web archive content held in the UK Web Archive Hadoop cluster provides the potential to perform a substantial historical completeness check of a large volume of material. Given the other issues identified in the recommendations below, this would help to validate the effectiveness of the DLS and associated ingest processes. It would help with identifying any other problems with DLS operation. Moving forward it would be prudent to consider how completeness checking can be instigated for future content.

A. ACTION: Investigate and clarify whether eBook content has been successfully ingested. It is noted that some investigation work has already been undertaken.
B. ACTION: Building on the work conducted to check the 2015 domain crawl, perform a further completeness check of web archive content. Ideally this should cover a larger volume of content (if possible the whole web archive collection). Any crawled items that do not appear in the store should then be investigated.
C. ACTION: Consider how best to incorporate completeness checking into future preservation activity, either in remedial work to the DLS or in terms of requirements and planning for DAMPS.

5.2 Integrity checking (see Metrics 2b and 2f)

Integrity checking of files stored on the DLS nodes, along with any automated repair of damaged or missing files, is an essential preservation function that must be reinstated as soon as possible. Remedial work was ongoing at the time of this iteration of the review, with roll out of a new OAC in progress. This must now be carefully monitored, and associated policy and reporting put in place.

D. ACTION: Monitor roll out of new OAC, particularly with regard to performance and projected time to check a complete DLS node.
E. ACTION: Retain OAC logs at least for the short term and confirm a policy for appropriate retention periods. Given the problems encountered to date, a minimum of 10 years is recommended.
F. ACTION: Verify that new OAC reporting is adequate to ensure correct/expected OAC operation.
G. ACTION: Develop policy on acceptable OAC performance/time to complete full integrity check.

5.3 Legal Deposit Regulations with respect to Access and Rendering Software Deposit (see Metric 3a)

This report describes some of the substantial challenges faced by the BL and the other LDLs in preserving digital materials under the Legal Deposit Regulations. Two particular issues stood out that warrant further consideration in the context of the upcoming review of the Regulations.

H. ACTION: Consider what changes could be made to facilitate the acquisition and preservation of software (and related metadata) necessary to enable the rendering (and ultimately the preservation) of legal deposit collections.
I. ACTION: Consider what changes could be made to increase the accessibility of legal deposit materials without damaging the interests of the publishers. This appears particularly relevant to

the web archive stream, where increased access appears to present a minimum of conflict with publishers.

5.4 Governance of preservation operations (see Metric 4b)

Appropriate management, monitoring, escalation and resourcing of staff and processes related to operational preservation is essential to ensure preservation activity is performed effectively, efficiently and meets with preservation policy. Substantial and positive changes have been instigated over the last two years to keep governance and processes aligned, and these should continue to be carefully monitored and reviewed to ensure expectations are met.

J. ACTION: Monitor the effectiveness of the governance of operational preservation, particularly in relation to the management and reporting of ongoing technical and process changes covered by the other recommendations in this report.
K. ACTION: Review the escalation and reporting of preservation issues within LDIG and associated LDL groups and consider if additional useful oversight could be provided by the other LDLs.
L. ACTION: In the context of Recommendation 5.1 (above), ensure roles and responsibilities that relate to completeness checking are clearly defined. It is noted that this may be complex, given the scope of performing checks right across the preservation lifecycle, across multiple staff remits, and across multiple preservation processes.

5.5 Management of DLS ingest and replication processes (see Metrics 2a and 6a)

Avoiding backlogs in content as it moves through various DLS processes is essential if completeness checks are to be performed to ensure all content has been preserved as it should be. Action has already been taken to monitor key ingest queues and node replication processes and to address backlogs. This should continue to be monitored closely and action taken as necessary.

M. ACTION: Ensure backlogs in DLS ingest and replication processes are kept to a minimum.
N. ACTION: Develop policy on reporting and escalation of processing backlogs with respect to acceptable levels/timescales.

6 Review of the Response to Initial Recommendations

This section of the report provides a review of the response by the BL and the other LDLs to the Initial Recommendations (see section 5) supplied by this project in May 2017. The original recommendations are presented in brief alongside evidence of the response. The findings of the review, based on this evidence, are also presented. Note that only the metrics that required further action are listed here.

6.1 Completeness checking across the lifecycle (Recommendation 5.1)

H. Investigate and clarify whether eBook content has been successfully ingested. It is noted that some investigation work has already been undertaken.

Key Evidence	Finding
• Missing content has been identified against manifests (see Glossary) (some manifests re-acquired as needed) and a list of missing items has been compiled. Publishers to re-supply missing content [evidenced by direct reports to the review]. • New development work and changes to the ingest workflow have been undertaken to improve cross checking of content following submission to the BL. This functionality is now operational. A live demo of this functionality was provided. Content is pushed by publishers to the BL by FTP, and a manifest is emailed to the BL. A short-term archive of this content is now created, allowing the BL to backtrack in the event of problems encountered further down the line. Cross checking of the submitted content against the manifest identifies duplicate or missing content automatically [evidenced by direct reports to the review and a live demonstration by relevant BL staff].	Investigation and remedial activities appear to be progressing well. There has been speedy and impressive IT development work to improve processes, cross checking and to ensure issues are (where reasonably possible) detected and addressed in future.

I. Building on the work conducted to check the 2015 domain crawl, perform a further completeness check of web archive (WA) content. Ideally this should cover a larger volume of content (if possible the whole web archive collection). Any crawled items that do not appear in the store should then be investigated.

