eBook Archiving: A Toolkit for Small Publishers

- Introduction
- What is a ‘simple’ ebook?
- Developing an Archiving Strategy: Why archive?
- Developing an Archiving Strategy: What content to archive?
- Developing an Archiving Strategy: What formats to archive?
- Developing an Archiving Strategy: What metadata to archive?
- Developing an Archiving Strategy: Archiving outbound links?
- Developing an Archiving Strategy: Where to Archive content?
- A very simple open archiving strategy for a small publisher

Introduction

This toolkit is intended to provide an overview and guidance for the archiving and preservation of ‘simple’ open access ebooks, targeted primarily at small publishers. As will become clear, archiving and preservation should not be considered the role of the publisher alone - and considerations of archiving needs to be taken by all actors, starting with the author. We believe that the advice and recommendations contained in this toolkit, and the underlying research, is important for authors and all participants in scholarly research and communication - but we focus on the roles and actions publishers can take.

In the following sections we will consider what we mean by a 'simple' ebook, and considerations for publishers in developing an archiving strategy for a publication - noting that it may be that different ebooks need different strategies.

But the primary recommendations for publishers in creating an archiving strategy are:

1. Identify the content that is important to archive. See the why archive and what to archive sections.
2. Archive standardised and widely adopted formats for all content (including metadata) - this will facilitate future format migration. See the what formats section.
3. Use multiple complementary archiving solutions - don’t rely on a single solution. See the where to archive section.
4. Adopt permanent identifiers and broadly adopted standardised terminologies within metadata and avoid bespoke terms and specifications. See the metadata section.
5. Describe the structure and nature of the content and links within the publication to allow layouts, fonts, descriptions of images and embedded content and links to be reconstructed if necessary. Applying accessibility standards to publications will help with this, and some file formats are better for this than others. See the Outgoing links section.

What is a ‘simple’ ebook?

Defining an ebook is not easy - and indeed an important part of the COPIM and OBF projects has been to push the boundaries of any such definition through the Experimental Publications work packages (see the Experimental Publishing pages within the Copim Compass).

For the sake of this toolkit we consider content that might reasonably be expected to be published and distributed as an ebook in PDF or EPUB format. Adapting slightly the ‘simple’ ebook structure developed by Gareth Stewart in his report on link rot (Stewart, G. (2026). Think, before you link: Link Rot, eBooks, and digital preservation. Zenodo. https://doi.org/10.5281/zenodo.19912435), we identify four main components of such a publication -

1. the core text - primarily text and equations, including their layouts such as in tables and poems;
2. embedded (non-textual) material - such as images, audio and video files;
3. outbound links to external material - which may include additional resources or datasets associated with the publication, as well as citations or other third party content that may be important to the understanding of the work;
4. the metadata associated with the work.

Figure 1 provides a representation of the main components of the type of ‘simple’ ebook publication we consider.

