

# Preservation Plan for Microsoft - Update

---

Digital Preservation Team

---

## Document history

Version	Date	Author	Status/Change
0.1	30/04/2007	Rory McLeod	Draft
0.2	04/05/07	Rory McLeod	Reviewed and approved by DPT members Paul Wheatley and Peter Bright.
0.3	24/5/07	Paul Wheatley	Minor changes following discussions between DPT and DOM Programme Manager.
0.4	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	19/06/2007	19/06/2010	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	JPEG2000	Monitor	19/06/2007	19/06/2008	
DPT	PDF 1.6	Monitor	19/06/2007	19/06/2008	
DPT	METS/ALTO	Monitor	19/06/2007	19/06/2008	

---

<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.

	Microsoft Update	Version 0.3	Date 24/05/07	
	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..	3
5	ANALYSIS OF CONTENT .....	4
6	INGEST PROCEDURE .....	4
7	FORMAT ANALYSIS .....	4
8	PRESERVATION PLAN .....	5
8.1	CONTENT TO BE PRESERVED.....	5
8.2	FUTURE USE- ACCESS COPY .....	5
8.3	FUTURE USE- PRESERVATION COPY .....	5
8.4	FUTURE USE- METADATA .....	5
8.5	PRESERVATION ACTIVITIES.....	5
8.5.1	Preservation action .....	5
8.5.2	Preservation watch .....	5
9	FUTURE PRESERVATION SUPPORT.....	6

	Microsoft Update	Version 0.3	Date 24/05/07	
	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 updated preservation plan for the Microsoft Live Book data ingested into the DOM system at The British Library (BL). It is updated here to reflect the vendor change from Internet Archive to CCS.

This document will refer repeatedly to the document MLB\_v2.doc that contains the detail of the project, and has already been signed off by the Microsoft Project Board. This document will serve only to update the sections that have altered under this supplier change.

### 2.1 Purpose

The purpose of this document is to approve the formats to be retained by the project, and identify the tools and methods used to preserve for the long-term the associated files for the Microsoft project. It will:

- Approve the formats based upon the previous work done
- 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.

## 4 Preservation plan timeframe and operational handover

This preservation plan is due for completion in April 2007. Operational aspects will be determined by the project.

	Microsoft Update	Version 0.3	Date 24/05/07
	Digital Preservation Team		

## 5 Analysis of content

As noted above, this is a revision to reflect any changes from the original documentation.

The content remains as stated in MLB\_v2.doc, namely 19<sup>th</sup> Century British Library out of copyright printed books. These books will be scanned by the new supplier CCS.

## 6 Ingest Procedure

Ingest procedures will follow that outlined in MLB\_v2.doc. Files must be checked at an object level by JHOVE for well-formedness and validity. The following files will be output by the Microsoft Digitisation Project:

- one METS file per book
- one PDF file per book
- one ALTO file per page (containing the OCR text)
- one JPEG2000 file per page

The proposal by the DOM Programme to use WARC containers to collect large numbers of related files (e.g. the JPEG2000 and ALTO files from a single book) for storage in the DOM System is endorsed as being a suitable mechanism. Whether one or more containers is used per book is an operational decision that is outside the scope of this document. It would be inappropriate for DPT to make further recommendations in this area due to the analysis operational areas will need to undertake.

SHA-1 hashes should be provided for the images and PDF to enable the detection of data loss during transit to DOM. These hashes should be recorded in the METS metadata. An additional SHA-1 hash of the METS file should be provided as a separate file. This file will consist of 40 hexadecimal digits representing the hash, two spaces, and the filename of the METS file (this is the de facto standard mechanism for recording SHA-1 hashes in standalone files).

## 7 Format analysis

The DPT have taken the view that since the budget for hard drive storage for this project has already been allocated, it would be impractical to recommend a change in the specifics as far as file format is concerned for this project. As such, we recommend retaining the formats originally agreed in MLB\_v2.doc. These are:

Linearized PDF 1.6 files for access, with the “first page” being either the table of contents, or the first page of chapter one, depending on the specifics of the book being scanned.

JPEG 2000 files compressed to 70 dB PSNR for the preservation copy.

METS/ALTO<sup>3</sup> XML for metadata. ALTO is an extension schema to METS that describes the layout and content of text pages. ALTO will be used to encode the output of the OCR process; it will describe the text and its position of the text on

---

<sup>3</sup> METS / ALTO XML Object Model, <<http://www.ccs-gmbh.com/alto>>. I.

	Microsoft Update	Version 0.3	Date 24/05/07	
	Digital Preservation Team			

the page. MODS will be embedded into the METS document to record descriptive metadata.

The use of MODS and ALTO is new; MLB\_v2.doc did not specify either. The use of MODS is consistent with current bibliographic and preservation standards. The use of ALTO provides richer resource discovery options. These changes do not change the previous decision to accept the METS file and its content as an acceptable preservation format.

## 8 Preservation plan

### 8.1 Content to be preserved

All content received will be ingested into the archival store.

### 8.2 Future Use- Access copy

PDF 1.6 files will be retained for access as per the original project specifications. Each PDF will represent an entire book, and will be linearised. The fast load page will be either the contents page or Chapter 1 (i.e. the first "content" page).

### 8.3 Future Use- Preservation copy

JPEG2000 files will be retained for preservation as per the original project specifications.

### 8.4 Future Use- Metadata

METS/ALTO files will be created to include both logical structural data (METS) and physical layout data (ALTO) as per the standards definition.

### 8.5 Preservation activities

#### 8.5.1 Preservation action

METS/ALTO, JPEG 2000, 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.

#### 8.5.2 Preservation watch

DPT will monitor the following formats annually for risks to their accessibility:

- JPEG 2000
- PDF 1.6
- METS
- Alto

No immediate risks are identified with the formats used within this project.

-The PDF is for access purposes but is a well-defined and widely used standard.

	Microsoft Update	Version 0.3	Date 24/05/07	
	Digital Preservation Team			

-The JP2 files fulfil the role of master file but a lack of industry take-up is a slight concern from a preservation viewpoint. However, the format is well defined and documented and poses no immediate risk.

-The JP2 format has yet to be added to the BL technical standards document however, a summer workshop of industry experts has been organised at the BL in London to discuss this matter. Any relevant findings will be added to this document at this time.

-Both METS and Alto are existing metadata schemas that are approved by DPT as preservation standards.

Peter Bright of the Digital Preservation Team will conduct the monitoring of these formats annually. 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.

## 9 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 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.