

BRITISH LIBRARY	Report	Version 1.2	Date 16 th October 2006
	DOM Programme		Pilot Project in Anticipation of Legal Deposit of Electronic Journals - interim report to JCLD

Pilot Project in Anticipation of Legal Deposit of Electronic Journals - interim report to JCLD

Contents

Pilot Project in Anticipation of Legal Deposit of Electronic Journals - interim report to JCLD.....	1
1 Introduction	3
1.1 Project Mission.....	3
1.2 Purpose of the Document	3
1.3 Achievements	3
2 Main Recommendation	4
3 Pilot Project interim findings on data receipt and ingest	5
3.1 Summary of the Pilot Project	5
3.2 Detailed preliminary findings	6
4 Recommendation.....	7
4.1 What the Pilot Project has shown so far	7
4.2 Recommendation	7
4.3 Issues needing to be considered by the Technical Panel	8
4.4 Other issues to be considered	10
5 Financial Impact Assessment on the BL and other LDLs	13
6 Publishers' observations on processes and financial implications	15
6.1 Data Preparation	15
6.2 Content Management.....	15
6.3 Content Submission.....	15
Appendices.....	17
Appendix 1: Publishers participating in the Pilot	17
Appendix 2: Pilot Project Interim Findings on Data Receipt and Ingest	19
Appendix 3: Publishers' inputs.....	23
Appendix 4: Functions considered in the Financial Impact Assessment on the BL.....	28

1 Introduction

1.1 Project Mission

The mission of the pilot project is to inform and influence the drafting of Legal Deposit regulations leading to the legal deposit of electronic journals (e-journals).

1.2 Purpose of the Document

This document reports the recommendations for such regulations. The recommendations are derived from the lessons learned from the Pilot Project in Anticipation of Legal Deposit between its start in June 2005 and the production of an interim report in April 2006. The technical investigations finished in June 2006 and the final report will not add significantly to them.

1.3 Achievements

The pilot project has, as far as we can tell at this interim stage, been successful in building up confidence (both in the British Library and in the publishers and others involved) that

- most of the practical issues in depositing e-Journals can be overcome, though some will require closer co-operation than was achieved
- the range of issues to be dealt with in the future is manageable.

2 Main Recommendation

It will not be possible to set up a secure, reliable and sustainable deposit system without agreement between the parties involved in legal deposit.

The main recommendation of the report is thus that a Joint Technical Panel on Legal Deposit should be set up, to review and agree on the range of technical formats and mechanisms which will be necessary or sufficient for the deposit process. The Panel should be set up in anticipation of the scaling-up of the Pilot Project, and should be fully operational before full Legal Deposit comes in to effect.

Submissions of e-Journals should adhere to a limited set of formats because:

- On the one hand, an adequate number of technical formats and mechanisms need to be supported so as not to pose an unreasonable barrier to deposit; but
- On the other hand, diversity must be limited to ensure the practical viability of legal deposit. Additional formats and mechanisms increase the cost of supporting legal deposit, inhibit interoperation, and increase the risk of loss of access to deposited material if obscure formats become inaccessible.

The Technical Panel should have participants drawn from:

- Legal Deposit Libraries (LDLs)
- Publishers
- Publisher Distribution Agents

The terms of reference and governance mechanism for the Technical Panel should be established. They should be created to facilitate the deposit process while minimising the impact on LDLs and publishers.

This Panel will need to be permanent because the technical standards for e-Journal transmission, deposit and preservation will evolve over time and will need constant

review. Its remit could be extended over time to include the other types of content approved for regulations under the legal deposit legislation.

3 Pilot Project interim findings on data receipt and ingest

3.1 Summary of the Pilot Project

The technical aspects of the Pilot Project were conducted by BL staff, whilst liaison with publishers was handled by Simon Inger of Scholarly Information Strategies Ltd. and by BL staff. The Pilot was to take a sample of 236 titles supplied by 23 volunteer publishers (listed in Appendix 1). It examined the practical problems of submission, ingest, storage and e-journal management.

