ISO/IEC JTC 1/SC 34/JWG 7

Joint JTC 1/SC 34-TC 46/SC 4-IEC/TC 100/TA 10 WG: EPUB

Convenorship: KATS (Korea, Republic of)

Document type:	Final Text submitted for TR publication
Title:	Call for final review for publication of ISO/IEC TS 22424-1 Digital publishing — EPUB 3 Preservation — Part 1: Principles
Status:	This document is final text for publication of ISO/IEC TS 22424-1 Digital publishing — EPUB 3 Preservation — Part 1: Principles reflecting comments from PDTS ballots. Through four-week review in SC 34 and SC 34/JWG 7 this document will be submitted to SC 34 for publication. If you have any further comments or suggestion on this document, please let project editor (Juha Hakala, juha.hakala@helsinki.fi) by August 23, 2019.
Date of document:	2019-07-17
Source:	Project editor (Juha Hakala)
Expected action:	INFO
No. of pages:	38
Email of convenor:	samoh@g.skku.edu, zzosang@gmail.com
Committee URL:	https://isotc.iso.org/livelink/livelink/open/jtc1sc34jwg7

ISO/IEC TS 22424-1:2019

ISO/IEC JTC 1/SC 34/JWG 7

Secretariat: JISC

Digital publishing — EPUB 3 Preservation — Part 1: Principles

Édition numérique — Archivage pérenne de l'EPUB 3 - Partie I : Principes

TS stage

Warning for WDs and CDs

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
copyright@iso.org
www.iso.org

Contents

Foreword iv

Introduction v

1 Scope 1

2 Normative references 1

3 Terms and definitions 1

4 Abbreviated terms 11

5 Packaging standards 11

6 Construction of OAIS information packages 13

6.1 General 14

6.2 Identification of information packages and their content 19

6.3 Structure of information packages 21

6.4 Generic Information package metadata 21

Annex A (informative) EPUB and digital preservation: issues and recommendations 23

Bibliography 26

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 34, *Document description and processing languages*.

A list of all parts in the ISO/IEC TS 22424 series can be found on the ISO website.

Introduction

This document facilitates the long-term preservation of EPUB publications by specifying in general level EPUB features which are mandatory for long-term preservation (such as font embedding) and features which should be avoided if possible.

This specification can be seen as a stepping stone towards a detailed specification which would be related to EPUB in the same way as PDF/A, specified in ISO 19005-1 – 19005-3, is related to PDF. If and when the EPUB community develops detailed guidelines for the production of archivable EPUB publications, this document could be used as one of the starting points.

Long-term preservation in general requires two things:

- making the object such as EPUB publication fit for preservation – including features to be used and features to avoid;
- the packaging of the object (and any metadata related to it) together with any additional data such as other versions of the object and other documentation into an OAIS Submission Information Package (SIP).

Packaging is covered in Part 2 of this technical specification.

EPUB

The EPUB standard

> *defines a distribution and interchange format for digital publications and documents. The EPUB® format 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 3.0.1].*

EPUB format was developed by the International Digital Publishing Forum, IDPF, which merged with the World Wide Web Consortium, W3C, in January 2017. Ongoing technical development of the standard, related extension specifications and ancillary deliverables are the responsibility of the W3C EPUB 3 Community Group[1], which published its charter in February 2017. According to the charter,

> *work on any future major revision of EPUB, e.g. an EPUB 4, is initially out of scope on the presumption that this will be taken up by a new W3C WG as a W3C Recommendation Track activity. The EPUB 3 CG will coordinate its work with such new WG, and meanwhile with the existing W3C Digital Publishing Interest Group (DPUB IG). [W3C]*

The International Digital Publishing Forum, IDPF, has ceased operations as a membership organization in January 2017, and its website[2] is now an archive. The latest version of the standard and information about future EPUB developments is available at the Publishing@W3C webpage, https://www.w3.org/publishing/.

The specification at hand covers EPUB 3 versions up to EPUB 3.0.1[3]. EPUB 3.1[4] was the first major revision of EPUB 3.0.1, but there are no implementations of version 3.1 and therefore it is not covered in this document. The most widely used version of the standard is still 3.0.1. EPUB 3.2, was published in

[1] https://www.w3.org/publishing/groups/epub3-cg/
[2] http://idpf.org/
[3] http://idpf.org/epub/301
[4] https://www.w3.org/Submission/epub31/

May 2019[5]. Unlike 3.1, it is fully backwards compatible with 3.0.1. It will be covered in the next edition of this document.

Differences between EPUB specifications 2.0.1-3.2 are well documented:

- EPUB 3 Changes from EPUB 2.0.1[6]
- EPUB 3.0.1 Changes from EPUB 3.0[7]
- EPUB 3.2 Changes from EPUB 3.0.1[8]

All EPUB specifications are available in the Web; 2.01 at http://idpf.org/epub/201, EPUB 3.0.1 at http://idpf.org/epub/301 and 3.2 at https://w3c.github.io/publ-epub-revision/epub32/spec/epub-spec.html.

All EPUB publications, including ones using version 3.2, can be validated using EPUBCheck version 4.2.0, which was released in March 2019.

From long-term preservation point of view, lack of backward compatibility between successive versions of a file format would be a problem because it makes migration more challenging. In addition, EPUB 3.1 has at least one feature which would have been problematic. In EPUB 3.1 foreign resources do not require fallbacks if they are not in the spine and not embedded in EPUB Content Documents. In EPUB 3.0.1, fallback guarantees that there is a version of the document that can be rendered; in 3.1 such guarantee no longer exists.

EPUB 3.0.1 was prepared by the IDPF. It consists of six interlinked documents:

- EPUB 3 Overview
- Publications 3.0.1
- Canonical fragment identifiers
- Content documents 3.0.1
- Media overlays 3.0.1
- Open Container Format 3.0.1

There are several extension specifications to these EPUB base standards. The list below is incomplete, as it contains mainly specifications that are relevant from the long-term preservation point of view. Some of them are still drafts:

- EPUB Accessibility specification 1.0[9] addresses evaluation and certification of accessible EPUB Publications, and discovery of the accessible qualities in such publications.
- EPUB Previews 1.0[10] describes how content previews can be included in EPUB publications.
- EPUB Distributable Objects 1.0[11] is a draft specification that defines a method for the encapsulation, transportation, and integration of distributable objects in EPUB publications.
- EPUB Scriptable Components 1.0[12] provides an interoperable publish and subscribe (pubsub) pattern by which interactive content can be created and incorporated into EPUB publications. Same as EPUB Distributable Objects, it is as of this writing (2019-05-13) a draft.
- EPUB Scriptable Components Packaging and Integration 1.0[13] is a draft that defines a method for the creation and inclusion of dynamic and interactive components in EPUB publications.

[5] https://w3c.github.io/publ-epub-revision/epub32/spec/epub-spec.html
[6] http://www.idpf.org/epub/30/spec/epub30-changes-20111011.html
[7] http://www.idpf.org/epub/301/spec/epub-changes-20140626.html
[8] https://w3c.github.io/publ-epub-revision/epub32/spec/epub-changes.html
[9] http://www.idpf.org/epub/a11y/accessibility.html
10 http://www.idpf.org/epub/previews/epub-previews-20150826.html
11 http://www.idpf.org/epub/do/
12 http://www.idpf.org/epub/sc/api/

- EPUB Multiple-Rendition Publications 1.0[14] defines the creation and rendering of EPUB publications consisting of more than one rendition of the same publication.
- EPUB Dictionaries and Glossaries 1.0[15] provides a means for expressing dictionary and glossary semantics in EPUB publications.

These extensions are not widely used and they have not been explicitly taken into account in this document. As regards accessibility, all EPUB publications are supposed to be accessible. However, accessibility features as such do not have an impact on long term preservation of EPUB publications and therefore this document does not make accessibility-related requirements.

EPUB 3 Core Media Types have been listed at https://idpf.github.io/epub-cmt/v3/. As of this writing [2019-05-13], the latest change has been made on April 1, 2018. Starting from EPUB 3.2, core media types are part of the standard.

In 2014, EPUB 3.0 specifications were republished as a standard, ISO/IEC TS 30135 parts 1-6, by the International Standards Organization. Each of these six ISO specifications is identical to its IDPF equivalent, for example TS-30135-1 has exactly the same content as the EPUB 3.0 Overview.

ISO/IEC TS 30135-7 is "Part 7: EPUB3 Fixed-Layout Documents" is from EPUB 3.0.1 (EPUB 3.0 does not have fixed layout specification). TS 30135 is therefore a combination of EPUB 3.0 and Fixed-Layout Documents specification from 3.0.1.

ISO/IEC JTC 1/SC 34 is currently updating the ISO standard to match fully the version 3.0.1.

EPUB is a rich document format with a lot of features. From the digital preservation point of view this is a challenge, not least because long-term preservation has not been a priority in the development of the standard. Preserving all aspects and features of EPUB publications may be difficult, since there are features which are difficult to preserve. Moreover, EPUB reading systems usually do not support all features of the specification and finding tools supporting rare features can be difficult.

In spite of these challenges EPUB is generally regarded as a suitable format for digital archiving. For instance, the Finnish National Digital Library initiative has selected just eight archivable file formats for text, EPUB being one of them. The selection criteria were openness/transparency, adoption as a preservation standard, degree of forward/backward compatibility, degree of protection against file corruption, frequency of version releases, dependencies/interoperability, and standardization. EPUB got an A, the best grade, from everything else except the second and third criterion. For those, the grade was the second best, a B [File formats, p. 40]. Based on these generic criteria, EPUB seems to provide a good basis for long-term preservation, although additional guidelines on how to use the standard are needed to guarantee EPUB files can be preserved efficiently.