Key Evidence	Finding
• A clearer picture of the ingest of Web Archive (WA) content to the DLS has emerged. Matching objects in the WA cluster to content in the DLS has confirmed to date that the vast majority of ingested objects are present in the DLS. Only 2262[12] were not matched in DLS, having either not been submitted for ingest as thought or lost in the chain of ingest processes. These remaining objects are being investigated. [evidenced by direct reports to the review]. • A new ingest workflow has been implemented for WA content that will provide more validation, including an inventory check between content on the WA cluster and the DLS [evidenced by direct reports to the review]. • Closer working between the WA team and IT has been beneficial in aiding understanding of challenges in both areas and facilitating more effective working [evidenced by direct reports to the review].	A thorough item-matching check across the NPLD WA content has identified some completeness issues which have now mostly been remedied. Remaining ingest of the small number of missing objects and a full integrity check are immediate goals. The WA team have developed a new workflow to enhance the tracking of WARC files during ingest. This new workflow will increase communication with the team that governs ingest of content into the DLS, will introduce a check of this content against the WA inventory, and will provide the opportunity to validate the results of those checks. Once implemented, this new workflow, supported by the newly drafted Completeness Checking Policy, will provide better monitoring and oversight for the WA team.

[12] Less than 0.01% of WA content by object count

J. Consider how best to incorporate completeness checking into future preservation activity, either in remedial work to the DLS or in terms of requirements and planning for DAMPS.

Key Evidence	Finding
• New Completeness Checking Policy governed by the Digital Collections Assurance Group mandates completeness checking of existing NPLD content streams. [evidenced by a draft policy submitted to the review]. • Ongoing projects to add a DLS ingest capability for new content streams (maps and sheet music) have not been adapted to address completeness checking and are clearly identified as out of scope within the Completeness Checking Policy. The intention is to roll out completeness checking to other content streams as is appropriate [evidenced by direct reports to the review]. • Completeness checking has not been identified as a specific requirement within the DAMPS Programme, but is within the broader aims of the project. It is expected that DAMPS will provide better support for completeness checking than is currently available [evidenced by direct reports to the review].	The Digital Collections Assurances Group, a body designed to improve oversight of digital processes, is now reviewing a new Completeness Checking Policy. Based on a review of a draft of the policy, this new documentation will provide a mandate for monitoring the completeness of digital objects acquired from depositors. It will further support a process and procedures (still under development) for addressing problems and escalating them when necessary. The new policy puts in place the foundation for ensuring the completeness of NPLD ingests from deposit through to preservation and access.

6.2 Integrity checking (see Recommendation 5.2)

K. Monitor roll out of new OAC, particularly with regard to performance and projected time to check a complete DLS node.

Key Evidence	Finding
• A live demo was provided of the re-written functionality for integrity checking and fixing missing/broken files. Running on the test server, the reviewers observed the OAC find artificially modified or deleted files and fetch replacements from an alternative node (recovery). Reviewers were able to request particular scenarios be executed, ask questions on the detail of what was seen and discuss how the new OAC addresses issues previously encountered with the old application [evidenced by a live demonstration by relevant BL staff]. • A complete integrity check of all DLS nodes had not completed by the end of this review, but good progress has been made. The Boston Spa node completed an integrity check in safe mode (i.e. without automated recovery – just detection of content files or signature files that fail to match their checksums (see Glossary)) with 2184 failed checks. The London node has completed in safe mode with 2713 failed checks, and is part way through a run with automated recovery. The Wales node has completed in safe mode with 328 failures. Scotland has not completed due to a delay associated with problems installing the new version of the OAC. A cross check of failures between the three nodes that have completed suggests that there are only four objects that are not present on any of those three nodes. Initial investigations suggest that most integrity check failures are associated with missing signature files [evidenced by direct reports to the review].	The re-write of the OAC appears to have been successful. Careful testing is ongoing, and integrity checks have completed on three of the four DLS nodes. The combined problems described in the Initial Assessment (see sections 4 and 5) of this review (detailing a lack of integrity checking and substantial backlogs in ingest and replication) could have resulted in significant loss of NPLD content. Integrity checking with the new OAC has revealed this not to be the case. The reviewers were apprehensive that a complete integrity check of at least one node might not be possible within the timescale of the review, which would have left uncertainty around the status of the preservation store. On the contrary, the performance of the new OAC has exceeded expectations.

L. Retain OAC logs at least for the short term and confirm a policy for appropriate retention periods. Given the problems encountered to date, a minimum of 10 years is recommended.

Key Evidence	Finding
• Integrity checking is undertaken once per quarter and OAC logs will be retained for one year. Management reporting information based on the logs (key stats, totals and summary figures) will be retained indefinitely [evidenced by direct reports to the review and a draft Completeness Checking Policy submitted to the review].	Given the mandated frequency of integrity checking (once per quarter) and logs retained for a year, there is an audit trail that includes at least three previous integrity checks. The reviewers believe this solution appears to be fit for purpose.

