

# Preservation Plan for ejournals

---

## Digital Preservation Team

---

### Document history

Version	Date	Author	Status/Change
0.1	07/03/2007	Rory McLeod	Draft
0.2	24/4/2007	Paul Wheatley	Update
0.3	26/04/2007	Peter Bright	Update
0.4	26/04/07	Rory McLeod	Update
0.5	30/04/07	DPT	Complete plan
0.6	15/05/07	Rory McLeod	Update
0.7	24/05/07	Rory McLeod	Updated to reflect DOM Board suggestion.
0.8	19/6/07	Paul Wheatley	Update

### DOM responsibility- Document expiry (document valid for three years)

Owner	Status	Start Date	Document Expiry Date <sup>1</sup>	Reviewed	Reviewed by (sign)
DOM Programme Manager	Checkpoint with DPT at 12 month intervals	24/05/2007	24/05/2010	2008 2009 2010	

### DPT responsibility- Preservation plan (this plan is reviewed annually)

Owner	Format	Status	Start Date	Review Date <sup>2</sup>	Reviewed by (sign)
DPT	NLM DTD	Monitor	24/05/2007	24/05/2008	
DPT	PDF 1.6	Monitor	24/05/2007	24/05/2008	

---

<sup>1</sup> At document expiry date (36 months), DOM programme requests an updated three-year document from DPT. During the life of this document, it will be reviewed annually by the DOM team.

<sup>2</sup> At preservation review date (12 months) the preservation plan is reviewed, updated and reissued to DOM by DPT. Changes are then incorporated into the DOM three year plan.

	ejournals	Version 0.7	Date 30/04/07	
	Digital Preservation Team			

1	FOREWORD.....	3
2	INTRODUCTION .....	3
2.1	PURPOSE.....	3
2.2	DOCUMENT REVIEW .....	3
3	CONSTRAINTS .....	3
4	PRESERVATION PLAN TIMEFRAME AND OPERATIONAL HANDOVER..	4
5	ANALYSIS OF CONTENT .....	4
6	FORMAT ANALYSIS .....	4
6.1	EJOURNAL OBJECT ANALYSIS: NLM DTD .....	5
6.1.1	External information .....	5
6.1.1.1	The available NLM.DTD Tag Sets .....	5
6.1.2	Analysis of the Tag Sets of schema .....	5
6.1.3	Versions .....	5
6.1.3.1	Compatibility of the standard.....	6
6.1.4	Preservation risks .....	6
6.1.5	Options .....	7
6.1.6	Conclusions .....	7
6.2	EJOURNAL OBJECT ANALYSIS: PORTABLE DOCUMENT FORMAT (PDF) .....	8
6.2.1	External information .....	8
6.2.2	Preservation risks .....	8
6.2.3	Options .....	9
6.2.4	Conclusions .....	9
7	PRESERVATION PLAN .....	10
7.1	INGEST ACTIVITIES.....	10
7.2	CONTENT TO BE PRESERVED.....	10
7.3	FUTURE USE- ACCESS COPY .....	10
7.4	FUTURE USE- PRESERVATION COPY .....	10
7.5	PRESERVATION ACTIVITIES.....	10
7.5.1	Preservation action .....	10
7.5.2	Preservation watch .....	10
8	FUTURE PRESERVATION SUPPORT.....	10

	ejournals	Version 0.7	Date 30/04/07	
	Digital Preservation Team			

## 1 Foreword

British Library Preservation Plans are living documents that will continue to evolve over time.

Key reasons for making revisions will include:

- Changes in the content profile necessitating update or expansion
- Better content characterisation facilities enabling more detailed description and analysis of the content
- New preservation technology enabling more detailed planning for preservation actions

The role and scope of Preservation Plans will change over time, as developments in preservation metadata, particularly representation information, are progressed.

In light of these issues, a detailed and frequent review schedule has been established to ensure the published Preservation Plans remain up to date and relevant.

## 2 Introduction

This document defines the preservation plan for ejournal content that is ingested into the DOM system at The British Library (BL)

Decisions have not yet been made as to what content will be ingested from which publishers. This plan will therefore address issues likely to arise from preserving common formats including PDF and NLM XML. This will be supported by analysis of sample content. When the content that will be ingested into DOM can be made available to DTP for analysis, the DOM Programme will commission updates to this document.

### 2.1 Purpose

The purpose of this document is to record the steps that will be taken and the tools and methods used to preserve for the long-term the associated files for the ejournal project. It will:

- Describe and analyse risks to the future accessibility of the material
- Provide a framework for future preservation decisions for material of this type
- Provide a practical plan for long-term preservation of the data

### 2.2 Document review

This document, and the principles herein, will be reviewed yearly and re-assessed where necessary.