The British Library's Digital Preservation Team has published an assessment of EPUB as a preservation format [Day]. It covers EPUB versions 3.0.1 and 2 and the overall view of EPUB is positive [Day, p. 2]:

> *EPUB 3 is currently the closest thing available to an open standard for e-books. In 2013, Bläsi and Rothlauf concluded that EPUB 3 had the "highest expressive power" of all formats in the e-book ecosystem, and that it included the superset of all features used in proprietary formats like KF8, Fixed Layout EPUB, and iBooks.*

EPUB is enjoying reasonable support in the e-book market. Many suppliers, publishers, and application developers who have supported EPUB 2 have implemented version 3.0.1. According to the EPUBTest

[13] http://www.idpf.org/epub/sc/pkg/
[14] http://www.idpf.org/epub/renditions/multiple/
[15] http://www.idpf.org/epub/dict/

web site[16], EPUB 3 support in reading systems is far from exhaustive, but market coverage is good – in January 2018, there were 59 reading systems supporting at least some of the features specified in EPUB 3.0.

E-book suppliers have produced EPUB 3 based formats that incorporate Digital Rights Management (DRM), and EPUB modifications that may restrict using the format on other than the suppliers' own platforms. For example, the Kindle Fire eReader, released in 2015, uses a new format called Kindle Format 8 (KF8), which is partly based on EPUB 3, with Amazon's DRM. [Day, 3]. Publisher/supplier specific DRM often restricts the use of e-books to that publisher's/supplier's rendering devices and/or applications, and is therefore a major obstacle to digital preservation [Day, p. 7].

The EPUB specification does not enforce a particular Digital Rights Management scheme, but DRM may be layered on top of the EPUB specifications. A producer can, for instance, use one of the three major rights management systems in the market (Amazon DRM, Apple FairPlay DRM for books bought from iBooks, and Adobe DRM), or some other DRM system along with some additional platform-targeting.

DRM protection should be removed from EPUB publications during pre-ingest by the producer or as a part of the ingest process by the OAIS archive. In practice, only national libraries may be able to do this, provided that legal deposit act and / or copyright act guarantee them such privilege. If migration is the chosen preservation strategy, existing EPUB publications will be converted into more modern EPUB versions when rendering tools for old versions are no longer available, and (eventually) migrated into other formats.

If preserved EPUB publications are not directly accessible by the public, removing DRM, digital watermarking, and other protection mechanisms from the archived documents is not a risk. When publications are delivered to the customers as Dissemination Information Packages (DIPs), the archive shall use a combination of administrative and technical means to protect the documents as required in the submission agreement. These means may include adding DRM protection mechanism into the DIP submitted to the user according to the requirements of the submission agreement. The agreement may also specify the customers the archive is entitled to serve; for instance, it is possible to require that the preserved documents can only be disseminated to the producer, and the producer will serve the end-users who do not have direct access the OAIS archive.

Digital preservation

The information society is dependent on successful long-term digital preservation. When an increasing percentage of information is produced and published only in a digital format, it is important to make sure that this information remains available in the distant future.

Digital preservation is not about preserving just bits, but about preserving access. The "business logic" is as follows:

- we need software and hardware to render content for human users
- software changes over time; there are new versions from old applications, and entirely new applications
- new or updated applications may not be able to render outdated file formats or format versions correctly
- digital preservation makes an effort to have all archived content in stable formats. Publications should also contain the smallest possible amount of features which are not commonly supported in software packages used to render the content in these formats, and also avoid

[16] http://epubtest.org/testsuite/epub3/

adding links to external resources since then the long-term access to the publication requires also persistence of these external resources.

- when necessary, data in old formats may be migrated into more modern formats or updated versions of the same format. For instance, an e-book in EPUB 3.0.1 format may be migrated to EPUB 5.2. when version 3.0.1 is no longer widely supported by reading systems.

- since the aim is to preserve the content, not the bits, the bits may change as a result of version updates and format migrations.

- Many OAIS archives preserve successive versions of archives publications, because migration may change the look and feel of the original document, or even its intellectual content.

In many countries, national libraries are responsible for preserving the published cultural heritage for the future generations, while national archives take care of governmental publications, irrespective of which format they are available in. All of these resources have to be preserved for decades, centuries even. Then again, publishers may guarantee continuous access to the subscribers of electronic serials and other licensed content. If this is so, either the publisher or a third-party should look after the publications and make sure they remain accessible or at least available.

Ordinary digital asset management systems are not suitable for long-term preservation; therefore it is a normal practice to separate short-term and long-term information management into different systems. However, this does not mean that digital archiving is independent of the routine life cycle of documents. Digital preservation is a long process that begins when publications are created.

Preservation metadata, which allows the publication to be found, rendered and authenticated correctly, is a prerequisite for digital preservation. Some preservation metadata elements can or should be provided by the original creator of the publication. It is also important to keep preservation requirements in mind when preparing a publication, if it is known that it has to be preserved for a long time. Any feature in a file format can be either essential, useful, neutral, questionable, or even downright counterproductive from a long-term preservation point of view. However, publishers are likely to use the features that let them achieve their own goals, and preservation may not be among them.

There already are archivable versions of some file formats. PDF/A (ISO 19005-1:2005)[17] is probably the best known example. It specifies how to use the Adobe Portable Document Format (PDF) for long-term preservation. An example of a counterproductive feature for preservation in PDF is font referencing; therefore in PDF/A all fonts shall be embedded in order to guarantee that the document can be rendered correctly.

PDF/A forbids also the use of encryption, because encryption is generally regarded as a risk for long-term preservation. But storing unencrypted documents is a risk as well, because if they are stolen, non-authorized usage is easy. Therefore, according to the Digital preservation handbook [Digital]:

> *Information security methods such as encryption add to the complexity of the preservation process and should be avoided if possible for archival copies. Other security approaches may therefore need to be more rigorously applied for sensitive unencrypted files; these might include restricting access to locked-down terminals in controlled locations (secure rooms), or strong user authentication requirements for remote access.*

In order to guarantee the correct processing of PDF/A files, there are specific requirements for PDF/A reading systems, such as support for embedded fonts. There are three versions of the specification: PDF/A-1 is based on PDF 1.4, PDF/A-2 adds features from PDF 1.5, 1.6 and 1.7, and PDF/A-3 contains

[17] https://www.iso.org/standard/38920.html

all the features of PDF/A-2 as well as allows the embedding of other file formats into PDF/A conforming documents [PDF/A].

The TI/A (Tagged Image for Archival) standard initiative intends to create an ISO recommendation to optimize the format specification for archival purposes. The motivation behind the initiative applies perfectly to other image formats, but there are valid points to the EPUB community as well [TI/A]:

> *The versatility of the TIFF format has made it very attractive for memory institutions for long-term archival of their digital images. However, since the TIFF format offers such a great flexibility, it is not guaranteed that in the future a standard TIFF reader will be able to read some TIFF images.*
>
> *The limitations of the baseline TIFF are too severe for many applications in digital archiving. It is important that, besides crucial technical metadata such as ICC color profiles (in case of color images) also important descriptive metadata is stored within the image file. Having descriptive metadata available (such as content description, iconography, copyright and ownership information etc.) is crucial for every archive. Having this information in the same file as the image data guarantees that this information will always be associated with the image.*

TIFF is not an EPUB core media type, but four other image types have been listed; GIF, JPEG, PNG, and SVG. It is significant from a digital preservation point of view how these formats and other core media types are used in the EPUB context. Image and audio files embedded in an EPUB publication may require migration before the EPUB publication itself has to be migrated into a more modern file format, if commonly available EPUB reading systems no longer support these file formats. This specification does not provide guidelines for creating archivable files in EPUB 3 core media types, due to the magnitude of such task. But EPUB community SHOULD follow the archival file format lists of national archives or libraries (for example the Library of Congress file format list[18] and the U.S. National Archives list[19]) when the core media file format list is updated. Publishers SHOULD also consider the persistence of file formats used when creating EPUBs for which the need for long-term preservation is foreseen. .

This specification does not require any changes to be made to the EPUB standard or to any future versions of it. However, with each new EPUB standard version it is necessary to check if the ISO 22424 needs to be revised, since any new EPUB features may be either useful, counterproductive, or irrelevant from a long-term digital preservation point of view. A similar approach is already in place for PDF/A: ISO 19005-1 applies to PDF 1.4, and ISO 19005-2 covers the subsequent PDF versions up to 1.7.

OAIS and related standards

ISO 22424 provides guidance on how to utilize the Open Archival Information System (OAIS) and current practices of OAIS archives in preservation of EPUB publications. The OAIS [ISO 14721] is equally relevant to both parts of the ISO 22424.

OAIS is a reference model for long-term data storage systems. It is used by memory institutions (libraries, archives, and museums) and many other organizations that need to preserve digital resources in the long-term. Although an ISO standard, the OAIS was originally developed by the CCSDS, The Consultative Committee for Space Data Systems[20], which still maintains the specification.

The model has five functional units:

[18] http://www.loc.gov/preservation/digital/formats/

[19] https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html

[20] https://public.ccsds.org/default.aspx

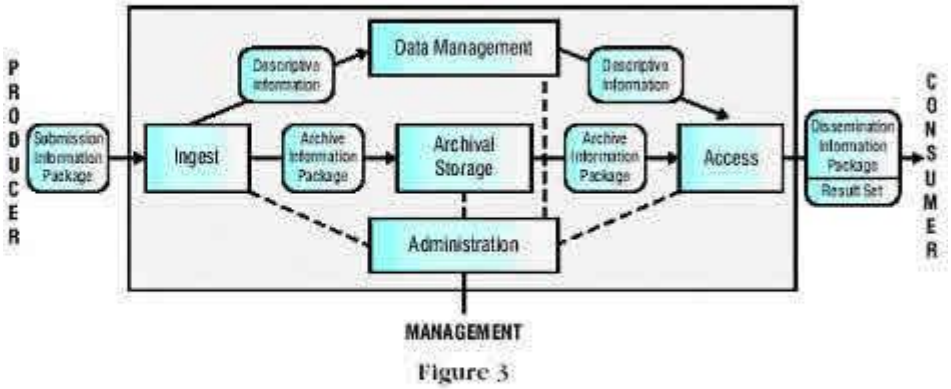

Figure 1. OAIS Model [Lavoie]

In the model, the *Ingest function* is responsible for receiving information from producers and preparing it for storage and management within the OAIS archive. The Ingest accepts information – in this case, EPUB publications – from producers in the form of Submission Information Packages (SIPs), performs quality assurance checks on the SIP, and generates an Archival Information Package (AIP) from one or more SIPs (or multiple AIPs from a single SIP). Finally, the Ingest function transfers the new AIPs to Archival Storage and the associated Descriptive Information (metadata) to Data Management.