M. Verify that new OAC reporting is adequate to ensure correct/expected OAC operation.

Key Evidence	Finding
• The OAC operations are being assessed against the ISO 16363 Standards to ensure they meet requirements, including reporting. The DPT has final sign off on the OAC development [evidenced by direct reports to the review and OAC User Stories document submitted to the review]. • The new Fixity Checking (see Glossary) Policy defines a series of specific data that will be reported to the Repository Service Management Group (RSMG), including a summary report of the status of each node [evidenced by a Fixity Checking Policy submitted to the review]. • Reporting was under development at the close of the review [evidenced by direct reports to the review].	Reporting facilities had not been completed by the close of the review but the foundation for reporting is in place and work appears to be progressing well.

N. Develop policy on acceptable OAC performance/time to complete full integrity check.

Key Evidence	Finding
• The new Fixity Checking Policy defines a schedule, reporting requirements, responsibilities and audit for integrity checking of NPLD content [evidenced by a Fixity Checking Policy submitted to the review]. • Discussion still in progress as to whether this Policy (or another, in parallel) should provide specific requirements for authenticity checking [evidenced by direct reports to the review].	The Policy requires a complete integrity check every quarter. Best practice depends on many factors, making peer comparisons difficult.[13] However, in the context of typical operations at other organizations, this is a strong commitment and is very encouraging.

6.3 Legal Deposit Regulations with respect to Access and Rendering Software Deposit (see Recommendation 5.3)

O. Consider what changes could be made to facilitate the acquisition and preservation of software (and related metadata) necessary to enable the rendering (and ultimately the preservation) of legal deposit collections.

Key Evidence	Finding
• Lack of software deposited with data (and lack of copyright exceptions for use of software to render/use NPLD content) continues to impose challenges, and these are expected to grow in the future [evidenced by direct reports to the review and a challenges document created by NLS submitted to the review]. • The acquisition and preservation of software has been discussed at LDIG. The need for software to help preserve and provide access to content will be mentioned in the Post-Implementation Review document, but is not one of the main pillars of the review [evidenced by direct reports to the review].	This issue is for the most part out of the hands of the LDLs, but it remains clear that preservation will become increasingly difficult, if not impossible, without greater leeway on collecting and applying access software to enable use of NPLD content. The issue is addressed further in section 8 Emergent Challenges.

[13] See http://www.dpconline.org/handbook/technical-solutions-and-tools/fixity-and-checksums and https://blogs.loc.gov/thesignal/files/2014/02/NDSA-Checking-your-digital-content-Draft-2-5-14.pdf

P. Consider what changes could be made to increase the accessibility of legal deposit materials without damaging the interests of publishers. This appears particularly relevant to the web archive stream, where increased access appears to present a minimum of conflict with publishers.

Key Evidence	Finding
• No further evidence gathered since phase two.	As above, this issue is largely outside the realm of the LDLs, but it seems clear that improving access would have a beneficial impact on preservation. This issue is addressed further in section 8 Emergent Challenges.

6.4 Governance of preservation operations (see Recommendation 5.4)

Q. Monitor the effectiveness of the governance of operational preservation, particularly in relation to the management and reporting of ongoing technical and process changes covered by the other recommendations in this report.

Key Evidence	Finding
• Ongoing remedial work – both work that began prior to this review as well as new work specifically addressing the review's recommendations – has been centrally co-ordinated and monitored. Key members of staff have taken ownership of specific tasks related to each recommendation and have provided clear evidence for how the Library has responded to the review [evidenced by direct reports to the review and a working documentation submitted to the review]. Three key management groups have responsibility for overseeing digital preservation activities: Collection Management Group (CMG), Repository Service Management Group (RSMG) and the Digital Collection Assurance Group (DCAG). The Digital Preservation Team is well represented on these groups and digital preservation is now a standing item on the CMG agenda. The RSMG is chaired by the Head of Digital Preservation and meets weekly to discuss operational digital preservation and other processes relating to the DLS operation [evidenced by direct reports to the review and references in new Policy documents submitted to the review]. • Despite a continued squeeze on resources and a heavily loaded IT department, support for outstanding digital preservation issues identified by this review has been prioritized [evidenced by direct reports to the review] • Recent work to address ingest issues (e.g. systems development to enhance eBook ingest cross checking) indicates improved communication (and resolution) of DLS operational issues [evidenced by direct reports to	This recommendation is on a longer time scale than the others, so the effectiveness of the response cannot be completely verified in the relatively short timeframe of this review. The significant test will be a future response from the BL to a significant and unexpected digital preservation issue. In the absence of that test, it appears that the BL has done all it can to be best placed in the event of future challenge.

the review and by the speed with which new solutions have been made operational across teams].	

R. Review the escalation and reporting of preservation issues within LDIG and associated LDL groups and consider if additional useful oversight could be provided by the other LDLs.

Key Evidence	Finding
• At the last LDIG meeting (21/11/2017), members discussed creating a new sub-group to report to LDIG which would focus on digital preservation. The Chair was actioned with considering an appropriate scope and structure. Further discussion is expected at the meeting of Chief Librarians on 1/12/2017 [evidenced by direct reports to the review].	The creation of a dedicated digital preservation LDIG sub-group would provide much needed focus on arising digital preservation issues, as well as useful oversight of key operations and reporting. The sub-group would also likely provide a useful forum for the different LDLs to work more closely on this topic, which is an encouraging development.

S. In the context of Recommendation 1 (above), ensure roles and responsibilities that relate to completeness checking are clearly defined. It is noted that this may be complex, given the scope of performing checks right across the preservation lifecycle, across multiple staff remits and across multiple preservation processes.