## 3 Constraints

Where project constraints are identified, they will be recorded here to produce a decision audit trail.

	ejournals	Version 0.7	Date 30/04/07	
	Digital Preservation Team			

## 4 Preservation plan timeframe and operational handover

This preservation plan is due for completion in April 2007. Operational handover will be determined by the DOM Governance Board.

## 5 Analysis of content

As noted above, only sample content was made available so this section of the plan will require update and expansion in a future revision.

Sample content was available from:

Elsevier, InformaInstitute of Physics, PublishingCambridge, University PressNature, PublishingOxford University, PressSageWiley-Blackwell, Future Science GroupRoyal Society of ChemistryINFORMS and American Medical Association

The publishers were asked to provide content in NLM DTD format. They were asked whether this was possible and whether their content would conform to NLM. All responded, a mix of full conformance, and part conformance to the DTD were received, the part conformance files appeared to have enough information so that they could be adapted or parsed by the full NLM.dtd. Although as well as this there were a number of xml files within the data that did not seem to meet either full or part conformance. DPT did not analyse this information further as the remit of this work was to assess the NLM.dtd.

Data received-

A Variety of PDF and XML schema in accordance with the NLM DTD was received. The NLM.dtd has four subsets and three of the four subsets were identified, those were Journal Publishing, Archive and Interchange and Article Authoring tag sets. The only Tag set not identified was the NCBI book Tag Set. As well as these files there were also some arbitrary data for embedded images such as .GIF and EPS.

Analysis of the data alongside conversations with BL Product Development showed that e-journal content is split into two markets. These are Science Technology and Medicine (STM) and Scholarly Journals. For this exercise, the data was dominated by STM publications. Expert opinion from our Product Development team indicates that the type of data that we received for this preservation plan is 90% representative of the data we would receive in a production environment.

Note; For the Scholarly market this figure would be reduced to closer to 20% due to the fact that for the top 25,000 titles peer reviews are split between four major platforms.

## 6 Format analysis

This section provides a description of the file formats covered by this preservation plan, along with an analysis of the risks to future accessibility and possible preservation actions that might be taken in the future.

	ejournals	Version 0.7	Date 30/04/07	
	Digital Preservation Team			

## 6.1 Ejournal object analysis: NLM DTD

### 6.1.1 External information

The National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM) created the Journal Archiving and Interchange Tag Suite with the intent of providing a common format in which publishers and archives can exchange journal content. The Suite provides a set of XML schema modules that define elements and attributes for describing the textual and graphical content of journal articles as well as some non-article material such as letters, editorials, and book and product reviews.<sup>3</sup>

#### 6.1.1.1 The available NLM.DTD Tag Sets

NCBI/NLM has created several distinct Tag Sets from the Suite of available XML Modules, each with its own purpose. A brief overview of each Tag Set is provided below. Please click on the links in the left hand column for full information.

<u><a href="#">Archiving and Interchange Tag Set</a></u>	Created to enable an archive to capture as many of the structural and semantic components of existing printed and tagged journal material as conveniently as possible, with no effort made to model any particular sequence or textual format
<u><a href="#">Journal Publishing Tag Set</a></u>	Optimized for the archives that wish to regularize and control their content, not to accept the sequence and arrangement presented to them by any particular publisher
<u><a href="#">Article Authoring Tag Set</a></u>	Designed for authoring new journal articles, where regularization and control of content is important
<u><a href="#">NCBI Book Tag Set</a></u>	Written specifically to describe volumes for the NCBI online libraries

### 6.1.2 Analysis of the Tag Sets of schema

Though each Tag Set contains substantial similarities, they are nonetheless different, and certain operations (such as future migrations) must accommodate these differences. However, in terms of the standard of documentation, market penetration, complexity, and so on, the different Tag Sets can be treated uniformly, and there is no need for this analysis to address them separately.

### 6.1.3 Versions

Versions range from 1.0 through to the current Version 2.3 published on March 28th, 2007. As the standard is actively maintained and updated it is not necessary at this point to recommend a preferred version. The data received from the publishers contained more than one version

<sup>3</sup> <http://dtd.nlm.nih.gov/>

	ejournals	Version 0.7	Date 30/04/07	
	Digital Preservation Team			

## Tag Sets

Each Tag Set is delivered as an XML DTD, W3C XML Schema, and RELAX NG, but only the XML DTD is intended for preservation.

### 6.1.3.1 Compatibility of the standard

The Suite and all Tag Sets are in the public domain. An organization that wants to create its own schema from the Suite may do so without permission. This makes it an open format which scores highly from a preservation viewpoint however the number of variations possible mean that a high degree of complexity could be introduced to the ingest process if not tightly controlled.

It is the DPT recommendation that only exact conformance to the NLM doctype be accepted in any live environment.

It is also recommended that if we encounter content that does not conform fully to the standard it purports to be that the DPT will conduct analysis on a next steps proposal on a case-by-case basis with colleagues from DOM.