The titles were held in EJOS, a commercial e-journal repository software system supplied by Endeavor (which is owned by Elsevier). A separate server was used to receive data using ftp: loaders were written and deployed to translate the raw incoming data into the database format for EJOS, and perform the loading operation. The metadata was held in the EJOS database and the content was under EJOS control.

The pilot thus tested the ability of publishers to submit data reliably as well as the BL's ability to process the data received.

3.2 Detailed preliminary findings

Findings from the Pilot Project from work done to the end of February 2006 are given in Appendix 2.

4 Recommendation

4.1 What the Pilot Project has shown so far

The pilot project has suggested a number of solutions that are applicable at the time of writing this report. However,

- for the present, these solutions will need to be validated by a wider audience, and
- for the future, the technical ground will continue to move and new solutions will have to be kept under continual review.

4.2 Recommendation

It is recommended that a 'Joint Technical Panel on Legal Deposit' should be set up, to review and agree on the range of technical formats and mechanisms which are necessary and/or sufficient for the deposit process. In each area a number of potential formats and mechanisms, designed to cover the majority of cases, will have to be supported for deposit.

All new approaches should be agreed by the Panel before implementation, via an agreed mechanism for doing so. A period of testing and validation of any new approach or standard should be undertaken before deposit commences, so that systems can be shown to be able to process submissions correctly.

4.3 Issues needing to be considered by the Technical Panel

There is a wide range of technical issues to be agreed before Legal Deposit becomes a reality. The British Library is anxious to seek to minimise the cost implications for publishers in implementing these.

4.3.1 Notification

- Supported notification mechanisms need to be determined, e.g. OAI-PMH, ONIX for Serials Serial Release Notification, RSS.

4.3.2 Acquisition methods

- Supported transport mechanisms need to be determined, e.g. ftp pull and ftp push.
- Supported compression and packaging file formats need to be determined, e.g. .tar, .zip, .gzip
- Sufficient steps to authenticate both supplier sites and items received should be determined.
- It is possible to verify whether all components of a delivery package have been received if a manifest which lists them is included with the submission. Supported manifest DTDs or schemas need to be determined.
- Supported notification mechanisms for successful acquisition and ingest need to be determined.
- Where multiple objects are wrapped in a directory structure (e.g. where individual objects representing 'articles' are wrapped in another object representing the 'issue'), supported directory structures, packaging mechanisms and filename conventions will need to be determined.

4.3.3 Content file formats

- Supported file types for the e-Journal content need to be determined, e.g. PDF, gif, XML.
- Supported content elements (such as special characters, mathematical formulas, and musical notation) should be determined.

4.3.4 Full-text DTDs or schemas

- Supported full text DTDs and schemas need to be determined, e.g. the NLM DTD.

4.3.5 Metadata DTDs or schemas

- Supported metadata header DTDs and schemas should be determined, e.g. the NLM DTD.

4.3.6 Minimum mandatory metadata

- Minimum mandatory metadata should be determined.

4.3.7 Stability criteria

- Supported policies and associated mechanisms for submitting replacement content, and for withdrawing any faulty data or perhaps illegally published data, need to be determined.

4.4 Other issues to be considered

4.4.1 Audit

- Publishers may find it useful to inspect an audit trail for the entire data load process for their content.

4.4.2 Content replacement and withdrawal

- Publishers may wish to replace or withdraw deposited content, but the LDLs may wish to retain it. Mechanisms to deal with this issue need careful consideration.

4.4.3 Authentication of submitters

- The LDLs will make every effort to ensure the authenticity of submitters, where processes and systems can provide support.
- One solution is that publishers and LDLs move to adopt the use of digital signatures in the submission process, where a digest of the object is signed. This would ensure authentication of the submitter as well as guaranteeing integrity of the content.