Key Evidence	Finding
• This issue was discussed with various BL staff who acknowledged the challenge. Responsibility has clearly been identified and assumed by various areas across the lifecycle. Oversight is provided by a number of management groups, and ultimately the DCAG [evidenced by observations of changing practices and direct reports to the review].	Appropriate governance across the lifecycle appears to have been established.

6.5 Management of DLS ingest and replication processes (see Recommendation 5.5)

T. Ensure backlogs in DLS ingest and replication processes are kept to a minimum.

Key Evidence	Finding
• Considerable effort has been focused on reducing backlogs, both in ingest and in replication. This effort appears to have placed further strain on a heavily loaded Application Support (IT) team [evidenced by direct reports to the review]. • Replication between the BL nodes are reported to be 'minimal'. Figures reported from the end of November 2017 reveal an NLS backlog of 200GB, 80000 files and a backlog of 300GB, 800000 files at the National Library of Wales (NLW). The ingest backlog for eBooks stands at 9000 items at the beginning of December. Figures were not available for the other streams [evidenced by direct reports to the review].	Replication and ingest backlogs do not yet meet the requirements specified in the relevant new policy (see below) but the aim is to be compliant by 1 January 2018. Reports seen by the DPC provided totals of current backlogs but did not reveal the length of time items had been held in backlogs, which will obviously be critical in meeting time constraints specified in the policy.

U. Develop policy on reporting and escalation of processing backlogs with respect to acceptable levels/timescales.

Key Evidence	Finding
• The NPLD Ingest and Replication Backlog Policy defines strict time limits within which ingest and replication should be completed, as well as detailing responsibilities, reporting and audit requirements [evidenced by an Ingest and Replication Backlog Policy submitted to the review]. • Currently, backlog reports are sent quarterly to LDIG and to the Service Assurance team.	This new Policy places strict time limits for ingest and replication backlogs and clearly defines when issues should be reported and escalated, with a clear statement of responsibilities. The BL aims to implement the policy from 1 January 2018.

7 Final Conclusions

The LDLs have been subject to a rigorous independent review based on international standards to assess the effectiveness of digital preservation practice as applied to the Non-Print Legal Deposit Collections, a requirement that derives from the UK Legal Deposit Libraries (Non-Print Works) Regulations 2013. Largely delivered by the BL on behalf of the other LDLs for NPLD Collections, this digital preservation practice has been found to be exemplary and world leading in many cases.

This conclusion has been made on the basis of two phases of review. An Initial Assessment which reported significant strengths but identified a small number of operational failings and a Review of the Response to the Initial Recommendations which showed the excellent progress that has been made by the BL on all of the recommendations made in the Initial Assessment. A substantial number of activities to address preservation shortcomings were already in progress when this review began. This existing momentum towards quality improvement has been enhanced through the independent scrutiny afforded by this assessment and has been expanded by the addition of actions specifically relating to the recommendations made throughout the review process. The overall impression gained from this assessment has been one of excellence. In the small number of cases where weaknesses were identified, responses have been timely and effective and have either met or exceeded reviewers' expectations. The reviewers have come to their independent judgement with full access to all relevant documents, processes and staff at the LDLs, especially at the BL. They are happy to confirm that the LDLs, and the BL in particular, have a firm commitment to continuous quality improvement, respond quickly and openly when shortcomings are identified, and that their digital preservation practice is exemplary in many respects.

The following list provides a summary of key observations about the existing remedial work and new actions triggered by the response to the Initial Recommendations:

- Outstanding preservation issues requiring immediate action have been addressed or are in good progress (in particular: integrity checking, backlogs, missing eBooks, WA ingest). (Indeed, some of these actions were already in development and would have progressed irrespective of the external review).
- Additional resources from across the BL have been devoted to addressing the various issues raised in the review and addressing them in an impressively short space of time.
- The impact of the failings outlined in the Initial Recommendations (see section 5) appears to be minimal. While full integrity checks on all DLS nodes are yet to complete, it appears that no more than four content files of all the NPLD content have been lost.
- New policies have been established to guide future activity where the direction has previously been lacking or unclear.
- The priority given to NPLD digital preservation requirements has increased significantly over the last two years at the BL.
- Changes to the governance of the DLS and associated processes have improved the oversight of operations key to digital preservation, made responsibilities more clear and established lines of escalation in the event of significant issues.
- The Digital Preservation Team has stronger representation in key decision-making areas – where prioritization of essential work has previously been an issue.
- Clear thinking and careful planning is driving forward the work on DLS replacement. It is clear that lessons have been learned from the wealth of experience from implementing and delivering preservation via the DLS. However, much still depends on the technology chosen to replace the DLS.

Nonetheless, these strong conclusions do not mean that the BL or other LDLs should relax their efforts or reduce their obligation to continuous quality improvement – on the contrary, they illustrate the importance of a sustained commitment to ongoing management of change in order to meet, and where appropriate exceed, the standards set by the new policies that are being introduced. The following areas will be critical over the short to medium term:

- implementation of completeness checking across the NPLD lifecycles;
- implementation of planned management information reporting on key digital preservation measures;
- continued commitment to meeting digital preservation requirements via the improved reporting, governance and prioritization of necessary digital preservation activities;
- sustained commitment of resources to operational (business as usual) repository maintenance activities, particularly with respect to integrity checking and addressing ingest and replication exceptions.