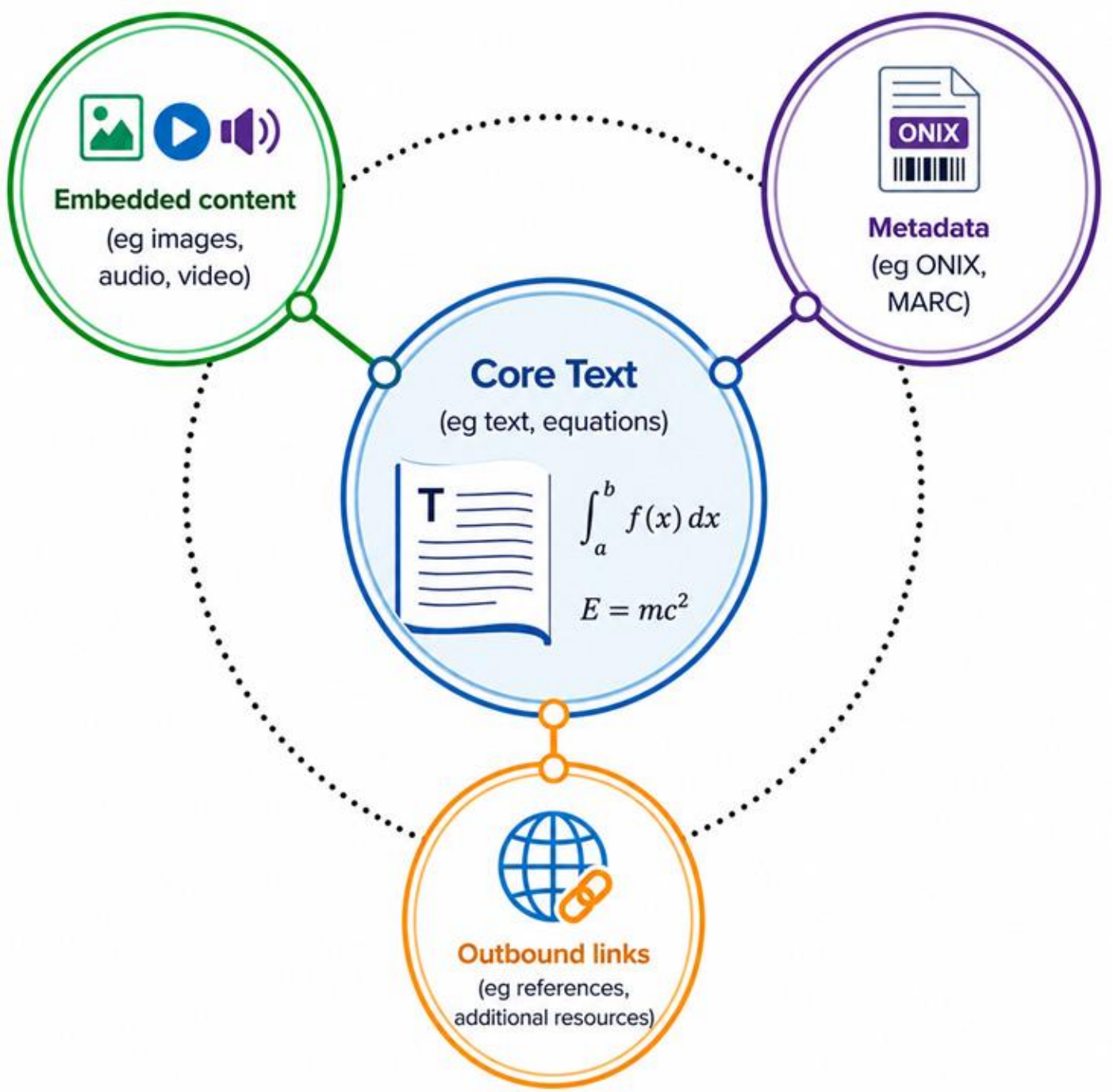

Figure 1: 'Simple' eBook - main components.

As Stewart notes: “The word ‘simple’ is used very loosely here(!), as evidently many layers of complexity could permeate in any of this model’s components. For example, metadata may be hosted externally in a separate system and linked to the eBook, embedded AV may be playable on an external web platform etc. There are many, many unusual combinations of doing things on the Web, which would be impossible to advise on in every instance.”

Furthermore, not all published formats of a single work will necessarily contain precisely the same content - the quantity and format of embedded material in the PDF and the EPUB editions of a work may vary, as may any embedded metadata.

It should also be noted that the content identified as important for archiving may differ from that included in any specific format used to publish the work. While it seems likely that the Core Text should be included in any archiving strategy, the other components may be considered more or less important for the understanding of the work and so be of greater or lesser importance for archiving purposes. Consequently, an effective archiving strategy is likely to differ from a strategy of archiving a single format of the published work.

Developing an Archiving Strategy: Why archive?

The first question for a publisher (or author) to address in developing an archiving strategy is: why archive the work at all?

One primary motivation is to enable a reader at some time in the future to be able to access the content of the publication.

What do we know about a future reader? If we consider a reader (say) 100 years from now, it seems reasonable to assume that:

1. the reader will be using a completely new technology that is not compatible with any of the (now redundant) formats used in today's publication
2. that every single url/link to content outside the digital ebook is broken (linkrot).

And so that future reader will need:

- a mechanism to discover that the work exists at all, and
- a way to both access the content and understand the author's work.

So we can state that **a primary archiving and preservation objective** for a publisher is:

> **To provide a future reader with ways to discover and engage with the publication, when the formats used in the publication are incompatible with the future reader's technology and all external links within the publication are broken.**

Solutions will necessarily require strategies to ensure:

1. the discovery of the work's existence - requiring the transition of book metadata over time,
2. the discovery of the publication files themselves - requiring the transition of the book files over time
3. access to file content by the new technologies - requiring format migration, and
4. the reconstruction of both the links to external content and access to (the archived version of) that content.