Modifying an EPUB publication so that it is suitable for digital archiving is from the OAIS point of view a part of pre-ingest and as such not a part of the OAIS model. The importance of the OAIS to the ISO 22424 is that the model provides a terminology, information package data model and an overall framework within which digital preservation can be performed.

Neither OAIS nor this specification describe the interface between a repository system used by the archive and systems used by producers. The Producer-Archive Interface Methodology Abstract Standard, also known as PAIMAS [ISO 20652], covers the first stages of the ingest process defined by the OAIS. It provides a basis for detailed specifications on how production systems communicate with OAIS archives. One such specification is DEPIP, the Data Exchange Protocol for Interoperability and Preservation [ISO/FDIS 20614]. The DEPIP is intended for systems used by libraries, archives, and museums. Other domains are likely to create their own API specifications.

Of all the functional units of the OAIS model, this specification covers only the Ingest unit. In addition there are tasks that are part of non-OAIS unit Pre-ingest, or things a producer shall take care of when preparing a SIP. Other OAIS units are beyond the scope, and therefore archival or dissemination related functions such as migration or creation of dissemination information packages are discussed only in passing. It is assumed that Ingest does not require any major changes, although if EPUB for some reason were no longer approved as preservation format, the archive would be obliged to migrate the EPUB publications into eligible file format. Even then the submission agreement might require the archive to disseminate the publication back to consumers in the original EPUB format.

OAIS submission agreements specify the principles of how documents should be prepared and submitted to the repository system. If the archive uses migration as the preservation method[21], submission agreements should specify file formats (and metadata formats) suitable for submission

[21] In this document, preservation method is assumed to be migration. In practice, emulation may also be applied if it is important to preserve the original look and feel of the publication. In an ideal world such migrations between the file formats would be lossless; in practice that may not be the case. Migrated document may look different even if the content is the same, and in the worst case semantics changes as well. Therefore archives often preserve also the original version of the archived resource, alongside more modern versions.

and/or archival, or refer to external documents listing these formats. File formats suitable for submission but not for archival are migrated during the ingest process, although the original files may be included in the AIP.

The submission agreements may also refer to SIP schema specifications, which provide more guidelines for document producers. Schemas may utilize long-term preservation standards such as METS (Metadata Encoding and Transmission Standard). Together the submission agreement and related documents should give a producer a clear idea on when and which publications should be sent to the repository system, which file formats and metadata specifications should be used, means of data transfer available etc. These requirements should cover both ingest and dissemination; that is, submission of documents to the repository system by the producer, and retrieval of the archived documents by customers.

This specification (ISO 22424 Part 1: Principles) outlines the general principles for the submission of EPUB publications from digital asset management systems to repository systems. The principles of archival storage or dissemination of archived documents are not covered here, because OAIS archives may apply various methods and processes to meet the requirements of submission agreements. Bit level preservation is also out of scope; the purpose of this specification is to make it easier for producers and OAIS archives to preserve access to EPUB documents.

The second part of this specification (ISO 22424 Part 2: Metadata requirements) provides a technical basis to meet the principles listed in this document by specifying metadata required for long-term preservation, and a method for packaging this metadata with the original EPUB container.

This specification is applicable to EPUB versions 3.x and as such it should be used cautiously with other (previous or later) versions of the standard. If there is a need to preserve documents that are in earlier EPUB versions, they do not need to be migrated, provided that a) submission agreement specifies those EPUB versions as archivable formats, and b) there are reading systems for these EPUB versions. Additional features in future EPUB versions should be analyzed from long-term preservation point of view. If such analysis reveals that they may constitute a risk, they should be avoided in submitted EPUB publications, or removed during ingest.

Annex A in this specification provides a summary of issues and recommendations related to the EPUB standard and its usage from long-term preservation point of view.

Digital publishing — EPUB 3 Preservation — Part 1: Principles

1 Scope

This document supports long-term preservation of EPUB publications via a dual strategy. First, it considers EPUB features from long-term preservation point. Some EPUB features are forbidden and some others required, depending on how they relate to a long-term preservation. EPUB publications constructed according to these guidelines should be suitable for preservation.

Second, this specification makes EPUB compliant with current practices of OAIS archives and technical requirements of repository systems. The former tend to rely on Open Archival Information Systems (OAIS) in their operations; the latter prefer to ingest electronic documents only in containers conforming to standards such as METS (Metadata Encoding and Transmission Standard).

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC TS 30135 (all parts), *Information technology — Digital publishing — EPUB3*

ISO 14721, *Space data and information transfer systems – Open archival information system (OAIS) – Reference model*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply. Unless stated otherwise, the terms have been adopted from ISO 14721:2012.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— IEC Electropedia: available at http://www.electropedia.org/

— ISO Online browsing platform: available at https://www.iso.org/obp

3.1
access functional entity
OAIS functional entity that contains the services and functions, which make the archival information holdings and related services visible to Consumers

3.2
administrative metadata
metadata that provides information to help manage a resource, such as when and how it was created, file type and other technical information, and access rights

[SOURCE: Understanding metadata]

3.3
archival information package
AIP
Information Package consisting of Content Information and associated Preservation Description Information (PDI), which is preserved within an OAIS

3.4

archive

OAIS archive

organization that intends to preserve information for access and use by a Designated Community

3.5

authenticity

property that an entity is what it claims to be

[SOURCE: ISO/IEC 27000]

Note 1 to entry: Authenticity is judged on the basis of evidence.

3.6

bit preservation

term used to denote a very basic level of preservation of digital resource as it has been submitted (literally the preservation of the **bits** forming a digital resource)

Note 1 to entry: This may include maintaining onsite and offsite backup copies, virus checking, fixity-checking, and periodic refreshing to a new storage medium.

Note 2 to entry: Bit preservation is not digital preservation but it does provide a building block for the more complete set of digital preservation practices and processes that ensure the survival of digital content and also its usability, display, context and interpretation over time.

[SOURCE: Digital preservation handbook, Glossary]

3.7

consumer

role played by those persons or client systems, who interact with OAIS services to find preserved information of interest and to access that information in detail

Note 1 to entry: This can include other OAISs, as well as internal OAIS persons or systems.

3.8

content information

set of information that is the original target of preservation or that includes part or all of that information

Note 1 to entry: It is an Information Object composed of its Content Data Object and its Representation Information.

3.9

context information

information that documents the relationships of the Content Information to its environment

Note 1 to entry: This includes reasons why the Content Information was created and how it relates to other Content Information objects.

3.10

core media type

a set of publication resource for which no fallback is required.

[SOURCE: EPUB Publications 3.0.1]

Note 1 to entry: Core media types have been specified in chapter 5.1. of the EPUB publications specification, version 3.0.1.

EXAMPLE core media types for still images are image/gif, image/jpg, image/png and image/svg+xml. Any other still image file format is foreign and requires a fallback, meaning the same resource expressed in another foreign format or core media type.

3.11
data, pl
reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing

[SOURCE: ISO 5127:2017]

Note 1 to entry: Data are often understood as taking the form of a set of values of qualitative or quantitative variables.

3.12
data dictionary
organized and constructed (electronic data base) compilation of descriptions of data concepts that provides a consistent means for documenting, storing and retrieving the syntactical form (i.e. representational form) and the meaning and connotation of each data concept

[SOURCE: ISO 24531:2013]

Note 1 to entry: PREMIS[22] is a data dictionary.

3.13
descriptive metadata
descriptive information
metadata about a resource for example for discovery and identification

Note 1 to entry: These can include elements such as title, abstract, author, and keywords.

[SOURCE: Understanding metadata]

3.14
designated community
identified group of potential Consumers who should be able to understand a particular set of information

Note 1 to entry: A Designated Community may be composed of multiple user communities. The community is defined by an Archive, though this definition may change later on.

3.15
digital preservation
series of managed activities necessary to ensure continued access to digital materials for as long as necessary

Note 1 to entry: Digital preservation refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological and organizational change

Note 2 to entry: Those materials may be records created during the day-to-day business of an organization; "born-digital" materials created for a specific purpose (e.g. teaching resources); or the products of digitisation projects.

[22] PREMIS Data Dictionary for Preservation Metadata (https://www.loc.gov/standards/premis/) is a leading metadata specification for metadata needed for long-term preservation.

EXAMPLE 1 **Short-term preservation -** Access to digital materials either for a defined period of time while use is predicted but which does not extend beyond the foreseeable future and/or until it becomes inaccessible because of changes in technology.

EXAMPLE 2 **Medium-term preservation -** Access to digital materials beyond changes in technology for a defined period of time but not indefinitely.

EXAMPLE 3 **Long-term preservation -** Access to digital materials, or at least to the information contained in them, indefinitely.

[SOURCE: Digital preservation handbook, Glossary]

3.16
digital rights management
DRM
packaging, distributing, controlling, and tracking content based on rights and licensing information

[SOURCE: ISO 19153:2014]

3.17
digital signature
signature
data appended to, or a cryptographic transformation of, a data unit that allows the recipient of the data unit to prove the source and integrity of the data unit and protect against forgery, e.g. by the recipient

[SOURCE: ISO/IEC 19784-1:2006]

3.18
dissemination information package
DIP
information package, derived from one or more AIPs, sent by an Archive to a Consumer in response to a request in the OAIS

3.19
distributable object
component of an EPUB publication that can be reused in other contexts

Note 1 to entry: A Distributable Object can be a complete EPUB Content Document (e.g., a chapter of a book), a section of such a document (e.g., an exercise or a promotional excerpt), a media resource (e.g., a video or interactive feature), or a combination of such resources that are not necessarily contiguous within the parent EPUB publication but are intended to be able to be distributed as a unit.