It remains important for the BL to complete the remedial work and maintain the focus on careful validation of its preservation operations as the DAMPS Programme progresses. The establishment of a new digital repository, and the complex process of migrating data and metadata to it, will hopefully provide a significant enhancement of digital preservation processes, but will also present a major challenge. A second review would provide a vital mechanism for verifying the standard of continuing digital preservation operations for NPLD materials.

In testing this ongoing commitment, the reviewers recommend that the LDLs undertake a second independent review, similar in design and scope to this one within 24 months following the establishment of the replacement for DLS.

8 Emergent Challenges

This section provides some brief thoughts on the challenges facing digital preservation at the BL and the other LDLs in the medium to longer term.

8.1 NPLD Regulations: constraints and opportunities

In delivering a significant programme of work to address the preservation of NPLD content, the LDLs are of course bound by the constraints of the Regulations themselves. These Regulations understandably place considerable restriction on how this preservation work (and subsequent access) is implemented. A recurring theme during almost all discussions held during this review was the difficulties faced in delivering effective and economical digital preservation within these constraints. Considerable onus is placed on the LDLs to adapt to the content and the considerable variations in size, shape, packaging and, most importantly, metadata, supplied (typically as is) by the publishers. The DPC has identified two key areas where the Regulations inadvertently constrain or defeat reasonable digital preservation actions. Small changes to the Regulations would make digital preservation practices more straightforward, more cost effective, and more responsive:

- The severe restrictions on user access to NPLD content impacts directly on digital preservation by limiting usage and user feedback on any problematic content. There is a huge volume of material within scope of preservation and, as a result, a daunting challenge of quality assurance. Lifting some access restrictions – such as on-site only access for web archive material – would provide considerable digital preservation dividends with minimal impact on publishers (due to the nature of content published openly on the Web in the first place). Digital preservation standards place considerable emphasis on the extent to which a facility can meet and adapt to the needs of a 'designated community' of users. The current Regulations are a significant impediment to the required feedback between the LDLs and the designated community.
- As NPLD content becomes more complex and published data becomes gradually more entwined with the software that creates and renders it (see section 3.2), the provision (and indeed preservation) of software will become more critical. The UK has fallen behind many other western countries where legal deposit of software, or even preservation exceptions to copyright restrictions regarding software, is in place. Software preservation (and the provision of software to enable access to preserved content) must be addressed within the regulations if the preservation of NPLD is to remain viable over the coming decades.

8.2 The LDLs: Partnership and Collaboration

Much of the focus on this review has been on the digital preservation operations that are largely based at the BL. During Phase 4 the DPC was asked by the LDLs to also consider the broader picture and look at how the BL works with the other LDLs. Whilst being collaborative in nature, for the most part it is a relationship also based on service provision. In interviews for this review, many staff at both the BL and NLS talked candidly about this relationship. Although collaboration has not been common across the board, there have been particular areas, typically collection and curatorially focused, where close working partnerships have paid dividends. Visibility and oversight were also discussed with a number of staff. Many at the BL and the NLS felt that there had been missed opportunities (by all parties) to exchange information on future planning, reducing the opportunity for feedback to the BL from experts at the other LDLs. As an example, little detail of the planning and requirements for the DAMPS Programme has been shared with key domain expert staff at the other LDLs. This review suggests that encouraging more collaboration and enabling constructive oversight of the BL from the other LDLs would be a very positive step. It also notes that the (at the time of writing) proposed establishment of a LDIG sub-group addressing digital preservation issues may act

as a catalyst to enable this. Communication remains a challenge, particularly for the BL. It is of course difficult to transfer knowledge of complex preservation operations; however, a number of staff at the BL and the other LDLs noted that conversations with the DPC reviewers had revealed significantly more about the actual workings of NPLD preservation than they had previously encountered. Clearly more could be done in this area.

8.3 Resourcing for the digital preservation of NPLD content

The adequate resourcing of digital preservation activities is obviously an important consideration, and one that this review was invited to consider. Assessing the adequacy or otherwise of funding for an operation that spans many teams, departments and indeed institutions, is a complex one, but the indications from this extensive review of staff, documentation, and in some cases process operation, has been largely positive. In particular, the BL's Digital Preservation Team, which plays an increasingly important role of policy direction and oversight, has seen a growing resource in recent times. However, it is clear that recent budget pressures at the LDLs have had an impact and this appears in particular to have placed a strain on IT operations at the BL. A constant challenge has been to balance work on new projects with maintaining a sufficient level of service in 'business as usual' activities – activities where key failings had been identified by this review's Initial Assessment. To an extent, this is a vicious cycle because project investment in process improvement will typically provide longer-term benefits in reducing business as usual activities. As noted in section 5, it is important for the BL to sustain sufficient resources to maintain appropriate standards of preservation. Over the medium and longer term there must be concern for the stresses and strains that will no doubt be encountered. These stresses and strains will only increase as the LDLs expand the range of content targeted for preservation and move from the theoretically lower hanging fruit of eJournals and eBooks to more complex data types. The BL has successfully implemented workflow and process improvement for these notionally simpler, but obviously in practice still very complex, content types. It has learned a great deal about the challenges of the content, the variability of the materials and the expected quality (or otherwise) of supplied metadata. But this learning process and the subsequent improvements to live workflows have been hard won. It is expected that the moves forward to addressing emerging content will be fraught with difficulty and this will require considerable resource to address effectively.