Of course one strategy is to do nothing now and assume that the future managers of the publisher will take full responsibility for all these actions when the time comes. But most small publishers recognise that neither the existence or the financial health of future managers is guaranteed! Consequently, any successful preservation strategy will need to rely on future (potentially

unknown) third parties undertaking some of this work for them. The important question for a small publisher then becomes:

What can I do today to make the job of reconstructing this publication for future readers by unknown third-party agents as easy as possible?

In the next sections we look at some of the main issues to consider for a publisher in developing such a strategy.

Developing an Archiving Strategy: What content to archive?

It seems unlikely that it is either necessary or feasible to archive every aspect of a publication. The first step is to consider how important the various components of the publication are to the understanding of the work by a future reader, and prioritise the most important components. We recommend four prioritised categories.

Category 1. Original content created for and core to the publication that is not being archived elsewhere.
This is most likely to include the core text written by the author as well as content such as images, audio/visual and other content which are embedded or linked to within the work. It may also include specific layouts for the content, such as in poetry or texts with line referencing. Metadata for the publication should also be included in this category.
This is the content that it will be most important to develop archiving solutions for.

Category 2. Non-original core content (such as third party images) that is very important to the understanding of the work, are embedded or linked to within the work, but for which there is no alternative archiving solution available. This, by definition, is not original content created for this publication - but readers will need access to this content to understand the author's work and original contribution, so addressing ways to ensure this content is archived and remains accessible to future readers will clearly be very important.

Category 3. Content important for the understanding of the work, but for which it seems reasonable to assume that archiving solutions are already in place? This might include third party content embedded in the work, important citations or original datasets related to the work but being archived elsewhere. In this case providing permanent links to the archived versions of the content or sufficient information for the archived work to be discovered by a future reader will be the priority, rather than necessarily archiving the content directly.

Category 4. Nice to have, but inessential content. Content that is deemed less essential for future readers to access can be considered as the lowest priorty for archiving. Providing permanent links to archived versions of this content (if available) is clearly still desirable, as is directly archiving that content if it is feasible to do so - but this category of content does not need to drive the archiving strategy.

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.

Developing an Archiving Strategy: What metadata to archive?

Metadata is critical for the discovery and effective dissemination of books. Just putting something online doesn't mean anybody will get to know about it - and this is particularly true for future readers searching and accessing works in ways we have not yet imagined. While more metadata is always to be encouraged, it is important to use persistent identifiers and controlled vocabularies whenever possible as this will increase the likelihood of interoperability and the successful transmission of metadata to new systems in the future.

What metadata to archive?

The report *International Metadata Recommendations, and Platform-Specific Requirements for Open Access Books and Chapters* (Steiner et al. 2026) identifies metadata fields that are "Essential", and those that are "Desirable", for the effective dissemination of open access books. These criteria are appropriate for archiving purposes also.

Essential bibliographic and access metadata include:

- title and subtitle (multilingual if appropriate)
- contributors (including standardised and persistent identifiers such as ORCID or ISNI where possible)
- copyright holder and licence,
- subjects (utilising recognised schemas such as THEMA where possible)
- landing page and full-text URLs and/or DOIs at book and chapter level (ideally, for archiving purposes, a link/reference to an archived version should be included)
- publisher details, and publication date.

Desirable elements include:

- abstract (multilingual if appropriate)
- cover image
- table of contents
- contributor affiliations (using standardised and persistent identifiers such as ROR where possible)
- funder details (using standardised and persistent identifiers such as ROR where possible)

What formats to archive metadata?

Many file formats, such as PDF and EPUB, allow extended metadata to be included in the book file itself - and clearly the more metadata included this way the better.

However, we recommend that when archiving content a separate metadata file be included alongside the primary ebook file(s) in an open and standardised format that can be accessed as plain text if necessary (such as ONIX, MARC or JSON). This helps ensures that the metadata can be openly shared across systems and platforms and that engagement with specific software or formats is not required to access the metadata

This section summarises the findings of several reports created within the OBF project:

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 . This report considers the archiving of metadata specifically in Chapter 2.