4.4.4 Terms of use

- Digital Rights Management tools may be required to control the use of legal deposit materials by the LDLs.
- The LDLs will need to have the right to create and keep as many copies of an item as are necessary for backup and redundancy.
- The LDLs may want to create and use copies of items in alternative formats for display purposes (e.g. a thumbnail version of a graphic).
- The LDLs may want the right to correct errors which would hinder later resource discovery (e.g. by spelling terms correctly in descriptive metadata when they are spelled incorrectly in the original item, so as not to impede indexing or retrieval).
- The LDLs may want to undertake 'Format migration' for preservation purposes from time to time.
- The LDLs will want to present items to users with their original 'look-and-feel' where possible.

4.4.5 Supply in encrypted form

- It is suggested that, initially at least, material to be deposited will not be in encrypted form or subject to software protection. However, unencrypted material may well be submitted by an understood and agreed method involving encryption and decryption, such as secure FTP.
- It is accepted that it may become necessary to submit material in encrypted form in the future, for security or other reasons. If this is the case then the form of encryption must be agreed by the Technical Panel, and the LDLs must be provided with an easy method of decryption.

4.4.6 Storage in encrypted form

- No content will be stored in encrypted form.

4.4.7 Content rejection policy

- A rejection policy for non-conforming material must be defined, along with a process to recover from the rejection of material.

5 Financial Impact Assessment on the BL and other LDLs

Commissioning the new system, managing the receipt of the data, and writing loaders for the various publisher DTDs have proved to be more time consuming than originally anticipated. The difficulty in receiving and processing publisher data was also found, as expected, not to be trivial. However, it would be unrealistic to extrapolate the cost of a live system directly from those of the Pilot Project, as the live system will benefit from the experience of the Pilot as well as from the weight of Legal Deposit law.

A brief analysis of the functions required to set up and operate a system handling Legal Deposit e-journals has been carried out. (The functions are shown in Appendix 4.)

The main assumptions made in deriving costings were:

- The system is part of the BL's Digital Object Management (DOM) system.
- Across the LDLs, ingest is carried out only once per issue
 - no allowance is made for whatever arrangements are needed to control the workflow behind this.
- Staff costs include significant generalised overheads, which cover the supply and maintenance of desktop equipment and IT infrastructure.
- Although the purchase costs of hardware needed for e-Journals storage has been costed in, other parts of the system such as storage installation and management, the provision and operation of servers supporting DOM as a whole, and whatever software and hardware is needed to provide end-use access have been omitted.
- No contribution to the general development and maintenance of the DOM system has been included.
- Preservation costs have been omitted.

The DOM system is a distributed multi-site system, with storage sites at at least 2 LDLs. It is likely that ingest and metadata creation will also be carried out at at least three sites, although some work will be centralised on the BL and partial costs could be recovered from the LDLs. The allocation of costs between different LDLs is an issue yet to be debated and is ignored in the costing.

The analysis shows

initial BL set-up costs	£371,000
annual BL operating costs	£250,000

6 Publishers' observations on processes and financial implications

Publishers were surveyed to gather information on the impact on their processes of e-Journal Legal Deposit, and any financial implications. Full details of the publisher responses are in Appendix 3.

6.1 Data Preparation

If the Legal Deposit Libraries prove able to cope with the data in the many formats that publishers will present them with, there will be little impact on publishers in this area.


If the proposed Joint Technical Panel on Legal Deposit agrees on a restricted range of technical formats and mechanisms, there will be an associated impact on some publisher costs. This is difficult to quantify at present. It is recommended that the impact on publisher costs is a consideration when the Panel discusses and agrees on any such restrictions.

6.2 Content Management

Publishers found no direct impact on costs which were specifically attributable to Legal Deposit.

6.3 Content Submission

Following a consideration of responses from publishers, an educated estimate of cost in this area (taking into consideration the wide range of publisher sizes and submission methods, and including staff and subcontractor costs) would probably be in the region of £70 per title per annum.

