

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 their 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 invoked, 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/backups 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 planning	Yes. CLOCKSS has a dedicated Trustee Committee 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 dedicated 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 URLs, 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 Kluwer, 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

Revision #1

Created 2026-04-28 15:13:52 UTC by Rupert Gatti

Updated 2026-04-30 16:10:58 UTC by Rupert Gatti