Steiner, T., Arias, J., Bennett, M., Booth, E., Edmunds, J., Gatti, R., Higman, R., Hillen, H., Laakso, M., Nason, M., O'Connell, B., Pogačnik, A., Rabar, U., Ramalho, A., Stone, G., van Gerven Oei, V. W. J., & Wake Hyde, Z. (2026). International Metadata Recommendations, and Platform-Specific Requirements for Open Access Books and Chapters (1.0). Thoth Open Metadata.

https://doi.org/10.5281/zenodo.18173982. This report identifies the most important metadata fields to provide to ensure broad discovery and dissemination of the work.

Stone, Graham, Rupert Gatti, Vincent W. J. van Gerven Oei, Javier Arias, Tobias Steiner, and Eelco Ferwerda.
‘WP5 Scoping Report: Building an Open Dissemination System’. Community-Led Open Publication Infrastructures
for Monographs (COPIM), 21 April 2021. https://doi.org/10.21428/785a6451.939caeab.

Developing an Archiving Strategy: Archiving outbound links?

As part of the OBF project the Garth Stewart at the Digital Preservation Coalition undertook a report looking specifically at issues around the archiving and preservation of outgoing links and the content associated with those links.

Stewart, G. (2026). Think, before you link: Link Rot, eBooks, and digital preservation. Zenodo. https://doi.org/10.5281/zenodo.19912435

The report provides helpful information for both publishers and authors to help them realistically mitage against the reprecussions of link rot for their publications. Key recommendations for authors and publishers, as laid out in Steps 1 to 5 of Part B of the Report, are:

1. Think before you link, and determine acceptable loss

Link rot is inevitable, so authors and publishers must be selective about which links they include. Not all links are equally valuable—some are critical to understanding the work, while others are supplementary.

Stakeholders should assess whether a link “must survive” and define an acceptable level of loss. This involves prioritising essential content, avoiding unnecessary links, and considering the longevity and reliability of the source. Authors should also evaluate how broken links would affect user experience, trust, and the integrity of the work.

Where possible, critical external content should be embedded or packaged with the eBook. Ultimately, this step promotes a risk-based, intentional approach to linking, recognising that some loss is normal and manageable if planned for.

2. Cite to resilient links

Focuses on improving the durability of links that are included. Authors should use simple, stable, and transparent URLs, avoiding long, complex, or shortened links that are prone to failure. Preference should be given to trusted sources, landing pages rather than deep file paths, archived and open-access content.

Persistent identifiers (e.g. DOIs) are encouraged where available, though these should not be assumed to be permanent.

Citing archived versions of webpages (e.g. from web archives) is best practice for critical content, as these are more stable and preserve content in a fixed state. Clear and consistent citation also acts as “preservation metadata,” helping future users locate content even if links break.

3. Describe more, bolster accessibility

To mitigate the impact of link rot, authors should provide detailed descriptions of linked content, including its context, purpose, and structure. This can include summaries, transcripts, or screenshots, especially for multimedia resources. Even brief descriptions can preserve meaning if the link fails. This approach not only strengthens long-term resilience but also improves accessibility for users and machines. Documentation can be embedded in the text or stored as metadata.

While this requires additional effort, it is far less costly than attempting to recover lost content later.

Overall, richer description ensures that the intellectual value of the work remains understandable even when links disappear.

4. Archive your URLs

For important or critical links, stakeholders should proactively create archived copies using web archiving tools or services.

These snapshots (e.g. via services like “Save Page Now” or similar tools) provide stable, time-stamped versions of content that can be cited or stored alongside the eBook. Where possible, multiple copies or formats (e.g. WACZ/WARC files) should be created to increase redundancy. These files could also be archived alongside the content and metadata files in the folder used for archiving the publication itself.

These archived versions can either replace or supplement live links, reducing dependence on the original source. Although this step requires more effort and resources, it significantly increases long-term access and protects against both link rot and content drift.

5. Monitor your URLs, enact digital archaeology

After publication, links should be periodically checked to identify breakage. Publishers and repositories can use automated tools to monitor link health, particularly for online eBooks. For open access works it is also not unusual for readers to report broken links when these are observed. When links fail, “digital archaeology” involves searching web archives or other sources to recover or replace missing content. While not all links can be restored, this process can often recover valuable material. This step acknowledges that preservation is ongoing, not one-time, and requires maintenance over time. It is especially important for high-value links, helping sustain the usability,

credibility, and completeness of the eBook as it ages.

Developing an Archiving Strategy: Where to Archive content?