### 6.1.4 Preservation risks

- Open/proprietary
  - NLM DTD is maintained and published by the National Library of Medicine. It has been placed into the public domain, and hence is freely accessible. It is not, however, subject to any formal standardization effort.
- Market penetration
  - NLM DTD is very widely used within the UK publishing industry, particularly within the Science, Technology and Medicine markets. The XML language within which the DTD is constructed has widespread industry penetration.
- Stability
  - DTD support in XML tools is widespread due to its inclusion in the XML 1.0 standard: NLM DTD is therefore expected to be stable for the foreseeable future.
  - NLM DTD is based on XML. XML is designed to be largely human legible mark-up language, whose primary purpose is to facilitate the sharing of data across different information systems. This gives us further confidence in the stability of the NLM DTD.
  - However, there is no support for newer features of XML — most importantly, namespaces. This gives us a slight concern that NLM DTD and indeed all DTDs could at some point be superseded by Document Structure Descriptions (DSDs) or XML schema definition (XSD). This will need to be monitored through DPT technology watch mechanisms.
- Exit strategy
  - NLM DTD exit strategy would be quite simple for extracting content from and representing all significant properties. The NLM DTD can be moved via transforms into any number of existing XML schemas.
- Complexity/Scale of format

	ejournals	Version 0.7	Date 30/04/07	
	Digital Preservation Team			

- NLM DTD is a reasonably simple tagged mark-up schema written for the medical publishing industry. Its human readability and XML format make this DTD both simple and accessible.
- Dependencies
  - External dependencies can include publisher specific information and variations between the four Tag Sets identified.

#### 6.1.5 Options

The strategies that could be pursued to ensure longevity of NLM DTD content are:

1. Migration to other XML (XSD), RELAX NG or DSD
  - a. This strategy would offer some insurance against if the case ever occurred where the NLM DTD was identified as at risk or obsolete. Whether migration between these identified formats could be easily and efficiently achieved would require careful consideration.
2. Migration to an alternative format
  - a. At present there does not appear to be a better-identified destination format or indeed mechanism for performing the migration. Any migrations between xml schemas would need to be tested thoroughly.
3. Alternative rendering options
  - a. NLM DTD can be rendered in any XML compatible software application and browser.
4. Monitor
  - a. There does not appear to be a pressing need to take preservation action with the NLM DTD now. Monitoring important tools, third party support, and changes to the format itself obviously remain important.

#### 6.1.6 Conclusions

The perceived threat of obsolescence to NLM DTD is considered at this time to be low. The NLM DTD has widespread industry support within both the Library and industry sector. The British Library has its own cataloguing teams who are experienced with this format and within our Product Development team we have private sector publishing experience. As well as industry support, the NLM DTD as a format is well defined and written in a mark-up language that is considered low risk for preservation and has high value as an access mechanism. Therefore, option 4 is the preferred outcome for this format at this time.

BRITISH LIBRARY	ejournals	Version 0.7	Date 30/04/07
	Digital Preservation Team		

## 6.2 Ejournal object analysis: Portable Document Format (PDF)

### 6.2.1 External information

The preservation issues surrounding the PDF format are discussed in some detail by various organisations in publicly available documents. A useful starting point is the documentation and various specifications published by Adobe<sup>4</sup>.

The AHDS Preservation Handbook on “Binary Text / Word Processor Documents” states: “...generally, the text and structural mark-up in a PDF file is not in an accessible file format and should be converted to a text-based format like XML where possible.”<sup>5</sup> It does not state what this format should be or how the variety of properties should be migrated and represented.

The FCLA require that PDFs submitted to their archive meet a number of requirements above and beyond the PDF/A standard<sup>6</sup>.

Ockerbloom discusses the options available for preserving PDFs in 2001, largely concluding that an ideal destination format does not yet exist<sup>7</sup>.

### 6.2.2 Preservation risks

- Open/proprietary
  - PDF is owned and licensed by Adobe, which publishes specifications of the format. The PDF/A format has undergone ISO standardisation. While the fact that PDF is a proprietary format causes some consternation across the preservation community, Adobe has shown interest in digital preservation requirements. Adobe has shown interest in working with the Planets Project.
- Market penetration
  - PDF is very widely used. A number of third party tools (both enthusiast developed, and commercial) provide reasonable support for rendering PDF files. JHOVE supports identification and validation of PDF files.
- Stability
  - As Goethals highlighted in the FCLA PDF Action Plan Background document<sup>8</sup>, PDF continues to grow in size and complexity.
- Exit strategy
  - There is no clear exit strategy for extracting content from PDFs and representing all significant properties.
- Complexity/Scale of format

---

<sup>4</sup> <http://www.adobe.com/devnet/pdf/>

<sup>5</sup> <http://www.ahds.ac.uk/preservation/ahds-preservation-documents.htm>