[SOURCE: EPUB Distributable Objects 1.0]

3.20
electronic book
e-book
non-serial digital document, licensed or not, where searchable text is prevalent, and which can be seen in analogy to a print book

Note 1 to entry: The use of e-books is, in many cases, dependent on a dedicated device and/or a special reader or viewing software.

[SOURCE: ISO 2789:2013]

3.21
EPUB container
ZIP based packaging and distribution format for EPUB publications

[SOURCE: EPUB Publications 3.0.1]

3.22
EPUB content document
publication resource that conforms to one of the EPUB content document definitions

[SOURCE: EPUB Publications 3.0.1]

3.23
EPUB navigation document
specialization of the XHTML content document, containing human- and machine-readable global navigation information

[SOURCE: EPUB Publications 3.0.1]

3.24
EPUB publication
collection of one or more renditions conforming to the EPUB specifications, packaged in an EPUB container

[SOURCE: EPUB Publications 3.0.1]

3.25
EPUB reading system
system that processes EPUB publications for presentation to a user in a manner compliant with EPUB specifications

[SOURCE: EPUB Publications 3.0.1]

3.26
fallback
mechanism with which versions of the same resource in different file formats can be linked to one another

[SOURCE: EPUB Publications 3.0.1]

Note 1 to entry: A reading system that does not support the file format of a foreign resource shall traverse the fallback chain until it finds a version it can render.

3.27
fixity information
information that documents the authentication mechanisms and provides authentication keys to ensure that the Content Information object has not been altered in an undocumented manner

[SOURCE: ISO 13527:2010]

3.28
foreign resource
publication resource that is not a core media type

[SOURCE: EPUB Publications 3.0.1]

3.29

identifier

data string or pointer that establishes the identity of an item, institution, or person alone or in combination with other elements.

[SOURCE: ISO 8459:2009]

Note 1 to entry: EPUB 3 specifies Unique Identifiers and Release Identifiers; the latter is a combination of a Unique Identifier and the last modification data of the rendition of the resource.

3.30

independently understandable

characteristic of information that is sufficiently complete to allow it to be interpreted, understood, and used by the Designated Community without having to resort to special resources not widely available, including named individuals

3.31

information

any type of knowledge that can be exchanged

Note 1 to entry: In an exchange, this is represented by data

EXAMPLE a string of bits (the data) accompanied by a description on how to interpret the string of bits as numbers representing temperature observations measured in degrees Celsius (the representation information)

3.32

information package

logical container composed of optional content information and optional associated preservation description information

3.33

ingest functional entity

OAIS functional entity that contains the services and functions that accept SIPs from producers, prepares AIPs for storage, and ensures AIPs and their supporting descriptive information become established within the OAIS

3.34

long-term

period of time long enough to raise concerns about the impact of changing technologies, including support for new media and data formats, and of a changing designated community, on the information being held in an OAIS

Note 1 to entry: This period extends into the indefinite future.

3.35

long-term preservation

act of maintaining information, independently understandable by a designated community, with evidence supporting its authenticity over the long-term

3.36

manifest

EPUB manifest element provides an exhaustive list of the Publication Resources that constitute the given Rendition, each represented by an item element.

[SOURCE: EPUB Publications 3.0.1]

3.37 metadata

data about other data, documents, or records that describe their content, context, structure, format, provenance, and/or rights.

[SOURCE: ISO 5127:2017]

3.38

METS

Metadata Encoding and Transmission Standard, a standard for presenting metadata using XML.

[SOURCE: Digital preservation handbook, Glossary]

3.39

migration

means of overcoming technological obsolescence by transferring digital resources from one hardware/software generation to the next

Note 1 to entry: The purpose of migration is to preserve the intellectual content of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology.

Note 2 to entry: Migration differs from the refreshing of storage media in that it is not always possible to make an exact digital copy or replicate original features and appearance and still maintain the compatibility of the resource with the new generation of technology.

[SOURCE: Digital preservation handbook, Glossary]

3.40

Open Archival Information System

OAIS

archive, consisting of an organization, which may be a part of a larger organization, of people and systems, that has accepted the responsibility to preserve Information and make it available to a Designated Community. It has a set of responsibilities, as defined in section 4, which allow an OAIS Archive to be distinguished from other uses of the term 'Archive'.

Note 1 to entry: The term 'Open' in OAIS is used to imply that this Recommendation and future related Recommendations and standards are developed in open forums, but it does not imply access to the Archive is unrestricted.

Note 2 to entry: The OAIS abbreviation is also commonly used to refer to the Open Archival Information System Reference Model standard which defined the term. The standard is a conceptual framework describing the environment, functional components, and information objects associated with a system responsible for long-term preservation.

3.41

package document

publication resource that describes one rendition of an EPUB publication, as defined in package document. The package document carries meta information about the Rendition, provides a manifest of resources and defines the default reading order.

[SOURCE: EPUB Publications 3.0.1]

Note 1 to entry: It specifies all tools required to render the document, provides an exhaustive list of resources belonging to the document, and defines their default reading order.

3.42
PDF
Portable Document Format, a set of formats and open standards maintained by the International Organization for Standardization for producing and sharing electronic documents

Note 1 to entry: Originally developed by Adobe Systems.

[SOURCE: Digital preservation handbook, Glossary]

3.43
PDF/A
versions of the PDF standard intended for archival use

[SOURCE: Digital preservation handbook, Glossary]

3.44
pre-ingest
actions required before data can be submitted into an OAIS archive, including negotiation of data acquisitions, checking rights and access criteria, licensing, and data submission

Note 1 to entry: This area also includes activities involving data producer support and training.

Note 2 to entry: Pre-ingest is not a function in the standard OAIS model, but activities in this area can form a significant part of a producer's responsibilities.

[SOURCE: UK Data Archive. Archive training manual[23]]

3.45
preservation description information
PDI
information necessary for the adequate preservation of Content Information that can be categorized as provenance, reference, fixity, context, and rights information

3.46
preservation metadata
metadata containing information needed to archive and preserve a resource

[SOURCE: Understanding metadata]

3.47
preservation planning functional entity
OAIS functional entity that provides the services and functions for monitoring the environment of the OAIS and that provides recommendations and preservation plans to ensure information stored in the OAIS remains accessible to, understandable by, and sufficiently usable by the designated community over the long-term, even if the original computing environment becomes obsolete

3.48
producer
role played by those persons or client systems that provide the information to be preserved

Note 1 to entry: This can include other OAISs or internal OAIS persons or systems. The producer does not need to be the publisher.

[23] http://www.data-archive.ac.uk/curate/archive-training-manual/pre-ingest

3.49

provenance information

information that documents the history of the Content Information

Note 1 to entry: This information states the origin or source of the Content Information, any changes that may have taken place since it was generated, and who has had custody of it.

Note 2 to entry: The Archive is responsible for creating and preserving Provenance Information from the point of ingest; however, earlier Provenance Information should be provided by the Producer. Provenance Information adds to the evidence to support authenticity.

3.50

publication resource

resource that has the content or instructions contributing to the logic and rendering of at least one rendition of an EPUB publication

EXAMPLE Examples of publication resources include a rendition’s Package Document, EPUB Content Document, EPUB style sheets, audio, video, images, and embedded fonts and scripts.

[SOURCE: EPUB Publications 3.0.1

3.51

reading system

system that processes EPUB publications for presentation to a user in a manner conformant with EPUB specification

[SOURCE: Modified from EPUB Publications 3.0.1]

3.52

reference information

information that is used as an Identifier for the Content Information

Note 1 to entry: This also includes Identifiers that allow outside systems to refer unambiguously to a particular Content Information.

EXAMPLE an ISBN is a type of Reference Information.

3.53

reference model

framework for understanding significant relationships among entities in an environment and for the development of consistent standards or specifications supporting that environment

Note 1 to entry: A Reference Model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a non-specialist.

3.54

reformatting

copying information content from one storage medium to a different storage medium (media reformatting) or converting from one file format to a different file format (file reformatting)

[SOURCE: Digital preservation handbook, Glossary]

3.55

refreshing

copying information content from one storage media to the same storage media

[SOURCE: Digital preservation handbook, Glossary]

3.56
release identifier
identifier that allows any instance of an EPUB publication to be compared against another to determine if they are identical, different versions, or unrelated

Note 1 to entry: Release Identifiers consist of a unique identifier and the last-modified date of the document.

[SOURCE: EPUB Publications 3.0.1]

3.57
remotely-hosted resource
objects hosted outside the EPUB Container.

3.58
rendition
one rendering of the content of an EPUB publication, as expressed by an EPUB package

[SOURCE: EPUB Publications 3.0.1]

3.59
repository system
long-term preservation system used by an archive

3.60
rights management metadata
information that identifies the access restrictions concerning the Content Information, including the legal framework, licensing terms, and access control

Note 1 to entry: This contains the access and distribution conditions stated in the Submission Agreement, related to both preservation (by the OAIS) and final usage (by the Consumer).

Note 2 to entry: It also includes specifications for the application of rights enforcement measures.

3.61
spine
EPUB spine element defines the default reading order of the EPUB Publication content by defining an ordered list of manifest item references.

[SOURCE : EPUB Publications 3.0.1]

3.62
structural metadata
metadata that indicates how compound objects are put together, for example how the pages of a document are arranged to form chapters

[SOURCE: Understanding metadata]

3.63
submission agreement
agreement reached between an OAIS archive and a Producer that specifies a data model and any other arrangements needed for the data submission session

Note 1 to entry: This data model identifies the format/content and the logical constructs used by the Producer and how they are represented on each media delivery or in a telecommunication session.

3.64

submission information package

SIP

information package that is delivered by a Producer to an OAIS to be used to construct or update one or more AIPs and/or the associated descriptive information.

3.65

unique identifier

primary identifier of an EPUB publication, which may be shared by one or several renditions of the same EPUB publication that conform to the EPUB standard and embody the same content.

[SOURCE: EPUB Publications 3.0.1]

3.66

XHTML content document

EPUB content document that conforms to the profile for HTML defined in XHTML Content Documents

[SOURCE: EPUB Publications 3.0.1]

Note 1 to entry: see EPUB Content Documents 3.0.1, chapter 2.