As part of the work undertaken within the OBF project we have identified many alternative archiving solutions and, in light of our recommendations for Open Archiving Criteria, conducted an analysis of some of the main alternatives. For five specific archives (CLOCKSS, Portico, Internet Archive, Zenodo and Figshare) we directly assessed their technical specifications and operations against the eight Open Archiving Criteria.

The table below, taken directly from that report, provides an overview of this work.

Overall, the main findings are:

- no single solution individually satisfied all eight open access archiving criteria identified;
- combinations of two or three different solutions collectively did. Strategically combining solutions and harnessing their different characteristics and structures is more effective than relying on a single solution;
- robust open archiving solutions that are both free and relatively easy to implement exist, and are available to even the smallest publishers.

Specifically, we found that combining two freely accessible open generalist repositories, the Internet Archive and Zenodo, will provide small publishers with a free-to-use and effective open archiving solution for thier publications and associated content, that is also built on open and non-profit infrastructures well aligned with the general COPIM principles.

We encourage publishers, when formulating their archiving strategy, to use the framework developed in the report and table below to assess the archiving alternatives available to them and how they can be combined effectively to create an open archiving solution that meets their own needs. For example, Thoth Open Metadata (itself an output of the COPIM/OBF projects) - as well as providing a free mechanism for publishers to create and output enhanced metadata in file formats conducive for archiving - has created an automated Open Archiving Network for publishers by uploading book content and metadata files to the Internet Archive, Zenodo and their own CDN.

Greater details of the analysis conducted can be found in the full report:
Steiner, T., Cole, G., Fry, J., Gatti, R., Higman, R., Stokes, P., & Turpin, H. (2026). *Applying Open Access and Open Data to the Archiving of Long-Form Scholarship: A Comparative Analysis of*

Existing Services Through the Lens of the Copim Open Archiving Criteria (1.0). Zenodo. https://doi.org/10.5281/zenodo.19882343

Table: Comparison of five archiving solutions with the Open Archiving Criteria.

Open Archiving Criterion	CLOCKSS	Portico	Internet Archive	Zenodo	Figshare
1) Openly accessible content (directly upon deposit)	No. CLOCKSS is a "dark archive" - content is generally not accessible to users unless a "trigger event" occurs, after which it is released under an Open Access license (Creative Commons or equivalent, selected by the publisher or the CLOCKSS Board). It seems noteworthy that CLOCKSS' "Triggered Content" section of released scholarly output currently only lists serials/journals - which seems to imply that no books have ever been released through a trigger event.	No. Portico is a "dark archive" that provides access to content only after a "trigger event". In case a trigger event is envoked, content is either released only to participating libraries, or made available open access (if the depositing publisher has indicated that to be their choice).	Yes. The core mission of the Internet Archive is "Universal Access to All Knowledge", and it accordingly provides free and immediate access to the vast majority of its collections.	Yes. Zenodo's core mission is to serve as an Open Science repository. While it allows for embargoed or restricted content, its goal is to make content public, with embargoes expiring automatically.	Yes. Figshare is an open-access repository that adheres to the principle of open data, with all publicly published content downloadable by anyone.

2) Openly accessible metadata	Partially. CLOCKSS publishes basic aggregate holdings metadata (e.g. titles, ISSNs) via open CSV/KBART lists, so the public can see what titles are being preserved. Extended content- and archiving-related metadata including relational descriptions (e.g. chapter-/book-level relations) is stored internally as part of the underlying LOCKSS software implementation, but these metadata sets are not available to the public.	Partially. Portico makes basic bibliographic holdings metadata openly available in several formats. Custom holdings comparisons are available to libraries on request so they can compare the coverage of their journal or book holdings to what is preserved in Portico. Portico generates custom reports for some community partners such as CHORUS.	Yes. Metadata for items and collections is usually stored in openly-available XML following Dublin Core, and can be output in formats like JSON, XML, or CSV.	Yes. Metadata is licensed under a CC0 dedication, exported via OAI-PMH, and can be harvested by third parties without restriction.	Yes. All metadata published on the Figshare platform is available under a CC0 dedication.
3) Openly verifiable processes: a) Publishing checksums to allow verification of content integrity	No, not publicly. According to CLOCKSS' documentation, the CLOCKSS system uses a "polling-and-repair mechanism" across its 12 nodes to continuously validate data integrity, but it does not publish checksums for public verification.	No. Portico maintains an internally-verifiable audit trail (not accessible to the public) and performs self-checks and third-party certifications, but it does not publish checksums for public verification.	Yes. Various checksums are recorded as part of each deposit's *_files.xml data file, which are made publicly available together with the user uploads.	Yes. Zenodo stores two MD5 checksums for every file (one stored in Invenio, one in EOS) and regularly checks files against these checksums to ensure consistency of archived content.	Yes. Figshare performs and displays MD5 integrity checks when files are uploaded to the platform, and its hosting provider (AWS) also performs regular data integrity checks.