<sup>6</sup> <http://www.fcla.edu/digitalArchive/pdfs/PDFGuideline.pdf>

<sup>7</sup> <http://www.rlg.org/preserv/diginews/diginews5-1.html#feature2>

<sup>8</sup> [http://www.fcla.edu/digitalArchive/pdfs/action\\_plan\\_bgrounds/pdf\\_1\\_3.pdf](http://www.fcla.edu/digitalArchive/pdfs/action_plan_bgrounds/pdf_1_3.pdf)

	ejournals	Version 0.7	Date 30/04/07
	Digital Preservation Team		

- PDF is reasonably complex, but is described in open documentation published by Adobe.
- Dependencies
  - External dependencies can include non-embedded fonts and referenced multimedia content.

### 6.2.3 Options

The strategies that could be pursued to ensure longevity of PDF content are:

5. Migration to PDF/A
  - a. This strategy would offer some insurance against storing PDFs which reference external content (e.g. Fonts), or storing PDFs which feature interactive elements that may be difficult to preserve in the future. Whether interactive content could be discarded or not, would require careful consideration. Obviously, if content could be received as PDF/A in the first place this would be preferable!
6. Migration to an alternative format
  - a. There does not appear to be an ideal destination format or indeed mechanism for performing the migration. Ockerbloom and others have discussed extraction of text and images, but this does not appear to be an ideal solution, or one that is required now. FCLA advocates creation of XML and TIFF in addition to retaining the original PDF<sup>9</sup>
7. Alternative rendering options
  - a. Alternative rendering tools are available including the open source xpdf<sup>10</sup>. Tools like Xpdf do not typically support all PDF functionality (such as rendering interactive PDF content) but do a good job of rendering typical static content. Tools such as Xpdf therefore offer some degree of redundancy as far as reliance on Adobe rendering tools goes.
8. Monitor
  - a. There does not appear to be a pressing need to take preservation action with PDFs now. Monitoring important tools, third party support, and changes to the format itself obviously remain important.


### 6.2.4 Conclusions

The perceived threat of obsolescence to PDF is considered at this time to be low. Migration from early versions of PDF to the current version (1.6) is not necessary. Opportunities for ingest of PDF/A or possibly even migration to PDF/A should continue to be considered. The PDF format and associated dependencies should be monitored by DPT. Therefore, option 8 is the preferred option for this format at the current time.

---

<sup>9</sup> [http://www.fcla.edu/digitalArchive/pdfs/action\\_plans/pdf\\_1\\_2.pdf](http://www.fcla.edu/digitalArchive/pdfs/action_plans/pdf_1_2.pdf)

<sup>10</sup> <http://www.foolabs.com/xpdf/>

	ejournals	Version 0.7	Date 30/04/07	
	Digital Preservation Team			

## 7 Preservation plan

### 7.1 Ingest activities

Ingested files will be characterised using JHOVE, in order to identify file formats and validate those file formats where possible. The output of this process will be recorded, resulting in event information describing the characterisation and format information describing the file formats encountered. Metadata will be stored as per the DOM ejournal profile.

### 7.2 Content to be preserved

All content received will be ingested into the archival store, subject to updates within the pre-ingest period.

### 7.3 Future Use- Access copy

Decisions have yet to be made as to the access copies that will be made available.

### 7.4 Future Use- Preservation copy

Both NLM DTD XML and PDF will be considered as the primary preservation copies where available.

### 7.5 Preservation activities

#### 7.5.1 Preservation action

Both NLM DTD XML and PDF are not considered to be at risk at the current time and no action will be taken to migrate or otherwise perform preservation actions on them. Over time it is expected that risks will be identified by Preservation Watch activities (see below) and recorded in revisions of this document under the section "Format analysis". When a risk is determined to be sufficiently serious, this section will be expanded to define appropriate preservation actions that will ensure accessibility to the material in question.

#### 7.5.2 Preservation watch

The following formats will be monitored by DPT for risks to their accessibility:

- NLM DTD XML
- PDF

The monitoring of these formats will be conducted annually by Peter Bright of the Digital Preservation Team. The DPT will perform a full review of the preservation plan and an update to any sections where concerns or changes have been identified that would affect the long-term stability of the data. This may also result in preservation actions where appropriate.

## 8 Future preservation support

It is expected that the procedures and technology for supporting preservation planning and execution within DOM will be significantly enhanced over the next few years. It is expected that facilities for characterisation will be enhanced, the

	ejournals	Version 0.7	Date 30/04/07	
	Digital Preservation Team			

systems for storing information about formats, tools that render formats and the environments tools run within will become available (possibly based on PRONOM) and tools for executing preservation actions will be made available from the Planets Project. As these developments become available, this preservation plan will be updated and expanded.