9 Appendix One: List of Interviewees

Initial Assessment (phase 2), British Library:

- Linda Arnold-Stratford, LDL Liaison Manager
- Alasdair Ball, Head of Collection Management
- Claire Caulfield, Digital & Prints Operations Manager
- Paul Clements, Head of Architecture and Design
- Ian Cooke, Head of Contemporary British Published Collections
- Chloe Crowder, Database Administrator
- Andy Davis, Legal Deposit Publications Liaison & Imp Coordinator
- Mark Dawson, Head of Service Assurance
- Lee Edwards, Head of IT
- Phil Hatfield, Lead Curator Digital Map Collections
- Andy Jackson, Web Archiving Technical Lead
- Steve Lenton, Head of Infrastructure Services
- Maureen Pennock, Head of Digital Preservation
- Amelie Roper, Curator Digital Music (on behalf of Richard Chesser, Head of Music Collections)
- Caylin Smith, Repository Manager (at time of interview)
- Neil Wilson, Head of Collection Metadata
- Kevan Wood, Head of Application Development
- Andy Woodward, Senior Software Engineer

Review of Response to Initial Recommendations (phase 4), British Library

- Linda Arnold-Stratford, LDL Liaison Manager
- Alasdair Ball, Head of Collection Management
- Jason Barry, Technical Services Manager, Ingest & Metadata
- Imran Choudry, Application Specialist, Ingest & Metadata, DLS Support
- Paul Clements, Acting Head of IT/Head of Architecture and Design
- Paul Connor, Solutions Architect
- Kevin Davies, Technical Analyst Digital Preservation
- Kevin Foody, Acquisitions Processing Manager
- Andy Jackson, Web Archiving Technical Lead
- Sharon Johnson, Head of Content Development Implementation
- Andrew McEwan, Head of Content & Metadata Processing
- Maureen Pennock, Head of Digital Preservation
- Kathy Raynor, Application Specialist, Ingest & Metadata, DLS Support
- Caylin Smith, Legal Deposit Libraries Senior Project Manager
- Philip Swift, Application Specialist, Ingest & Metadata, DLS Support
- Paul Thurlow, Digital Processing Team Manager
- Tim Wood, Application Specialist, Ingest & Metadata, ILS Support

Review of Response to Initial Recommendations (phase 4), National Library of Scotland

- Chris Fleet, Map Curator
- Graeme Forbes, Head of Collection Management
- Graham Hawley, General Collections Manager

- Lee Hibberd, Digital Preservation Officer
- Stuart Lewis, Head of Digital
- David White, Interim IT Infrastructure Manager

10 Appendix Two: Metrics

1. Content Preservation

- Preservation risks are mitigated by identifying, assessing and taking action to ensure content is understandable, sustainable and accessible to users *(adapted from DSA R3, DSA R10 and DSA R11)*.
    - Appropriate checks are applied to ensure quality and completeness of content on deposit/acquisition and subsequently ingest, where possible.
    - Appropriate format identification, validation and other content characterization processes are applied as appropriate.
    - Technology watch and other appropriate monitoring activities are implemented.
    - Content preservation risks are assessed and identified.
    - Action is taken to mitigate preservation risks and ensure users can access and exploit content.
    - Content preservation processes and risk mitigation actions are carefully managed and documented. Issues are documented, reported, and escalated as appropriate.
    - Have formal documentation of policies and procedures for implementing digital preservation across the organization.

2. Integrity and Authenticity

- Bit-level integrity and authenticity is ensured via replication and integrity checking processes *(adapted from DSA R7 and DSA R9)*.
    - Content is replicated to minimize impact of any bit loss.
    - Integrity checking ensures any bit loss can be identified.
    - Appropriate mechanisms are in place to minimize risk of accidental deletion or malicious damage.
    - Technology choices and storage architecture minimize the risk of loss due to (common) hardware failure. Storage refreshment is managed conservatively.
    - Appropriate encryption, signing and related processes ensure authenticity of content and metadata.
    - Bit-level preservation processes and risk mitigation actions are carefully managed. Issues are documented, reported, and escalated as appropriate.

3. IPR and Regulatory Constraints

- IPR and LDL Regulatory constraints are monitored and managed to minimize impact on preservation of the collections *(adapted from DSA R2)*.
    - LDL regulatory constraints do not place undue restriction on effectiveness of preservation or burden on resourcing of preservation.
    - IPR and Data Protection constraints are well understood and managed.

4. Organizational Infrastructure

- The repository has sufficient resources managed through a clear system of governance to effectively carry out effective digital preservation *(adapted from DSA R5)*.
    - Staffing and funding are sufficient to sustain the repository, enable effective preservation and ensure permanent access to the collections.
    - A clear and effective system of governance is in place to manage and develop the repository over time.

5. Expert Guidance

- Mechanisms are in place to ensure skills and expertise of relevant staff are up to date *(adapted from DSA R6)*.
    - Staff development/training mechanisms ensure staff have appropriate and up-to-date digital preservation expertise.
    - Appropriate support is provided by sources of external guidance, support, and review.
    - Good awareness of state-of-the-art in technology and techniques for digital preservation.

6. Technical Infrastructure

- The repository functions on well-supported operating systems and software and is well managed with the application of appropriate processes and standards *(adapted from DSA R15.)*
    - Repository application is well managed and fit for purpose.
        - Well-managed process for software configuration management.
    - Metadata profiles are fit for purpose and meet functional needs of preservation and access.
        - Appropriate standards are applied to preservation processes (e.g. architecture, metadata).