3) Openly verifiable processes: b) Transparent version control (for both content and metadata)	No. CLOCKSS tracks and records all changes, including version updates and errata. New versions can be added to the archive, but content is never deleted.	No. According to Portico's documentation, an audit trail is maintained, keeping the original file and all related information if a transformation occurs. This information appears not to be available to the public, but can be accessed by the depositing publisher as well as designated auditors from the Portico network's participating libraries.	Yes. For user-uploaded items, a history of changes can be viewed by changing the URL from 'details' to 'history'. New versions of files can be uploaded and will be updated.	Yes. Zenodo supports file versioning. Records are not versioned, but changes to files will create a new version of a given deposit, together with a new DOI, to ensure the original version remains unchanged for citation purposes.	Yes. Figshare supports version control for both files and metadata, with previous versions displayed and accessible on each item's landing page.
3) Openly verifiable processes: c) Clear mechanisms for checking and maintaining the content	Partially. CLOCKSS claims to have a unique "polling-and-repair mechanism" by which its 12 peer systems continuously validate the integrity of their shared data - but this can not be checked by the public	Partially. Portico claims to conduct regular fixity and integrity self-checks and undergoes independent third-party audits and certifications to guarantee quality and security. These mechanisms are not publicly accessible, though.	Partially. The Internet Archive duplicates/backs up all files at various locations. Its internal storage system, Petabox, is also mentioned. It does not seem to be openly verifiable, though. States that verification checks are undertaken periodically - but not clear how often of when.	Yes. Files are regularly checked against their MD5 checksums to ensure content constancy, and backups are performed nightly. Zenodo also performs file format checks.	Partially. Figshare relies on its hosting provider, AWS S3, which performs regular data integrity checks. Nightly backups of data files and metadata are also performed.

4) Adherence to Accepted Good Practice in Digital Archiving Operations: a) Satisfies industry standards, e.g. CRL TRAC audit, ISO:16363, the Core Trust Seal, or the DPC's Rapid Assessment Model (RAM)), which signal a commitment to and expertise in long-term preservation	Yes. Satisfies criteria for membership of Keepers Registry. CRL TRAC-audited (2018). Certified CoreTrustSeal repository.	Yes. Satisfies criteria for membership of Keepers Registry. CRL TRAC-audited (2010). Alignment with OAIS (ISO 14721)	Yes. Satisfies criteria for membership of Keepers Registry. Not formally certified against ISO 16363, TRAC, or CoreTrustSeal, but its operational model reflects many of the core principles of trusted digital repositories.	Yes. Certified CoreTrustSeal repository. Aligned with OAIS (ISO 14721). Meets core expectations for fixity, authenticity, and traceability in archival standards.	Partially. Compliance with OSTP and NIH "Desirable Characteristics for Data Repositories". While Figshare's hosting provider, Amazon Web Services, itself is not certified against ISO 16363, TRAC, or CoreTrustSeal, it delivers the core technical controls required for bit-level preservation and secure archival storage.
4) Adherence to Accepted Good Practice in Digital Archiving Operations: b) Institutional reliability and long-term sustainability	Yes. It is a financially secure 501(c)(3) non-profit with a diversified funding stream from hundreds of publishers and libraries.	Yes. Portico, a non-profit service of ITHAKA, has a diversified funding stream from ca. 1,300 libraries and 1,300 publishers, with financial contributions roughly split 50:50 between both stakeholder groups. It also conducts annual financial audits.	Partial. As a 501(c)(3) non-profit, its sustainability is tied to individual donations and grants from foundations, and it has an intention to store materials in perpetuity. Facing significant legal challenges which may be problematic for long-term sustainability.	Yes. Zenodo's long-term viability is tied to its host institution, CERN, which has a projected experimental program for at least the next 20 years.	Yes. Figshare is a for-profit company that provides a 10-year service-level agreement (SLA) guaranteeing persistent availability.

