Developing an Archiving Strategy: What formats to archive?

Archiving something is clearly better than not archiving at all, so a starting point is to archive what you have!

But given the opportunity to develop or select between alternatives, the project team developed "Good, better, best" guidance around alternative publication formats for archiving purposes, which is summarised below.

Barnes, M., Cole, G., Fry, J., Gatti, R., & Higman, R. (2023). 'Good, Better, Best': Practices in Archiving & Preserving Open Access Monographs (1.0). Zenodo.
https://doi.org/10.5281/zenodo.7876048 .

There are a number of useful and more detailed reports on file formats and their suitability for archiving available, including:

> DPC Digital Preservation Handbook: File formats and standards
>
> Library of Congress Recommended Formats Statement 2022-2023

The primary issues to consider in assessing the suitability of a specific format for archiving are:

- adoption: the extent to which use of a format is widespread
- technological dependencies: whether a format depends on other technologies
- disclosure: whether file format specifications are in the public domain
- metadata support: whether metadata is provided with the format

Existing formats that can satisfy all these criteria well are PDF, EPUB, HTML and XML - although precisely how the publication is structured within these standards matters. Formats that are proprietary or niche are unlikely to be good candidates for long-term preservation.

Overall summary

- PDF = stable, fixed, access-friendly but not so good for embedded content
- EPUB = flexible, containerised, good if self-contained

- HTML = web-native, archivable at scale but context-dependent
- XML = best for long-term preservation and reuse (if well formatted), but not for direct reading

Together, these formats serve complementary roles, with XML/EPUB supporting preservation and reuse, and PDF/HTML supporting access and dissemination. It should also be noted that it is possible to include multiple formats together in a single folder as part of the archiving process - and many of the archiving platforms allow for this.

PDF

The most commonly used format presently used for both the publication and preservation of eBooks. PDF is now an open standard, and the broad adoption of the format and the sheer number of pdf documents in existence means that accessibility of future systems to pdf content seems very likely.

The primary characteristic of PDF is that it displays content formatted as if on a printed page - thus it is particularly valuable where that format is intrinsically important to the work itself (such as in poems, or when lines are referenced).

Ideally the PDF should be well formatted and structured with searchable text, embedded fonts, content tagging, alt-text and good metadata - as generated, for example, for compliance with accessibility standards and embodied in the PDF/UA specification.

PDF offers options for embedding multi-media content - but the difficulty is that preservation software will not pick up the existence of that media.

The PDF/A standard was created specifically for archiving and preservation - however this restricts external dependencies, and so is not ideal when these are important for the publication.

However badly formatted PDFs, lacking any of the above features, can also be generated. While they may display well enough today they will be less appropriate of successful for archiving purposes. The good news is that work undertaken to enhance accessibility of the publication will be valuable for archiving and preservation purposes as well.

EPUB

The EPUB format consists of XHTML files that carry the content, packaged in an archive file along with any additional images and supporting files. The container file (based on the ZIP format) is able to include separate files for embedded content - which facilitates the migration of the content over time.

The difficulties with EPUB are that they don’t maintain the formatting information in the same way as PDFs do - if that is important for the publication. Utilising the full features of the EPUB for archiving purposes can also generate a very large file size, not suitable for easy transmission as an ebook - so some publishers generate separate EPUBs for distribution and archiving purposes.

XML

XML is not technically a file format, but a language that can be used to define any number of specific formats, which are defined by an accompanying XML Schema Definition (XSD) and Document Type Definition (DTD). EPUB3 is one such XML format. Following a well defined standard (such as EPUB3 or TEI) is necessary for successful long-term preservation and later rendering. As with PDF, if XML files are created in nonstandard ways, this can jeopardise future usability and prevent proper rendering.

HTML

HTML, and XHTML, is a text-based markup languages widely used in websites and for the online rendition of publications. When combined with DOCTYPE declaration and presentation stylesheet(s) these can function well for preservation purposes.

Revision #1
Created 2026-04-28 15:10:34 UTC by Rupert Gatti
Updated 2026-04-30 16:10:29 UTC by Rupert Gatti