7. Infrastructure Security

- The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users *(adapted from DSA R16)*.
    - An appropriate disaster recovery plan is in place to address a major fail of services.
    - Appropriate mechanisms are in place to ensure the cyber-security of the repository, and these are independently verified.
    - An appropriate succession plan, contingency plans, and/or escrow arrangements are in place in case the repository/organization ceases to operate.

11 Appendix Three: Glossary

This report is presented on the assumption that most of its readers are familiar with many aspects of the UK and Ireland Legal Deposit Library system and the Regulations within which it operates, and therefore have a basic grasp of most of the relevant agencies and acronyms. Names are given in full during first usage, and acronyms are used thereafter. The following acronyms occur frequently and this list is presented as a simple reference to ease understanding.

11.1 Frequently used acronyms

- BL – British Library
- DAMPS – Digital Asset Management & Preservation System (a Programme to replace the DLS with a new preservation repository)
- DCMS – Department for Culture Media and Sport
- DLS – Digital Library System
- DP – Digital Preservation
- DPT – Digital Preservation Team (at the British Library)
- DPC – Digital Preservation Coalition
- DSA – Data Seal of Approval (now formally known as the Unified Requirements for Core Certification of Trustworthy Digital Repository, but specific metrics are still preceded by 'DSA' identifier)
- FTP – File Transfer Protocol
- GML – Geography Markup Language
- IPR – Intellectual Property Rights
- LDLs – Legal Deposit Libraries
- LDIG – Legal Deposit Implementation Group
- METS – Metadata Encoding and Transmission Standard
- MODS – Metadata Object Description Schema
- NLS – National Library of Scotland
- NLW – National Library of Wales
- NPLD – Non-Print Legal Deposit
- NTF – National Transfer Format
- OAC – Object Authenticity Checker
- OAIS – Open Archival Information System (reference model)
- PICASA – Preservation of Ingested Collections: Assessments, Sampling, and Action
- PREMIS – Preservation Metadata: Implementation Strategies
- WARC (Web ARChive) – file container used to store web crawls
- WA – Web Archive

11.2 BL management groups

- CMG – Collection Management Group
- RSMG – Repository Service Management Group
- DCAG – Digital Collection Assurance Group

11.3 Frequently used terms

Authenticity: the digital material is what it purports to be; all changes to a digital object can be traced and demonstrated to be authorized (see Digital Preservation Handbook).

Checksum: a unique numerical signature derived from a file; used to facilitate a **Fixity Check** (see below; see also Digital Preservation Handbook).

Completeness check: a process or series of processes to ensure that digital objects received are the same as the digital objects sent by a depositor and that no individual object is missing any required information; this information is ideally verified using an accompanying manifest (see also **Manifest**). The British Library defines 'complete' in the NPLD Completeness Checking Policy.

Crawl (Web crawl): the capture of a website or set of websites from the Web using a tool called a web crawler; a 'domain crawl' refers to the yearly capture of all UK web domain content under e-Legal Deposit (2013) Regulations; also referred to as **Harvest** or **Web Harvest** (see Digital Preservation Handbook).

Database (Database System or Database Management System): a database management system, often referred to as a database, is software (or several pieces of software working together) that allows users to access and manage data. Database systems are often used to manage data collected through research or business, such as scientific data or an inventory; the most common type of database system is relational, organizing data into tables, but non-relational databases are becoming increasingly popular, particularly for web-based platforms such as for online shopping and social media. (For an overview of databases and different types of databases, see the DPC *Technology Watch Report: Preserving Transactional Data)*.

Designated community: an identified group of potential consumers who should be able to interpret and understand a collection of preserved digital objects; these consumers may consist of multiple communities and may change over time. The term derives from ISO 14721:2012, the OAIS Reference Model (see also **OAIS**).

Digital preservation: the series of managed activities necessary to ensure continued access to digital materials for as long as necessary; including all actions required to maintain access to digital materials beyond the limits of media failure or technological and organizational change (see Digital Preservation Handbook).

Fixity check: a method for verifying that a digital object has not been altered or corrupted. It is most often accomplished by computing checksums (or hash values) (see also **Checksum**, see also **Authenticity**).

Geospatial data: also termed 'geographic information' or 'spatial data'; describe features on the earth, typically datasets such as transportation networks, property boundaries, coastlines, aerial imagery, or terrain models. (For an overview of geospatial data, see the DPC *Technology Watch Report Preserving Geospatial Data*).

GML (Geography Markup Language): **XML** based **Geospatial data** modelling language and file format

Hadoop cluster (web archive cluster): an Apache Hadoop cluster is an open-source digital storage and processing solution designed to manage large volumes of data by distributing digital information across multiple computers (referred to as distributed storage).

Harvest: see **Crawl**

Ingest: the process of preparing deposited digital objects to be put into a digital repository, including all actions required to ensure the digital objects can be accessed for as long as necessary.

Integrity check: see **Fixity check**

ISO16363: this International Standard defines a recommended practice for assessing the trustworthiness of digital repositories. ISO 16363:2012 can be used as a basis for certification (see ISO 16363:2012).

Manifest: a verifiable list of all files used to check that all digital objects received are accounted for and uncorrupted; should contain information such as file names, locations and sizes, format types and checksums (see also **Checksum**; see also **Completeness check**).