4 Abbreviated terms

AIP	Archival Information Package
DIP	Dissemination Information Package
DRM	Digital Rights Management
OAIS	Open Archival Information System
PDI	Preservation Description Information
SIP	Submission Information Package

5 Packaging standards

An archiving process includes several distinct steps. A producer – which may be the publisher or other body acting on behalf of the publisher, such as the archive itself - creates a Submission Information Package (SIP) and transfers it to a repository system in an OAIS archive. The archive performs a quality control process to the SIP and, if the package meets the criteria set in the submission agreement, accepts it, creates an Archival Information Package (AIP) and transfers the package to archival storage. During ingest some of the files or metadata records within SIP may be migrated to new formats or additional metadata may be added.

The OAIS archival storage function stores, maintains, and retrieves AIPs. Maintenance may include for instance frequent error checks to protect the data against bit rot. In order to keep the documents understandable it may also be necessary to migrate[24] them in new formats, or to update the AIP with additional metadata related to emulation. Migration and other preservation related tasks may be carried out by the producer, OAIS archive and / or third parties. The party or parties responsible should be specified in the submission agreement.

[24] From OAIS point of view, migration is a complex process which involves export of the document (as a migration DIP) and then migration during "ingest as new manifestation".

The OAIS Access function allows users to retrieve information from a repository system in the form of Dissemination Information Packages (DIPs) which can include all or parts of the data and metadata of an AIP. Differences between SIPS, AIPs and DIPs can be substantial, depending on the preserved content, requirements of submission agreement, national legislation and institutional practices. OAIS does not require a 1:1 relationship between information packages, so one AIP can contain documents and metadata from multiple SIPs or vice versa.

Transfers of package states (SIP to AIP to DIP) do NOT mean that the content SHALL change. The change from SIP to AIP can be minimal, that is, the content information remains the same, but some administrative metadata is added into the AIP about the actions taken during the ingest process. If an EPUB publication is created according to the requirements in this document there should be no need for reformatting the EPUB publication itself. During ingest it is enough to check the validity of the document, and if there are no issues, it can be stored "as is". Some archives may choose to apply even simpler initial ingest procedures (that is, avoid even validity checks) if the producer is well known and reliable, such as other OAIS archive.

This specification covers only the initial stage of the archiving process, namely the creation of Submission Information Package (SIP). SIP consists of data objects and representation information with which the data is interpreted. Both the data (documents) and representation information (metadata) MUST conform to the standards and specifications the producer and the archive have agreed upon in the submission agreement. If a SIP does not meet the requirements, ingest to the repository system fails. Note that a SIP MAY contain unarchivable resources, provided that they have been encoded in an appropriate manner.

The content and structure of all information packages in repository systems MUST be standardized. There are several packaging standards available, but the most commonly used one is the Metadata Encoding and Transmission Standard (METS[25]) developed by the Library of Congress. ISO/IEC JTC1/SC34 JWG 7 decided to recommend the use of METS as the container standard, although this specification does allow the use of other container standards as well.

Since container standards – including METS - are rich specifications there is a need to create profiles to specify how they should be used. This specification provides a METS profile for EPUB in Part 2. Other container standards are not taken into account; if METS is replaced by another container specification, profiling needs to be done separately.

Some digital preservation projects have developed tools for creating SIPs that meet the project requirements, which makes it a lot easier to submit information to the repository system. Producers SHOULD nevertheless have at least basic understanding of digital preservation, since pre-ingest steps from document creation to SIP submission should not risk the authenticity of the documents to be submitted to the OAIS archive.

[25] http://www.loc.gov/standards/mets/

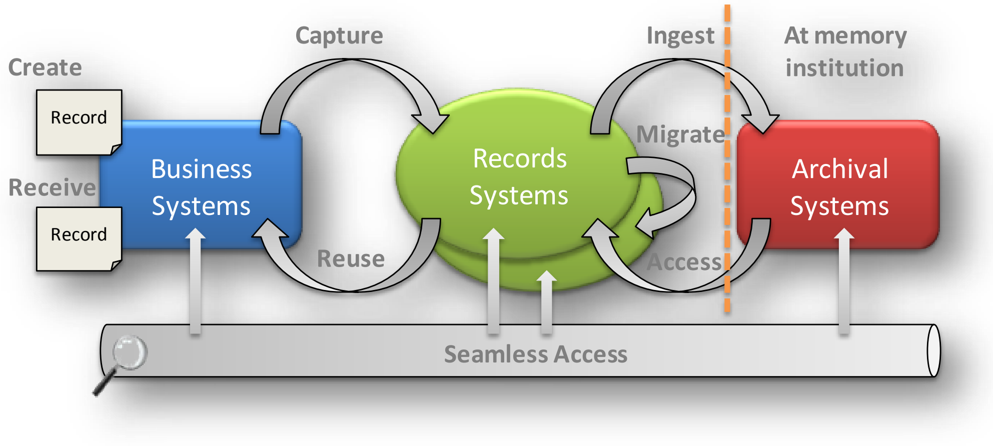

Figure 2. Information flow between live and archival systems
[E-ARK Common specification, p. 13]

Different disciplines, even if they all use OAIS, will develop interfaces optimized for their own needs. And if the payloads are not the same, technical metadata standards will also differ. Domains have also adopted different packaging and preservation metadata standards. Almost all digital archiving projects in the library domain rely on METS and PREMIS specifications. Some libraries use BagIt[26] instead of METS for storage of ditigal objects, but BagIt specification does not require knowledge of the semantics of the resources in the container, whereas METS supports such metadata. Therefore BagIt is not an alternative to METS for long-term preservation.

Compared with libraries, the film industry started digital preservation efforts a bit later and may eventually develop different preferences[27]. And even if the same standards were used, they may be applied in a non-interoperable way even within the same domain. Therefore creating set application profiles is important in digital archiving.

6 Construction of OAIS information packages

According to the Open Archival Information System (OAIS) model[28], information package is "a container that contains two types of Information Objects, the Content Information and the Preservation Description Information (PDI)". Content information is the data that needs to be preserved and preservation description information is the metadata and other information that is needed in order to preserve, find and understand the data in long-term.

Preservation description information consists of reference information, provenance information, context information, fixity information, and access rights information. See the OAIS specification for an in depth explanation of these.

According to the OAIS specification (pages 4-35),

> *[i]t is necessary to distinguish between an Information Package that is preserved by an OAIS and the Information Packages that are submitted to, and disseminated from, an OAIS. These variant packages are needed to reflect the reality that some submissions to an OAIS will have insufficient PDI to meet final OAIS preservation requirements. In addition, they MAY be organized very differently from the way the OAIS organizes the information it is preserving. Finally, the OAIS MAY provide information to Consumers that does not include all the PDI*

[26] https://tools.ietf.org/html/rfc8493
[27] https://www.cen.eu/news/calls/Calls/CEN-Call_for-tender_Digitalcinema.pdf
[28] http://public.ccsds.org/publications/archive/650x0m2.pdf

> *with the associated Content Information being disseminated. These variants are referred to as the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP). Although these are all Information Packages, they differ in mandatory content and the multiplicity of the associations among contained classes.*

The principles listed below provide SIP production guidelines for document producers (publishers or third parties creating EPUB publications). The creation of the principles has been inspired by the draft common requirements published by the E-ARK project (see Introduction to the Common Specification for Information Packages in the E-ARK project, version 1.0[29]). Although E-ARK has served as a model for this specification, these requirements have not been aligned with those of E-ARK, and therefore there may be significant differences between the specifications.

6.1 General

6.1.1 EPUB publications SHALL be sent to a repository system as well-formed and complete Submission Information Packages (SIPs)

- This specification does not assume that publishers create SIPs. The OAIS producer MAY be a third party acting on behalf of the publisher, such as hosting platform or other production vendor or even the OAIS archive itself.
- This specification and its accompanying document are mainly concerned with the structure and content of SIPs. The way EPUB publications are archived and disseminated (the structure of Archival Information Packages and Dissemination Information Packages, or AIPs and DIPs) depends on the submission agreements made between the archive and the producers, and on the operational principles of the archive, and is beyond the scope of this document. It is possible that an EPUB publication is migrated into another format during Ingest, and disseminated again as an EPUB publication. The archive may also preserve (in bit level) the original file.
- Submitted EPUB publications SHALL be conformant with EPUB requirements[30] and conformance SHOULD be validated.
- Submitted EPUB publications SHALL either contain or at least facilitate access to all the data and metadata required to render the content information successfully.
  - i. Preview publications MAY be submitted, even though they are by definition not complete, if the final documents are sent when ready. Depending on the submission agreement, the archive MAY preserve just the final version, or both versions of the resource. Identifiers SHALL be used in such a way that the OAIS archive will be able to link all versions of the publication and delete preview versions, if that is the agreed preservation policy.
  - ii. Distributable objects SHALL NOT be submitted individually. They MAY be embedded within EPUB publications, but the archive is not obliged to deliver them as DIPs unless the submission agreement mandates that.
  - iii. Fonts SHALL be embedded into the EPUB publication in full and un-obfuscated, if font license allows that. If submission agreements allow submission of EPUB publications with obfuscated or non-embedded fonts, there is a risk that such publications become unusable in the future.
  - iv. Related resources such as audio and video SHOULD be embedded in the EPUB publication.

[29] http://www.dasboard.eu/specifications/common-specification

[30] Conformance requirements for EPUB publications and reading systems have been specified in chapter 3.1 of EPUB Publications, version 3.0.1.

   v. Remotely-hosted resources SHOULD be avoided, but if used, it is necessary to ensure that all remote data is available to the archive so that the data can be incorporated into the AIP during ingest, and permission to do this SHALL be explicitly agreed upon in the submission agreement, especially if the publisher is not in full control of remote data.
   vi. Descriptive and other metadata SHOULD be embedded in the SIP. METS mdRef element MAY only be used if a) referred metadata is part of the same SIP, or b) the archive is able to retrieve any linked external metadata and incorporate it into the AIPs in an appropriate format.
   vii. Permission to use remote resources and metadata SHALL be specified in the submission agreement[31]. The permission SHALL specify acceptable metadata and file formats.