	Report	Version 1.2	Date 16 th October 2006
	DOM Programme	Pilot Project in Anticipation of Legal Deposit of Electronic Journals - interim report to JCLD	

Appendices

Appendix 1: Publishers participating in the Pilot

BioMedCentral
 Blackwell
 Cambridge University Press
 Elsevier
 Equinox (via Atypon)
 IFIS (via Atypon)
 Institute of Physics
 International Union of Crystallographers
 Lippincott Williams and Wilkins
 London Mathematical Society
 Maney (via Ingenta)
 Mineralogical Society (via Ingenta)
 Multilingual Matters
 Oxford University Press
 Portland Press
 Radcliffe Medical Press (via Ingenta)
 Royal College of Obstetricians & Gynaecologists (via Atypon)
 Royal Society
 Royal Society of Chemistry
 Symposium
 T&F Informa
 The Way
 Wiley

Atypon (which bought Extenza, the original participant, during the project) and Ingenta are 'aggregators', acting on behalf of several publishers.

Appendix 2: Pilot Project Interim Findings on Data Receipt and Ingest

Notification

OAI-PMH has been investigated with the titles from the Institute of Physics (selected as an example). Other methods, such as RSS, have not yet been examined.

Acquisition Methods

During the pilot, a BL server was set up and appropriate permissions were exchanged with publishers. This enabled ftp pull and ftp push to be compared.

The submission methods encountered were as follows:

FTP Pull	2
STAX (Web Site pull)	1
Web Site Pull	1
FTP Push	15
Email	1
CD	1
No Submission	2

Files were received both in native and in compressed formats (as .tar, .zip, and .gzip files). No problems were met.

At the end of January 2006, data had been received from 20 of the 23 publishers, though 1 other had supplied only 1 of 2 formats available.

File Format	Number of publishers	
	Supplied	Due
PDF and XML full text	11	
PDF, no headers	2	1
PDF, XML header	7	
PDF and SGML full text	1	1

Issues around authentication were not explored beyond the use of passwords to specified ftp servers.

Metadata File Formats

The file formats for the headers (or for metadata contained in full text files) encountered were as follows.

Header Format	Number of publishers
Effect file	1
No header	3
SGML full text	1
XML header	7
XML full text	11

Metadata DTDs or Schemas

The DTDs or schemas for the files submitted were:

DTD used	Number of publishers
Blackwell XML	1
BMC XML	1
CUP XML	1
Elsevier XML	1
IOP XML	1
IUCr XML	1
NLM DTD	7
OVID SGML	1
RSC XML	1
SSH2 SGML Catchword	3
T&F XML	1
No XML or SGML	3

To take the content supplied by publishers into EJOS, a loader had to be written or adapted for each publisher or aggregator so as to trap the metadata in a uniform way. Only the Catchword loader, which covers the submissions from Ingenta, worked 'out of the box' in a problem free manner - all other loaders required redevelopment to a greater or lesser extent.

Whilst the project did not seek to impose use of specific DTDs on its participants, it encouraged standardisation, with the primary focus on the NLM DTD.

Content File Formats

The file formats for the 'main' content files encountered were as follows.

Content Format	Number of publishers
PDF	10
PDF and SGML Full text	2
PDF and XML Full text	11

Graphic elements were included as JPEG, PNG, GIF, SML and postscript files. Specialist input included mathematical notation: musical and other symbologies had not been encountered by the end of February. Video files were included as MPEG files. No sound files have been found.

Minimum mandatory metadata

No conclusions have yet been drawn about any common metadata observed across all the submissions, or any minimum metadata required.

Stability criteria

No errata, corrections, or withdrawals have been found in the submissions examined.

Appendix 3: Publishers' inputs

[This appendix was based on information supplied by Simon Inger.]

The Pilot Project in Anticipation of Legal Deposit of E-Journals has conducted a survey of its publisher participants to ask them how they currently perceive the impact of electronic legal deposit on their processes and any financial implications thereof.

The survey was sent out to all 23 publishers, and responses were received from 18 of them, representing an excellent cross-section of the scholarly publishing industry as a whole.

Summary

Overall the publishers anticipate little difficulty in complying with E-Journal Legal Deposit as they currently understand it, although for many it is too early in the pilot project to be able to know with a deal of confidence what particular obstacles will reveal themselves. This anticipation is caveated well by this publisher comment:

"[Our] comments are based on the stated premise that the LD libraries will not impose a uniform content structure on publishers and that the metadata requirements will be in line with our current metadata structure – the more the LD libraries veer from this, the more costs we'll incur."