METS (Metadata Encoding and Transmission Standard): an XML schema (see **XML,** see **Schema**) for packaging digital object metadata, including descriptive, administrative, and structural metadata. The standard is maintained by the Library of Congress (see more at the Library of Congress).

MODS (Metadata Object Description Schema): an XML schema (see **XML**, see **Schema**) for packaging bibliographic data that may be used for a variety of purposes, and particularly for library applications. The standard is maintained by the Library of Congress (see more at the Library of Congress).

NTF (National Transfer Format): file format used predominantly by the Ordnance Survey for the transfer of **Geospatial data**

Normalization: a process of converting multiple file format types to one or a few file formats types; for example, converting all text document files into PDF format. This strategy helps to avoid file format proliferation and enables more effective preservation (see more about normalization in the Digital Preservation Handbook).

OAIS (Open Archival Information System): an Archive, consisting of an organization, which may be part of a larger organization, of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community (see **Designated community**). It meets a set of responsibilities, as defined in the OAIS standard, that allows an OAIS Archive to be distinguished from other uses of the term 'Archive'. The standard is a conceptual framework describing the environment, functional components, and information objects associated with a system responsible for long-term preservation. As a reference model, its primary purpose is to provide a common set of concepts and definitions that can assist discussion across sectors and professional groups and facilitate the specification of archives and digital preservation systems. Although produced under the leadership of the Consultative Committee for Space Data Systems (CCSDS), it had major input from libraries and archives. (For an overview of the standard, see the DPC *Technology Watch Report* *The Open Archival Information System (OAIS) Reference Model: Introductory Guide,* 2nd edn).

PREMIS: consists of the PREMIS Data Dictionary for Preservation Metadata, an XML schema (see **XML,** see **Schema**), and supporting documentation. The Data Dictionary is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. The standard is overseen by the PREMIS Editorial Committee and maintained by the Library of Congress (see more at the Library of Congress).

Raster: an image format that can be used to represent mapping data (see **Geospatial data**) organized on a regularly spaced, multidimensional grid of cells or lattice points, characterized by the number of dimensions (most often two, but occasionally three); the number of bands (i.e. the number of layers); and the data type of the cell values in each layer. (For more about raster formats, see the DPC *Technology Watch Report* *Preserving Geospatial Data*)

Replication: making exact copies of files to store in different locations, or different 'nodes', in order to reduce the risk of losing content; replication typically includes a means of checking that all copies remain identical and uncorrupted so that damage to files can be detected early and be repaired using a 'good' copy (see **Fixity checking**)

Representation information: information needed to facilitate the use or understanding of a preserved digital object by a particular **Designated Community**. For example, it might include a system of co-ordinates for a map, a dictionary for a set of data entries, or some software required to view an obscure file format. The term derives from ISO 14721:2012, the OAIS Reference Model (see also **OAIS**).

Seed URL: the starting point for a web crawler (see **Crawl**), usually generated by an archivist or curator. For example, a seed URL might be www.bbc.co.uk/. A web crawler would start at this URL and follow links on the page identified by the URL to capture associated pages. Crawling will be performed to a specified depth, or number of links.

Schema (XML Schema): a logical plan for structuring metadata so that it can be processed by many different systems.

Spatial databases: a database system (see **Database**), usually relational, for managing geospatial data (see **Geospatial data**) capable of storing multiple datasets along with dataset relationships, behaviours, annotations, and data models; these databases can support large volumes of data (raster and vector), complex data models, and multi-user editing and versioning. These features pose a challenge to long-term preservation because it is often not possible to extract and transfer data into other systems without losing some information. (For more about spatial databases, see the DPC *Technology Watch Report Preserving Geospatial Data*).

Structured data: a term that loosely refers to any data formatted to be easily used by machines, such as data in databases (see **Database**) or data formatted in a machine-readable language such as XML (see **XML**), as opposed to un-structured data, created by humans, such as the bodies of emails or Word documents.

Validation (also known as File Format Validation): a process used to ascertain if a particular digital file conforms to the file format it purports to be.

Vector: a geospatial data format (see **Geospatial data**) that models features on the earth's surface as points, lines, and polygons. Other information can be associated with vector data; for example, a line representing a street might have attached information like 'street name', 'number of lanes', 'speed limit', etc. Associated data may either be stored directly within the vector dataset or stored externally in a spreadsheet or database, which must also be preserved for versions of vector files to be usable in the future. (For more about vector formats, see the DPC *Technology Watch Report Preserving Geospatial Data*).

Web crawl or **Web harvest**: see **Crawl**

Web Curator Tool (WCT): an open-source software application for selective web archiving designed for use in libraries and other collecting organisations. It can be used by non-technical curators but also provides control over the web crawling process. It is integrated with the Heritrix web crawler and supports key function like permissions, quality review, and the collection of descriptive metadata. WCT was developed by the National Library of New Zealand and the British Library and is available under the terms of the Apache Public License.

WARC (Web ARChive format): a file format (.warc) used to hold web crawls (see **Crawl**); a type of file 'container', or 'wrapper', that holds multiple types of digital information and metadata in one computer file (see ISO 28500:2017).

XML (eXtensible Markup Language): a widely used standard (derived from SGML), for representing structured information, including documents, data, configuration, books, and transactions. The standard is maintained by the World Wide Web Consortium (W3C) (see more at W3C).