Rationale for the development of an EPUB/A standard

# Terms of Reference for the project

Building on previous work[[1]](#footnote-0):

* Create EPUB/A, a subset of EPUB 3 with features suitable for long-term digital preservation.
* Agree which EPUB features are required for long-term preservation and features which should be avoided if possible.
* Specify metadata elements which are required or recommended for long-term migration and 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.
  + Enable stakeholders who have adopted EPUB understand key concepts in digital preservation (e.g. why it is important and how, at a high level, it is done)
  + Enable stakeholders responsible for digital preservation to understand why EPUB is important, and by whom and how it is deployed

NB: 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.

# Candidate scope (list or statements)

As a general principle, things that make EPUB3 files more accessible also make them more preservable.

To be preserved, publications should contain features commonly supported in software packages used to render the content.

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 during migration for preservation may cause significant changes in the meaning or even lose it completely. There are things that can be done (e.g. good use of semantics, HTML, and CSS) to make fixed layout EPUB publications more preservable.

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 grow 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.
  + If there are a large number of resources, embedding them all may make an EPUB file too unwieldy. In these cases the resources should be packaged alongside the EPUB file.

External resources integral to the work should not be embedded and instead need to be independently captured, described, and preserved in addition to the EPUB. Links to external resources should use unique, permanent identifiers.

Bespoke functionality can lead to preservation challenges as well as security and privacy issues.

* Scripted content should be avoided. Interactive EPUB documents (i.e. those containing software components) are 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.
* Fragment identifiers are problematic since URI fragments are media type dependent. These should be avoided.

DRM protection must be removed from EPUB publications by the publisher prior to deposit into an archive:

* LCP, an open protocol for DRM has not be as widely adopted as had been hoped. There therefore isn’t in practice one standard open DRM for EPUBs.
* Even if there were a widely available open DRM these would still contain some proprietary code.
* Publishers deliver DRM-free content to platforms and preservation services, and are happy to do this for trusted archives.
* Retailers, rather than publishers, apply DRM before delivering copies to readers. This can be done for security reasons and also to tie content to specific retail/reader platforms to support business models.
* Retailers are not authorized to provide content for preservation. Publishers do so.
* Publishers should not be required to make the additional step of adding DRM to EPUBs prior to preservation. This is additional work and would also make preservation much harder.
* Social DRM (e.g. watermarks) should also be removed prior to preservation.
* Encryption and obfuscation of EPUBs are also problematic from a preservation perspective.
* DRM is sometimes added right before acces by an individual, so this might be relevant for eventual release of the publication by an archive but this use case is out of scope for EPUB/A.
* In practice, only archives cooperating with the publishers or national libraries areable to remove DRM protected content deposited for preservation. This requires agreements with rightsholders or legislation such as legal deposit and / or copyright exceptions.

Descriptive and management metadata are required, including information about any compression methods or any reading systems for which the file is optimized and:

* Title
* Creator
* Creation Date or Start Date/End Date
* Publisher/producer/distributor
* ISBN (or other standard unique identifier)
* DOI
* Core media type resource identifiers
* Metadata format and its versions
* Identifiers for metadata records (e.g. Dublin Core or PREMIS)
* 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

# Purpose and justification

EPUB 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.

EPUB is widely deployed in book publishing and if the archival version of this is standardized it has the potential to become widely used to preserve human knowledge.

# Timeline with deliverables

* ISO\_IEC\_TS\_22424-1 and ISO\_IEC\_TS\_22424-2 were delivered in 2020
* Informal discussion has taken place throughout 2023
* Request for approval of New Work Item – September 2023
* Registration of the Approved Work Item (AWI) – October 2023
* Call for requirements and candidate issues – November 2023
* First meeting of working group – January 2024
* Second meeting of working group – February 2024
* Third meeting of working group – March 2024
* Fourth and final meeting of working group – April 2024
* Deliver working draft to share more widely – May 2024
* Committee stage = Committee draft (CD)
* Enquiry stage = Enquiry draft (ISO/DIS)
* Approval stage = Final Draft International Standard (ISO/FDIS)
* Publication stage = International Standard (ISO)

# Candidate project leader and participants

JWG 7 committee members led, for this work item, by Dr. Alicia Wise.

# Stakeholders

We have identified stakeholders including:

* Academic libraries
* Advocates for ebook accessibility
* Book production and supply chain partners
* E-Reader vendors
* National libraries and archives
* Preservation services
* Publishers
* Relevant software vendors

1. Digital publishing — EPUB3 preservation — Part 1: Principles (ISO\_IEC\_TS\_22424-1\_2020\_ed.1) and Digital publishing — EPUB3 preservation — Part 2: Metadata requirements (ISO\_IEC\_TS\_22424-2\_2020\_ed.1) [↑](#footnote-ref-0)