4) Adherence to Accepted Good Practice in Digital Archiving Operations: c) Succession plannning	Yes. CLOCKSS has a dedicated Trustee Committe that defined a Succession Plan. This includes the formation of a 4-library network that would continue to preserve the existing CLOCKSS content by running four LOCKSS nodes; the four libraries are Stanford University (USA)., University of Alberta (Canada), University of Edinburgh (UK), and Humboldt University, Berlin (GER).	Partial. Portico has a dedicate Succession policy, in which it outlines that the organisation will endeavor to find a successor non-profit organization, should it ever cease to operate. No actual organisation appears to have been identified.	Partial. The Internet Archive has two independent branches in Canada and Europe that also mirror the main IA repository content. In case the US-based institution should cease to exist, any of the other two branches may carry forward the IA's operations. No specific information on institutional succession planning could be found from the documentation. The Internet Archive's approach might thus be described as implicit and infrastructural, not procedural.	Partial. Zenodo's policies state that in case of closure of the repository, best efforts will be made to integrate all content into suitable alternative institutional and/or subject based repositories.	Unclear. Bound by its host company Digital Science, a subsidiary of Holtzbrinck Publishing Group.

4) Adherence to Accepted Good Practice in Digital Archiving Operations: d) Multiple geographically-redundant copies	Yes. CLOCKSS operates 12-node LOCKSS repository network at academic institutions worldwide, spread across 4 continents. 2 Australian National University (Australia), Humboldt University-Berlin (Germany), Indiana University (USA, Indiana), National Institute of Informatics (Japan), OCLC Online Computer Library Center (USA,Ohio), Rice University (USA, Texas), Stanford University x 2 (USA, California), Università Cattolica del Sacro Cuore (Italy), University of Alberta (Canada), University of Edinburgh – EDINA (UK), University of Virginia (USA, Virginia) - Total of four continents	Yes. A master copy containing all archival packages is kept in Princeton, NJ (USA) and is maintained using an Oracle database. All archival packages are replicated to a file system in the Texas Advanced Computing Center (TACC) as part of a partnership with Texas Digital Library (TDL). Publication content has a second online replica housed on a dedicated server in the National Library of the Netherlands. Non-publication content (e.g. D-Collections, Preserved Collections) has a second online replica located in an Amazon Web Services (AWS) Glacier repository. A separate complete copy of the original supplied files (pre-processing) are also archived in AWS Glacier.	Yes. The Internet Archive has six primary data centers in three countries, including a full, second live copy in a Canadian data center as a backup outside the US, incl in the EU. It stores at least two copies of everything.	Yes. All data is stored in CERN Data Centres, with separate replicas stored in Geneva and Budapest.	Yes. Figshare is hosted on Amazon Web Services (AWS) S3, which is designed to sustain the concurrent loss of data in two facilities and offers cross-region replication.

5) Support for retrieving and archiving associated content: a) “Additional materials” provided to supplement the main content	Yes. CLOCKSS preserves "supplementary materials," including datasets, multimedia, and additional documentation.	Yes. Portico preserves e-journals, e-books, and "D-Collections" (digitized historical collections), as well as audio and video content. Any file format is accepted as supplement.	Yes. IA archives a vast range of content types beyond scholarly papers, including music, TV news, software, and images.	Yes. Zenodo accepts a wide variety of research artifacts, including text, spreadsheets, audio, video, and images.	Yes. Figshare is built to host "non-traditional research outputs" such as figures, datasets, media, papers, and code. It also hosts supplementary material for publishers.
5) Support for retrieving and archiving associated content: b) Web pages represented by URLs within the main content	No/Unclear. While CLOCKSS uses web harvesters / crawlers to programmatically discover and collect content from websites based on static URIs, it is not clear from the documentation if retrieval and subsequent archiving of associated content would be processed.	No. Portico's documentation does not explicitly mention a policy or process for archiving associated content.	Yes. The Internet Archive's Wayback Machine is specifically designed to archive web pages and their associated data such as Outlinks. Its Archive-It program allows for targeted web archiving.	No. Zenodo's documentation does not explicitly mention a policy or process for archiving associated content.	No. Figshare's documentation does not explicitly mention a policy or process for archiving associated content.
6) Clearly-stated policies around removal of content	Yes. Content is not deleted from the archive. Corrected or retracted versions can be added to it, maintaining a permanent record.	Yes. Removal Content is held in perpetuity and released only in the event of a "trigger event," not removed. Portico has a clearly-described Content Modification and Deletion policy detailing removal processes.	Yes. Content can be removed, e.g. if copyright infringement has been claimed, or removal is requested by a website owner.	Yes. Content may be removed for reasons including spam, copyright infringement, scientific misconduct, and transfer to another repository.	Yes. Figshare maintains the right to remove data that violates its Terms of Acceptable Use.

