



ISO/IEC JTC 1/SC 34/JWG 7 "Joint JTC 1/SC 34-TC 46/SC 4/WG: EPUB"  
Convenorship: KATS  
Convenors: Cho Yong-Sang Dr, Oh Sam Gyun Professor



## [Contribution for PWI] Looking forward to working with you to produce the EPUB/A specification

Document type	Related content	Document date	Expected action
General / Other	Meeting: <a href="#">VIRTUAL 29 Mar 2023</a>	2023-03-28	

### Description

This document is prepared by Alicia Wise who is a leader of preparation PWI regarding EPUB/A. This document will be discussed in JWG7 online meeting on March 29, 2023.

# Preserving EPUB3 files – current position and next steps

23 March 2023

# About EPUB3

- EPUB3 provides a means of representing, packaging and encoding structured and semantically enhanced Web content — including HTML, CSS, SVG and other resources — for distribution in a single-file container.
- EPUB is generally regarded as a suitable format for digital archiving because it is open/transparent, a standard, generally forward/backward compatible, there is a degree of protection against file corruption, the frequency of version releases, and it is generally interoperable.

Two helpful documents prepared by Juha Hakala  
under the auspices of the Joint Technical Committee ISO/IEC JTC 1, Information technology,  
Subcommittee SC 34, Document description and processing languages

TECHNICAL  
SPECIFICATION

**ISO/IEC TS  
22424-1**

First edition  
2020-01

---

---

**Digital publishing — EPUB3  
preservation —**

Part 1:  
**Principles**

*Publications numériques — EPUB3 preservation —  
Partie 1: Principes*

TECHNICAL  
SPECIFICATION

**ISO/IEC TS  
22424-2**

First edition  
2020-01

---

---

**Digital publishing — EPUB3  
preservation —**

Part 2:  
**Metadata requirements**

# The two documents:

- The principles document (ISO/IEC TS 22424-1):
  - Is more publisher-facing
  - Introduces concepts in digital preservation, definitions for frequently used terms, and provide a map to the many standards we may need to consider.
  - Describes which EPUB features are mandatory for long-term preservation (such as font and image embedding) and features which should be avoided if possible.
  - Covers the initial stage of the archiving process, namely the creation of submission information package (SIP) which preservationists use to generate archival information packages (AIPs)
  - Dates to 2020 and covers EPUB versions up to 3.0.1
- The metadata requirements (ISO/IEC TS 22424-2):
  - Is more archivist-facing
  - Specify metadata elements which are required or recommended for long-term preservation and the ways in which the EPUB publication and related metadata can be packaged.
  - Focus on metadata needed to support migration rather than emulation preservation approaches.
  - Provide examples of how metadata elements should be expressed using either 1) METS and PREMIS Data Dictionary for Preservation Metadata and/or 2) EPUB version 3.0 and 3.0.1. The two are important because METS is more widely used by archives, and publishers are more likely to submit EPUBs with embedded metadata plus an ONIX metadata feed.
  - Do not cover detailed metadata requirements to support the preservation of embedded media types.

# Key EPUB preservation challenges

- As a general principle, things that make EPUB3 files more accessible also make them more usable and more preservable.
- Reflowable EPUB publications are designed so that their look and feel can change with no impact on semantics, which is better for digital preservation. Fixed layout EPUB publications can be better for presentation, but any change in the appearance of the document may cause significant changes in the meaning or even lose it completely.
- Fonts, audio, and video should be embedded in the EPUB file
  - Image and audio files embedded in an EPUB publication may require migration before the EPUB publication itself and should adhere to relevant standards for their type.
  - The EPUB community can change the list of EPUB Core Media Types any time, independent of the EPUB specification updates.
  - Core media types that are considered to be non-archivable should be avoided whenever possible. For instance, it is better to use a JPEG or a PNG than a GIF image.
- Links to external resources should be avoided or else those external resources also need to be captured, described, and preserved in addition to the EPUB.
- To be preserved, publications should contain features commonly supported in software packages used to render the content.
- Bespoke functionality can lead to preservation challenges:
  - There are various ways EPUB supports interactivity, which is a key challenge. Interactive EPUB documents (i.e. those containing software components) are more challenging to preserve. It is less likely that they can be migrated to newer formats, and more likely to require software emulation in order to preserve access
  - Publisher/ supplier specific DRM often restricts the use of e-books to that publisher's/supplier's rendering devices and/or applications and is an obstacle to preservation. DRM protection should be removed from EPUB publications during pre-ingest by an archive. In practice, only archives cooperating with the publishers or national libraries may be able to do this, provided that legal deposit act and / or copyright act guarantee them such privilege.
  - Fragment identifiers are problematic since URI fragments are media type dependent. These should be avoided.
  - Encryption, obfuscation, and foreign resources are also problematic.
- Descriptive and preservation metadata are required, including information about any compression methods or any reading systems for which the file is optimized

# Metadata requirements

- Package creator / submitter information
- Package status
- Package identifier
- Work and publication identifiers
- Core media type resource identifiers
- Foreign resource identifiers
- Identifiers for metadata records (e.g. Dublin Core or PREMIS)
- Relevant dates (e.g. publishing date, SIP creation date, EPUB or SIP modification dates)
- Metadata format and its versions
- Administrative metadata (e.g. nature and formats of embedded media)
- Technical metadata (e.g. file formats and versions, digital signatures and checksums)
- Rights metadata
- Structural metadata
- Preservation metadata (e.g. preservation steps that have been taken in the life of the file)
- Structure of submission information packages
- Content of submission information packages
- Digital signatures (which are not mandatory but there is guidance should they be created)

# Our Task

- Create EPUB/A, a subset of EPUB 3 with features suitable for long-term digital preservation.
- The specification should be complemented by an explanation of why the EPUB 3 features not included in the EPUB/A format may jeopardize digital preservation, and a justification for those features that are required.
- Once EPUB/A has been published, there is a need for tools for creating, validating and rendering EPUB/A files, **but it is not our task to develop them.** Since EPUB/A will be a subset of EPUB, developing these tools should not be a difficult technical task.

# Stakeholders

- Academic libraries preserving EPUB3 files
- National libraries and archives with legislation to require / underpin their preservation activities (e.g. National Library of Finland).
- Preservation services ingesting EPUB3 files for long-term preservation (e.g. CLOCKSS)
- Publishers of many types:
  - Academic (commercial, library, society, university presses)
  - Educational
  - Trade
- Aggregators (e.g. EBSCO)
- Distributors (e.g. Ingrams)
- E-Reader vendors (e.g. Amazon, Kobo)
- Others?

Looking forward to working with  
you to produce the EPUB/A  
specification

Alicia Wise

[awise@clockss.org](mailto:awise@clockss.org)

@wisealic