The findings can be grouped into three areas:

Data Preparation

This area covers the work that publishers have to undertake to prepare their content for submission. At the moment the publishers can foresee very little investment in data preparation that they would attribute to future legal deposit; changes are being made anyway as part of their ongoing improvements to data. However, several publishers commented that the pilot project had not yet got to a stage of providing detailed feedback on submitted formats and consequently it is too early to tell if there will be an impact here.

"...requirements for data preparation and delivery are not clear yet, so it's difficult to quantify the impact of this on our processes."

"Too early to tell about many of the costs since the pilot has yet to test the ingest of much of the XML full text and more importantly the multimedia elements"

Other publishers have gone down a *de facto* standards route on the assumption that this will mitigate future development:

"We deliberately adopted a developing standard archival XML DTD (NLM 2) to minimise future costs/overheads in working with e-content partners"

Clearly, if the Legal Deposit Libraries prove able to cope with the data in the multiplicity of formats that publishers will present them with, there will be little impact on publishers in this area.

If the proposed Joint Technical Panel on Legal Deposit agree on a restricted range of technical formats and mechanisms, then there will be an associated impact on some publisher costs. This is difficult to quantify at present. It is recommended that the impact on publisher costs are a consideration when the Joint Technical Panel on Legal Deposit discusses and agrees on any restrictions in the range of technical formats and mechanisms.

Content Management

This area covers the work that publishers have to undertake in managing their content post production and pre distribution. Many of the publishers surveyed outsourced this component to third parties, usually their content hosts (like Ingenta, Atypon or MetaPress). There was no perceived change required in this area, except for where publishers had developed some multimedia components which were perhaps not embedded within a standardised work-flow and hence not easily incorporated into an ultimate data feed to LDLs. There would be work necessary, although most felt that this should not be attributed solely to Legal Deposit.

Content Submission

This area covers the actual sending of data to the LDLs and is the area in which there is greatest variation.

There have been calls from some publishers for an LDL-led drive towards the development of industry standards for content submission, although others fear that such moves lead to the slippery slope of compliance with those standards and the incurrence of additional cost.

It would appear that for the larger publishers using their own content submission systems, the overhead of content submission will be low. Although they almost universally put this figure at zero, there is, of course some ongoing management and staff cost, but it may be as low as £20 per annum per title.

Mid-sized publishers use a mixture of in-house and out-of-house content submission methods and the out-of-house ones are a mixture of submissions from typesetters and online hosting partners. Publishers have put the cost of this part of compliance as between zero and £150 per title per annum. However, those offering the lowest figures caveat that with the observation that their content management partners don't currently seek to charge for this data submission, but may well do in the future once properly established into a regular work-flow.

An educated estimate of cost in this area would probably be in the region of £70 per title per annum when considering the wide range of publisher sizes and submission methods, including staff and subcontractor costs.

Appendix 4: Functions considered in the Financial Impact Assessment on the BL

Phase	Activity	Sub-activity
Set-up	Ingest system software	Licence Develop DOM interface
	Identify titles Identify publishers Discuss LD with publishers Get sample data Create Information Provider Profile (IPP)	description create metadata create non-descriptive metadata (including submission process) create / modify adaptor to process input create / modify tools for validation, navigation, rendering
Processing	Test sample data Determine validation steps Process submission &/or harvest Ingest	Handle normal ingest Handle problem ingest: claims, validation, rejection Create metadata Start internal transactions
	Store (includes redundant copies) Management Information Preservation Access (includes DRM control)	indexing resource discovery delivery rendering
Support & Maintenance	Share items with other LDLs	
	Ingest	Maintain IPP
Pervasive	Technical support of publishers and readers Technical Standards Panel S/w and h/w maintenance	
	Sample checks Audit QA plan Security Workflow Management DRM Communication with the publishers	