- The SIP SHOULD[32] be checked for viruses and malicious software before submission to the repository system.
- EPUB publications in SIPs SHOULD NOT be encrypted, because that compromises long-term preservation. If data is submitted in an encrypted format, the archive SHALL receive necessary decryption information/details within the SIP, as agreed in the submission agreement or elsewhere. When the archive disseminates the archived data to its customers, it can be encrypted again.
- DRM protection, if any, SHOULD be removed by the producer before the document is submitted. If the content in the SIP is DRM protected, the archive SHALL receive the necessary information/details to remove the DRM protection within the SIP, as agreed in the submission agreement or elsewhere. Such permission may be producer-specific, based on the submission agreement, or a generic permission, based on e.g. the Copyright Act.
- If data is compressed, the user of the compression method SHALL be specified using the Compression metadata element in the EPUB's encryption.xml file.
- The submission agreement SHOULD specify at least one EPUB reading system capable of rendering the submitted EPUB publications successfully. Knowing the reading system requirements in advance makes it easier for the archive to design and implement the ingest process. Although submitted publications will usually be validated only with automated tools[33], the archive should be able to validate that the received EPUB can be presented to the customers, and check for instance the look and feel of archived EPUB publications before and after migration. This is possible only if the archive can operate the reading systems that can render the archived publications successfully.
- Each SIP SHOULD specify EPUB reading system or systems, which can render the EPUB publication in the SIP. If this information is missing, reading system or systems SHALL be specified in the submission agreement.
- Multiple-rendition EPUB publications may be designed for multiple reading systems, in which case the submission agreement may require the archive to carry out at least occasional checks in all of these reading systems. If so, all these reading systems SHOULD be listed in the submission agreement.

[31] If there are remote resources or associated metadata linked to the SIP with a LINK element, these external resources will be retrieved as part of the ingest process and included in the AIP. If external resources cannot be retrieved, the ingest process fails. Submission agreement SHOULD specify how to handle such a situation. For instance, the agreement can require the producer either sends a new SIP with all the data and metadata embedded into it, or makes sure that the archive is allowed to access remote data and metadata.

[32] Some producers may not be able to make virus checks, but all OAIS archives SHALL be. Virus checks are commonly done during ingest.

[33] One such tool is Epubcheck, available from https://github.com/idpf/epubcheck

- If a submitted EPUB publication has been optimized for a certain reading system, the system SHOULD be described in the document's technical and/or preservation metadata, since such information is valuable for preservation and archival access purposes.
- If the optimal EPUB reading system is no longer available, the archive SHOULD, with permission and support from the producer, either find another suitable reading system or modify the ingest process so that the EPUB publications affected by this change can be used by another EPUB reading system.[34]

6.1.2 Regardless of its type or format, it SHALL be possible to include any data or metadata in SIPs

- It SHOULD be possible to maintain the SIP and EPUB specifications independently, i.e. so that any change to SIP does not automatically mean that the EPUB format needs to be updated and vice versa. The exception from this rule is that any existing and future features in EPUB specification which are relevant from long-term preservation point of view such as font embedding SHALL be taken into account in the SIP specification.
- This document does not set a priori constraints either to the current or future versions of EPUB with regard to the choice of metadata and file formats or either's versions (see note 1 below on EPUB Core media types).
- The submission agreement SHOULD specify metadata formats and file formats approved for submission and archival. For EPUB publications, at least Dublin Core metadata format and all EPUB core media types SHALL be supported by the archive in order to guarantee efficient processing of EPUB publications.
- Submission agreements SHOULD specify what kind of executables can be embedded in the submitted EPUB publications (see note 2 below on interactive e-books and EPUB publications).

NOTE 1 EPUB community may change the list of EPUB Core Media Types any time, independent of the EPUB specification updates. New core media types may be approved and old ones deprecated. If core media types are not checked from long-term preservation point of view, some new EPUB core media types may turn out to be non-archivable.

File format lists in submission agreements may cover all EPUB core media types or – if the producer does not use all the core media types - just a subset of them. When a core media type is deprecated, the producer (if it still exists) and the archive should decide whether the file format in question is migrated or kept as is (and emulated). If the latter, it may be necessary to migrate the deprecated file format when DIP packages are created.

NOTE 2 E-books are likely to contain more interactive features in the future. From preservation point of view it is therefore a problem that there are various ways in which EPUB 3 can support interactivity. On the other hand, some EPUB reading systems may not support interactivity at all, and even if it is supported, different reading systems may not behave identically, partly because EPUB is not specific about how support should be implemented. EPUB 3 `object` element enables the use of arbitrary embedded executables that are not inherently supported in EPUB 3 reading systems. A common use case would be to include proprietary applets or Adobe Flash applications. However, in a majority of cases, interactive publications will be created through the use of in-book source code. Because JavaScript is the de facto standard scripting language for SVG and HTML5, EPUB 3 content documents can be assumed to be scriptable only if they contain JavaScript code. The standard does not define which versions of JavaScript (ECMAScript) are required to get the support. Content creators should comply with the most commonly supported features in web browsers for

[34] While this standard is about the "state" in which the EPUB publication itself shall be in order to be archivable, the SIP may include a lot of other information (metadata, executables, other renditions of the EPUB publication, additional documentation etc) which may make it easier to preserve the intellectual content in the long-term.

best results [Daly]. Usage of common tools and techniques will also make it easier to preserve the publication in the long-term, either via emulation (a common solution for software preservation) or migration.

- Archives offering long-term preservation services for EPUB publications SHOULD keep track of EPUB core media types and consider the possibility of including them on the list of archivable formats. If this is not viable, the archives SHOULD maintain clearly defined and well tested migration pathways from non-archivable core media types into archivable formats. Then the archive would not need to migrate these images during ingest and it would be possible to preserve EPUB publications unchanged[35].
- If there is a foreign resource embedded or linked to a submitted EPUB publication, a fallback chain ending in a core media type resource SHOULD be provided even if the foreign resource is in an archivable format. (Note that this requirement is stricter than those in EPUB 3.x specifications, which require a fallback only in certain situations.)
- The producer MAY include foreign resources (and metadata formats) in submitted EPUB publications if they have been specified as suitable for ingest and/or archivable in the submission agreement, or if their METS encoding in SIPs makes it possible to ignore them during ingest (see below).
- If foreign resources and metadata are originally in un-archivable formats (formats that have not been specified as acceptable in the submission agreement), they SHALL be migrated during pre-ingest. The SIP may contain either just migrated publications, or both the original and migrated publications. Note that the preservation method MAY be either emulation or migration, so this requirement does not mean migration-only approach.
- Core media types and foreign resources not specified in the submission agreement MAY be submitted if and only if the submission agreement allows it. METS encoding of these files in SIPs SHALL make it possible to skip their validation against the generic ingest criteria during the ingest process (since otherwise the SIP shall not pass the validation) and therefore passed directly to AIP. The specifics of this type of encoding SHALL be defined in the submission agreement.
- If there are alternative versions (renderings) of the publication to be included in the SIP which are not archivable, they SHOULD be migrated into acceptable file formats prior to submission by the producer or a third-party preparing a SIP on behalf of the producer. For instance, if PDF is specified as not archivable but PDF/A is, the producer should create a PDF/A version of the document, which will then be submitted to the repository system alongside the EPUB publication of the same work.
- If these non-archivable originals are submitted, their METS encoding in SIPs SHALL make it possible to skip validation against the generic ingest criteria during the ingest process (since otherwise the SIP would not pass) and therefore passed directly to AIP. The specifics of this type of encoding SHALL be defined in the submission agreement[36].

NOTE EPUB 3 Fixed Layout Properties

In digital preservation the usual aim is to preserve intellectual content. Preserving also the original look and feel of the document is more challenging, although that may be required for some resources or collections. Reflowable EPUB publications are designed so

[35] An OAIS archive does not need to migrate non-archivable file formats during the ingest process. Depending on the preservation strategy, migration may only happen when a real risk to the format emerges – such as the loss of applications capable of rendering it - or when the document is disseminated for the first time.

[36] Ideally, a well-designed and built repository system should be able to validate any file format. In practice, there are file formats validation tools cannot process. If there is a need to preserve these files in bit-level, they have to be ignored during validation.

that their look and feel can change with no impact on semantics, which is a good thing from the digital preservation point of view, since in these documents EPUB content presentation adapts to the user preferences and display properties, which will change in the future.

In fixed layout EPUB publications the intellectual content and the design of the document cannot be separated: any change in the appearance of the document may cause significant changes in the meaning or even lose it completely. Therefore fixed-layout EPUB publications give the content creators greater control over presentation. This control is based on a set of metadata properties with which the intended rendering dimensions can be specified [EPUB 3 Fixed]. However, if the document is migrated, these metadata properties may be lost, and even if that does not happen changes in hardware (e.g. display technologies), operating systems, and middleware may change the original look and feel of the document. Therefore emulation of the original hardware and software environment is likely to be the best approach for preserving such documents.

Submission agreements SHOULD specify if submission of fixed-layout EPUB publications is allowed and if so, how they are treated during ingest. One solution is to include in SIPs also reflowable versions of these publications. If this is not possible or practical, SIP SHOULD contain metadata supporting emulation of the EPUB publication or publications in the package.

6.1.3 It SHOULD be possible to transfer SIPs by any means, methods, or tools from the submitting organization to the repository system

- Although there are no general limitations (it is possible to use e.g. FTP or UPS), submission agreements MAY limit the options available by specifying the protocols to be used during submission.
- SIPs SHALL be composed so that their structure and content does not limit the use of any particular transfer method.

6.1.4 The archive SHALL have a way to verify the identity of the submitting organization/person, no matter how the information packages are transferred

- If submission is taken care of by a third party service and the producer is a different organization of person, the archive SHALL be able to verify the identity of both of them.
- There are various ways to implement this requirement, including digital signatures, secure channels, recording relevant information within the SIP as metadata, or even manual exchange of data on secure media.
- Part 2 of this specification provides an example of how a digital signature can be used for verification.