7) Collation of usage statistics	Partially. A dark archive, only 'triggered' content is accessible. The privacy policy mentions collecting "Aggregated Data" and "Usage Data" from its website once triggered, but there is no public-facing mechanism for tracking or reporting content-level usage statistics.	No. As a dark archive, Portico has reported very low usage overall. Usage statistics reports can be generated manually upon request for participating libraries.	Yes. The Internet Archive tracks and shares "views" and "downloads" for items and collections through a public API. No mentions of COUNTER or Make Data Count standards	Yes. Zenodo tracks and shares usage statistics, including visits and downloads, via a public API and is compliant with COUNTER and Make Data Count standards.	Yes. Figshare tracks and displays views, downloads, citations, and Altmetrics for hosted materials. It is compliant with Make Data Count and COUNTER standards.
8) Independence from private or government-controlled entities: **a) Governance**	Yes. CLOCKSS is a non-profit 501(c)(3) organisation. It is governed by a board with equal representation from libraries and publishers and answers to its community. The CLOCKSS Board of Directors comprises representatives from large commercial publishers incl. Elsevier, Wiley, and Wolters Kluver, as well as aggregators such as OCLC - all of who may have an influence on the organisation's overall direction.	Partially. Portico is a service of the non-profit organization ITHAKA. It operates through a community model with guidance from the participating library and publishing communities. It has its own dedicated Advisory Committee, but is also overseen by ITHAKA's Board of Trustees. Ultimately, Portico's decision-making may thus also be influenced by ITHAKA's strategic objectives.	Yes. The Internet Archive is a 501(c)(3) non-profit, not controlled by a private or government entity. Its management board comprises librarians and open archiving advocates.	Partially. Zenodo is a service of CERN, an intergovernmental organization. Governing board made up of CERN staff-members.	No. Figshare is a for-profit company and a portfolio business of Digital Science, a subsidiary of Holtzbrinck Publishing Group.

8) Independence from private or government-controlled entities: b) Legal jurisdiction.	Partially. Registered in the US and subject to US legal jurisdiction, but partner repositories hosted by universities based in multiple jurisdictions.	Partially. Registered in the US and subject to US legal jurisdiction, but partner repositories hosted by universities and national libraries based in multiple jurisdictions.	Partially. Registered in the US and subject to US legal jurisdiction, with fallback options established in Canada and Europe.	Yes. As an intergovernmental organization CERN enjoys certain privileges and immunities, including e.g. immunity from jurisdiction of the national courts to ensure independence from individual Member States.	No. Holtzbrinck Publishing Group is registered in Germany.
8) Independence from private or government-controlled entities: c) Technology	Partially. Architecture based on open source LOCKSS system. Independent repositories hosting content.	Partially. Portico's software stack can be hosted on and moved to different platforms if needed, but is dependend on a mix of open-source and proprietary software (e.g. Oracle) to support it. Content hosted using multiple alternative solutions, including Oracle, AWS, and a dedicated server provided by the KB (Koninklijke Bibliotheek, National Library of the Netherlands).	No. Bespoke system (unclear if fully open source?)	Partially. Invenio RDM (CERN-developed, open-source, can be self-hosted)	No. Built on Amazon Web Service

A very simple open archiving strategy for a small publisher

1. Create an information pack for authors - help them to identify the material most critical for archiving, and encourage them to use PIDs, robust links, and to link to archived content
2. Create a folder to hold all the material associated with the publications that you wish to upload to an archive
3. Include within the folder
    1. the highest quality versions of the ebook publication you are able to generate (see the what formats section above)
    2. a file with as complete metadata for the publication as you have, in a format that can be easily read (see what metadata section)
    3. any additional or embedded resources it is important to make available to reader
4. Create web archives of important links within the work using a suitable web archiving service, and make clear within the publication where these can be found.
5. Upload the folder to a couple of general archiving platforms, ideally based in different countries and using different technologies (eg Internet Archive and Zenodo).
6. Reconsider and enhance each step over time as you can - archiving is an ongoing process!