6.1.5 There is no 1:1 relation between OAIS information packages

- SIPs SHALL be composed so that their structure and content SHALL NOT prescribe or limit SIP -> AIP -> DIP conversions.
- During ingest, it SHALL be possible to transform one SIP into 1-n AIPs, or many SIPs into 1-n AIPs. For instance, a SIP might consist of all yearbooks of a publisher (e.g. 15 EPUBs) which are then archived in separate AIPs. Relevant data and metadata SHALL always archived; number of AIPs created during ingest depends on the internal practices and processes of the archive, which are not within the scope of this specification.

6.1.6 A SIP MAY contain 0-n EPUB 3 publications, and one EPUB 3 publication MAY be submitted to the repository system in 1-n SIPs

- A SIP MAY contain only metadata about EPUB publication, not the publication itself.
- A SIP MAY contain multiple EPUB publications; for instance, different renderings of the same document[37]. If so, the SIP SHALL contain descriptive and administrative metadata which allows these publications to be ingested separately.
- A SIP MAY contain alternative renderings (such as PDF or DOCX) of the publication, but if so, the SIP SHALL contain all administrative metadata required for processing of these versions, and explaining the relations between these renderings.
- A single EPUB publication MAY be split into multiple SIPs if there is a valid reason to do so, such as the complexity or large size of the document.

6.1.7 The information package type (in this case, SIP) SHALL be indicated

- Only packages which are marked to be SIPs will be ingested. AIPs, DIPs and unlabeled packages are not suitable for ingest.

6.1.8 SIP packaging method SHALL not restrict the application of any preservation method

- Although the most common preservation method is migration, some archives MAY choose emulation as the primary approach, which will have an impact on the OAIS Preservation Description Information required.
- Some information objects (such as programs) are not suitable for migration. Submission agreements SHOULD specify a preservation strategy for such resources.

6.1.9 The packaging method SHALL NOT limit the size of the SIP

- Some archives can have problems in e.g. validating and ingesting very large data objects. If there is a risk that the SIPs are becoming too large for the submission method used or the ingest process used by the archive, an appropriate splitting mechanism SHOULD be applied. Describing such mechanisms is beyond the scope of this specification.

6.2 Identification of information packages and their content

6.2.1 It SHALL be possible to identify any SIP uniquely both during and after the ingest process

- Since multiple SIPs may be submitted to the repository system simultaneously, there is a need to identify all packages in a (globally) unique manner. Identification will also make it possible to relate the packages with appropriate submitters, earlier submissions etc. Such identification helps to streamline the whole submission process and any potential communication between the archive and the submitting organization.
- Once the ingest process has been completed and 1-n AIPs have been formed, the SIP itself is no longer needed, but sometimes it is necessary to acquire more information about submitted publications from the producer, and SIP identifier is often necessary for that. Therefore both the SIP identifier and the AIP identifier(s) which the producer receives after the SIP has been ingested SHALL be persistent.
- There are circumstances in which AIP identifiers SHOULD be not only persistent, but also globally unique. For instance, an OAIS archive can cooperate with other archives by exchanging AIPs in order to share the bit level preservation costs.

[37] OAIS archives may have different ideas of what “interrelated” means. For instance, archives tend to prefer large SIPs which may contain large number of documents gathered for years, while libraries archive publications on an individual basis.

- The entire SIP or parts of it SHALL be resubmitted in a revised format if the ingest process fails due to errors in the package. To keep track of the packages, SIPs SHALL have unique identifiers.

6.2.2 Information objects (EPUB publications, PREMIS preservation metadata record, etc.) within SIPs SHALL be identified uniquely and persistently

- Identifiers have many vital uses in digital preservation. They are used as access keys to the archived content in repository systems and facilitate information exchange with external systems. Identifiers also enable linking different versions of an archived document to each other. Moreover, with identifiers it is possible to link documents and descriptive/administrative metadata records that describe them. These links enable the archive to e.g. create dissemination information packages with the requested content.
- Submission agreements SHALL specify identifier systems used, their location (EPUB document or SIP) and who is responsible of creating them (producer, archive or a third party). For instance, if the use of EPUB release identifiers is forbidden because the repository system does not support them, another means of identifying releases is needed.
- International standard identifiers, such as ISBNs for books and DOIs for articles, SHALL be used as EPUB unique identifiers whenever possible. Any exceptions (such as using other identifier systems for releases which do not have ISBNs) SHOULD be specified in the submission agreement.
- It SHOULD be possible to express the identifiers (also) as actionable HTTP URIs. Usage of persistent identifiers (Handles, DOIs, URNs, or ARKs) is recommended.
- If there are multiple renditions of a work in an EPUB publication, requirements in the EPUB Multiple-Rendition Publications 1.0 specification SHALL be followed. Each rendition of an EPUB publication in a SIP SHALL have its own identifier.
- The SIP SHOULD contain separate descriptive and administrative metadata records for each rendition, and these records SHALL have their own identifiers.

NOTE 1 According to EPUB Multiple-Rendition Publications 1.0, the need to include more than one rendition of the content in an EPUB publication has grown as reading systems have become more sophisticated. In addition to optimizing the layout, adapting the content to specific reading systems may involve changing the content itself. Adaptation may also involve the prose of a textual work; instead of publishing several single-language EPUB publications multiple translations may be published as a single multiple-rendition EPUB publication.

NOTE 2 Standard work identifier such as ISTC (International Standard Text Code) would be an ideal means of linking all manifestations to each other. Unfortunately there is no widely used standard identifier for textual works, and therefore this document does not require work level identification. However, if such identifier is available and supported in all applications involved, it is a good idea to use it. Work identifiers can be used to detect duplication of intellectual content in the archive, and if they are used in producers' and publishers' systems as well, it is possible to check overlap and possible gaps.

6.2.3 EPUB Fragment Identifiers SHOULD not be used in EPUB publications sent to a repository system, unless the submission agreement explicitly allows their use

- EPUB Canonical Fragment Identifiers define a standardized method of referencing content within an EPUB publication through the use of URI Fragments. From the digital preservation point of view, fragment identifiers can be problematic if the preservation strategy is not emulation, since URI fragments are media type dependent. Following migration the fragment identifiers may no longer be functional, because the new media type does not support them.

- If fragment identifiers are allowed, the producer and the archive SHOULD take this into account in preservation planning, and design migrations so that the functionality provided by the fragment identifiers is preserved.

6.3 Structure of information packages

6.3.1 Submission information packages SHALL be built so that their components can be logically and physically separated from one another

- For each rendition of the EPUB content document, there SHALL be a manifest file, which identifies and describes a set of resources that collectively compose a given rendition of a document, and EPUB spine, which provides a default reading order for a given rendition.
- EPUB Open Container Format (OCF) defines a file format and processing model for encapsulating a set of related resources (for instance, renditions of the same resource) into a single-file (ZIP) EPUB Container[38].
- The structure of each EPUB ZIP archive SHALL be described using the EPUB container.xml file (which describes the locations of root files of available renditions of the EPUB publication, and the rendition's package document and navigation document).
- EPUB Package document and navigation document SHALL contain all metadata needed for rendering the publication, including the recommended reading system.

6.4 Generic Information package metadata

6.4.1 Metadata in information packages SHALL be based on standards

- METS or another agreed upon container format SHALL be used as the container standard, since this makes ingest to existing repository systems easier.
- The submission agreement SHALL specify at least one mandatory metadata format for descriptive metadata. The format does not need to be Dublin Core; although EPUB publications always contain some Dublin Core metadata elements (see below), they MAY contain more complete metadata in another format, such as ONIX.
- The minimum required descriptive metadata for EPUB publications are title, identifier, and language from the Dublin Core Metadata Element Set. Each rendition of a publication SHOULD also have at least the last modified date property from the DCMI Metadata Terms. Each rendition SHOULD also have the publication date encoded as DCMI Date, if the publication date is required to distinguish between publications.
- SIPs submitted to a repository system MAY[39] contain preservation metadata, although such metadata will normally not be created in production systems, but in repository systems during ingest. PREMIS SHOULD be used for preservation metadata, as it is the most widely used and supported standard for this kind of metadata.
- The submission agreement SHALL specify the syntax of metadata and its location (in the EPUB document, or in the SIP container), metadata formats used and metadata elements required or recommended.

[38] EPUB specifications do not require or recommend any specific ZIP tool. It is possible to use for instance ePubPack (https://sourceforge.net/projects/epubpack/) to create EPUB ZIP containers from a folder.

[39] Adding preservation metadata during pre-ingest might be tricky since preservation metadata is the core of any preservation system and it's use is highly regulated within repository systems. Errors in preservation metadata prepared by the submitter may cause serious problems in the preservation process.

- Since problems with text forms and encodings are common in repository systems, textual metadata SHOULD be provided also in Romanized form, using the EPUB alternate-script property to transcribe it if the metadata is originally in non-Roman script.

6.4.2 Metadata SHOULD allow (automatic) validation of the structure and content of SIPs in terms of integrity, fixity, and syntax

- SIPs SHALL contain message digests for all files of the SIP, and for the package itself.
- File format identification and validation metadata (created with EpubCheck[40] or other validator tool) SHOULD be included in the SIP, if a validator is available.

6.4.3 It SHALL be possible to edit metadata in information packages

- If ingest has failed because of erroneous or missing metadata, the producer or a third-party responsible for the submission SHALL be able to modify the SIP so that it meets the metadata requirements in the submission agreement.
- Producers and archives MAY use crowdsourcing and entity extraction activities to update descriptive metadata; an archive MAY choose to update this metadata also in the AIPs in the repository system although all the other components in the packages remain unchanged.

Annex A
(informative)

EPUB and digital preservation: issues and recommendations

The British Library's EPUB Format preservation assessment includes a preservation risk summary [Whibley, p. 7-8]. The risks mentioned in the BL assessment are marked with [BL].

A.1 EPUB standard: issues

- Lack of stability in the e-book sector
  - EPUB does not have universally widespread support across e-book devices [BL].
- Lack of EPUB format stability [BL]
  - Some EPUB versions have not been downward compatible. EPUB 3 is different from EPUB 3, and EPUB 3.1 from EPUB 3.0.1. The next version, 3.2, is based on 3.0.1, not on 3.1.
  - Due to rapid technical development, future e-books (and EPUB specification) are likely to differ from the current one.
  - Proprietary changes and non-standard use of specifications have been used and will be used to restrict usage of EPUB publications to specific manufacturer hardware/software
- Challenging EPUB features
  - From the long-term preservation point of view, the challenging features in EPUB include the possibility of using DRM, encryption and obfuscation, foreign resources and non-embedded resources.
  - While migration may be the best approach for most EPUB publications, interactive documents (containing software components) and fixed-layout documents are likely to be more suitable for preservation via emulation, so any OAIS archive preserving EPUB publications should be familiar with both preservation techniques.
- Lack of archivable EPUB version
  - The standard is becoming richer and richer, and publishers and other users may find it more difficult to specify and avoid counterproductive features from the long-term preservation point of view. Pre-ingest (modifying the EPUB publication so that it can be preserved easily) may be difficult unless it has been taken into account from the beginning.

Recommendations:

- W3C should actively promote the EPUB format, because it is the only open e-book standard and it is based on open standards such as HTML5 and CSS.
- EPUB community and digital preservation experts should develop a subset of EPUB ("EPUB/A") suited for long-term preservation.

A.2 EPUB usage: issues

- Ecosystem specific EPUB implementations
  - Major players in the e-book market (e.g. Amazon, Apple) have built EPUB based but closed (non-interoperable) ecosystems for e-books. E-books in vendor-specific formats, such as Amazon's KF8 should be migrated to EPUB before they are submitted to a repository system. Technically this is possible since EPUB is a "more or less obvious superset of what is possible in the different formats". The only exception is the fixed-

  layout document specification in KF8; it is based on percentage information, not on absolute pixel positions as in EPUB 3. [Bläsi, p. 38].
  - These players have also created vendor-specific DRM solutions, which prevent the use of archived EPUB publications with other vendor's reading devices, unless the DRM protection has been removed during pre-ingest or ingest.
- Encryption and obfuscation [BL]
  - Encryption may prevent the rendering of documents.
  - Where not easily substituted, obfuscated fonts may lead to loss of critical information.
- Incomplete support in EPUB viewers [BL]
  - Support for all aspects of the EPUB standard appears to be mixed, although impact of this is unclear. In the short-term, if the EPUB publication has been optimized for a specific reading system or systems, metadata embedded in the SIP should specify these systems. In the long-term, functionalities that are not widely supported may be lost.
- Losing information
  - Where not easily substituted, non-embedded fonts may lead to loss of critical information.
  - Metadata (and data) may not be embedded, but just linked to the SIP. During ingest, retrieval of linked information may fail.
- Invalid or badly formed EPUB files [BL]
  - May affect the ability to render files now or in the future.
- Documents relying on EPUB features that may be difficult to preserve
  - Fixed-layout documents: digital preservation usually concentrates on preserving the intellectual content, not the original look and feel of the document since that is regarded as difficult in the long-term. Preserving fixed-layout EPUB publications for the long-term may therefore be more demanding and require emulation instead of migration.
  - Interactive documents that contain embedded applications supporting the required functionality may require a combination of migration and emulation methods for preservation; the former for intellectual content, the latter for software components.
- Legal issues [BL]
  - It may be illegal to remove DRM, de-obfuscate embedded fonts, or to migrate the document to some other e-book format.
- Interactivity and animations
  - With EPUB 3, there are two possibilities to realize built-in animations and interactive features. One is to use a CSS construct for transformations; another, more versatile approach is to use embedded JavaScript, Adobe Flash, or other software components that may enable complex interactive behaviour [Bläsi, p. 32]. Although EPUB 3 allows the use of JavaScript, it does not standardize the use of JavaScript elements in e-books. This can easily lead to proprietary extensions as well as incompatible EPUB 3 reading systems that support a different or incompatible subset of scripting elements [Bläsi, p. 17].
  - Different e-book formats support interactivity in different ways, and apart from EPUB, features may be undocumented. Therefore migrating interactivity features between e-book formats is either difficult or impossible.
- Non-archivable core media types
  - Depending on the chosen preservation strategy, some current or future core media types may be regarded as unsuitable for digital preservation. For instance, GIF is not an archivable format according to the requirements of the Finnish National Digital Library initiative. [File formats, p.25].
- Non-archivable foreign resources
  - Foreign content may be both non-archivable and unsupported by EPUB reading systems the archive is able to use.

  - For the time being, there are no video codecs among the core media types. There is a recommendation that reading systems should support either H.264 or VP8. Neither of these are archivable or even ingestible formats in the Finnish National Digital Library specification, which approves JPEG 2000 sequence and MPEG-4 AVC as archive formats and DV (Digital Video), MPEG-1, MPEG-2, and WMV (Windows Media Video) as ingestible formats.
- External references [BL]
  - Externally referenced content (metadata, core media types, or foreign resources) SHALL be retrieved during pre-ingest and embedded into the SIP, during ingest and embedded into the AIP. If retrieval fails, the AIP is incomplete. If the submission agreement allows such policy, the archive can store the incomplete AIP and try to retrieve the missing content post-ingest. If the second attempt is successful, the AIP is ingested again into the repository system, and the missing content is added.
- Missing or poor fallback documents
  - If a foreign resource cannot be rendered, there SHOULD be a core media type fallback document. However, even if a fallback resource is present it may not produce the same rendition than the original resource and there is no guarantee either that the original semantics will be preserved.

Recommendations:

- EPUB 3 covers the superset of the expressive abilities of all the other e-book formats. Therefore there is no technical or functional reason not to use and establish EPUB 3 as an interoperable open e-book format standard [Bläsi, p. 8]. Having a universally supported e-book format would benefit current e-book users and make long-term preservation of e-books easier.
- Readium[41] project is developing a robust and efficient reader for EPUB publications. Such tools will make it easier to use rich EPUB documents, and EPUB community should continue investments on Readium and similar initiatives.
- The EPUB community should create EPUB/A, a subset of EPUB 3 with features suitable for long-term preservation. The specification should be complemented by an explanation why the EPUB 3 features not included in the EPUB/A format may jeopardize digital preservation, and a justification for those featureas that are required.
- When new EPUB core media types are added, the archivability of these file formats should be taken into account. EPUB community could co-operate with the digital preservation community to achieve this goal.
- Legal aspects of long-term preservation of EPUB 3 documents should be investigated.
- Open source licenses such as SIL Open Font License[42] should be used when possible.
- Foreign resources should be used with caution, until the archivability of the utilized file formats has been verified.
- 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.

[41] http://readium.org/

[42] http://scripts.sil.org/OFL_web

Bibliography

[1] BLÄSI, Christoph and Franz Rothlauf: *On the interoperability of eBook formats*. [online]. Brussels: European and International Booksellers Federation, 2013. Available from: http://wi.bwl.uni-mainz.de/publikationen/InteroperabilityReportGutenbergfinal07052013.pdf [viewed 2017-06-12].

[2] DALY, Liza: *EPUB 3 and interactivity.* [online]. EPUBZone, 2014. Available from: http://epubzone.org/news/epub-3-and-interactivity [viewed 2017-06-21].

[3] WHIBLEY, Simon: *EPUB format preservation assessment*. Version 1.2. [online]. British Library, 2015. Available from: http://wiki.dpconline.org/images/a/a9/EPUB_Assessment_v1.2.pdf [viewed 2018-03-14].

[4] *Digital Preservation Handbook*. 2nd, revised ed. [online]. Digital Preservation Coalition, 2017. Available from: http://www.dpconline.org/handbook. [viewed 2017-07-25].

[5] E-ARK: *Common specification for information packages. Version 1.0* [online]. E-ARK Project, 2016. Available from: http://www.dasboard.eu/specifications/common-specification. [viewed 2017-08-23].

[6] *EPUB Publications 3.0.1.* International Digital Publishing Forum, 2014 [online]. Available from: http://www.idpf.org/epub/301/spec/epub-publications.html [viewed 2018-10-30].

[7] *File formats. Version 1.6.1.* [online]. The National Digital Library, 2017. Available from: http://digitalpreservation.fi/files/File-Formats-1.6.1-en.pdf. [viewed 2018-09-10].

[8] ISO 14721. *Space data and information transfer systems – Open archival information system (OAIS) – Reference model.* [online]. ISO, 2012. Available from: https://www.iso.org/standard/57284.html [paywall]. [viewed 2017-07-10]. Also available as CSSDS specification from: https://public.ccsds.org/pubs/650x0m2.pdf [viewed 2017-07-17].

[9] ISO/FDIS 20614. *Information and documentation – Data exchange protocol for interoperability and preservation*. [online]. ISO, 2017. The standard, when completed, will be available from: https://www.iso.org/standard/68562.html. [viewed 2017-07-17].

[10] ISO 20652. *Space data and information systems - Producer-archive interface - Methodology abstract standard [PAIMAS].* [online]. ISO, 2006. Available from: https://www.iso.org/standard/39577.html. [paywall]. [viewed 2017-07-01]. Also available as CCSDS 651.0-M-1:2004 from: https://public.ccsds.org/publications/archive/651x0m1.pdf [viewed 2017-06-22].

[11] LAVOIE, Brian: *Meeting the challenges of digital preservation: The OAIS reference model*. [online]. Dublin, OH: OCLC Research, 2000. Available from: http://www.oclc.org/research/publications/library/2000/lavoie-oais.html [viewed 2017-07-17].

[12] *PDF/A* [online]. Wikipedia, 2017-05.17. Available from: https://en.wikipedia.org/wiki/PDF/A [viewed 2017-07-18]. Archived version available from: https://web.archive.org/web/20170713180152/https://en.wikipedia.org/wiki/PDF/A

[13] *TI/A Standard Initiative homepage*. [online]. TI/A Standard Initiative, 2016. Available from: http://ti-a.org/ . [viewed 2017-07-18].

[14] W3C. *EPUB 3 Community Group Charter*. [online]. W3C, 2017. Available from: https://www.w3.org/2017/02/EPUB3CGcharter. [viewed 2017-07-24].