

‘GOOD, BETTER, BEST’: PRACTICES IN ARCHIVING & PRESERVING OPEN ACCESS MONOGRAPHS



Photo by Bilakis from Pexels (Pexels License)



Research
England

30TH APRIL 2023 • The COPIM Project

Authored by: Miranda Barnes, Gareth Cole, Jenny Fry, Rupert Gatti, and Ross Higman, with contributions by Graham Stone (Jisc), Paul Wheatley (DPC) and Ilkay Holt (British Library)

<https://doi.org/10.5281/zenodo.7876048>

Table of Contents

CHAPTER 1: BACKGROUND AND SUMMARY OF METHODS & APPROACHES	4
Background	4
Methods & Approaches	11
CHAPTER 2: ‘GOOD, BETTER, BEST’: PRACTICE GUIDANCE FOR OA MONOGRAPH PUBLISHERS	13
A basic guidebook for the small and scholar-led press.....	13
Introduction	13
File Formats.....	14
Metadata.....	28
Selecting Content: What is essential?.....	34
Existing routes to digital preservation archives	39
Archiving & Preservation Workflows	48
Copyright, Reuse & Licensing.....	50
CHAPTER 3: CASE STUDY 1 – MANUAL INGESTION.....	60
Experimenting with repository workflows for archiving	60
CHAPTER 4: CASE STUDY 2 – AUTOMATED INGESTION	71
Options for computer-assisted repository archiving for small and scholar-led presses publishing open access monographs	71
CHAPTER 5: DEVELOPMENT OF DISSEMINATION TOOL IN THOTH	79
Case studies of content deposit in the Internet Archive and Figshare By Ross Higman	79
CHAPTER 6: ENHANCED VS. COMPLEX DIGITAL MONOGRAPHS: IMPLICATIONS FOR ARCHIVING & PRESERVATION.....	92
Considerations for complex & experimental monographs.....	92
CHAPTER 7: A BRIEF INTRODUCTION TO THE THOTH ARCHIVING NETWORK	100
A community solution for open access monograph archiving.....	100
CHAPTER 8: LOOKING AHEAD TO COPIM’S OPEN BOOK FUTURES.....	107

A new grant to significantly expand and accelerate COPIM’s open access infrastructures	107
Archiving and Preservation within Open Book Futures	109
 APPENDIX I: TOOLS AND RESOURCES	 111
Open Access Books Toolkits & Guides	111
Toolkits.....	111
Guides	111
Copyright, Reuse Licenses & Third-Party Content	114
Publishing.....	114
 Digital Preservation	 115
Digital Preservation Archives	115
Digital Preservation Software	116
Digital Preservation Guides.....	117
 APPENDIX II: GLOSSARY	 119
 REFERENCES	 126

Chapter 1:

Background and Summary of Methods & Approaches

Background

[COPIM](#) (Community-led Open Publication Infrastructures for Monographs) is an international partnership of researchers, universities ([Coventry University](#); [Birkbeck, University of London](#); [Lancaster University](#); and [Trinity College, Cambridge](#)), established open access the [ScholarLed consortium](#), which includes [Mattering Press](#), [meson press](#), [Open Humanities Press](#), [Open Book Publishers](#) and [punctum books](#)), libraries ([UCSB Library](#) and [Loughborough University Library](#)) and infrastructure providers (the [Directory of Open Access Books](#) and [Jisc](#)).

COPIM is also collaborating closely with institutions such as the [British Library](#) and the [Digital Preservation Coalition](#), and with the [Next Generation Library Publishing](#) project, in addition to consortium members. As well, a broad spectrum of academics, publishers, librarians, software developers, funders and others contribute as part of the working groups, events and projects that COPIM is setting up and running. COPIM's funders are the [Research England Development \(RED\) Fund](#), and [Arcadia](#) — a charitable fund of Lisbet Rausing and Peter Baldwin.

The Project is dedicated to investigating the difficulties that impede the progress of small publishers interfacing with large-scale organisations and processes. Through the work of this project, the consortium is in the process of developing a significantly enriched, not-for-profit and open-source ecosystem for open access (OA) book publishing, supporting, and sustaining a diversity of publishing initiatives and models, particularly within Humanities and Social Sciences (HSS) publishing.

Work Package 7 is tasked with exploring potential archiving and preservation solutions for the small and scholar-led publishers of open access monographs. We

first determined the lay of the land for these presses with our [Scoping Report](#) (2022), as well as the workshops and interviews that fed into the report's publication. Though key findings made clear that no single solution would be possible, and multiple approaches will be necessary, it was also evident that existing pathways to preservation would not serve all publishers and presses equally.

There is a straightforward route to digital preservation for large and medium-sized presses who are able to subscribe to the existing services of large-scale digital preservation archives, such as CLOCKSS and Portico. This relationship offers smooth solutions both in terms of the technical challenges of preservation and the longer-term question of policy impact. However, not every publisher has the resources, financially, technically, or in terms of staff, to initiate and maintain a subscriber relationship with a digital preservation archive. This means the substantial contribution to knowledge produced by these presses is at risk from disappearing entirely.

We know from our own discussions with small and scholar-led presses, as well as from the Jisc landscape study report from 2017 (*Changing ecologies: a landscape study of new university presses and academic-led publishing*, Adema & Stone), that there is a persistent lack of consistent preservation practice among this subsection of OA monograph presses. The survey results for academic-led presses (ALPs) indicated that “[m]ost ... do not have a systematic preservation strategy.” The ALPs’ existing solutions included hard drives, servers, or cloud storage, though several deposited into a digital preservation archive via a third party, such as DOAB or OAPEN.

New university presses (NUPs) were also asked a series of questions around their operations, including preservation. While most indicated they had some kind of preservation solution in place, there was ambiguity surrounding certain answers, and at least one NUP stated they had no existing preservation policy. University presses tend to have the support of their parent institution and the subsequent existing infrastructure, but not every university press benefits equally. While this

landscape study survey took place a few years in the past, it is still recent, and these findings indicate that the challenge of preservation is not limited to the scholar-led or independent press. (In fact, digital preservation more broadly remains a stubborn challenge for smaller institutional libraries: “Despite a growing number of resources that support digital preservation work, among current best practices it is difficult to find scalable workflows for institutions with limited staff and funds.” (Velte & Winkle, 2020)) For many respondents, the survey questions were what first alerted them to the issue of preservation. Respondents also indicated they would ideally want a “Shared service for preservation”, rather than each press having to solve this issue in isolation. (Adema & Stone, 2017, p. 39)

We also know from the *UKRI gap analysis of open monograph infrastructure* (2021) that there is an “ambiguity concerning who is responsible for the preservation of OA books.” (Ferwarda, Mosterd, Snijder & Mounier, p. 7) This echoes feedback received in the above landscape study, in which respondents expressed a desire for “a centralized place [for] archiving all the open access content across a multiplicity of publishers.” (Joy in Adema & Stone, p. 72) The UKRI’s gap analysis report indicates that a recommended action in this area would be to “Develop [an] approach to preservation of OA books in liaison with UK legal deposit libraries and international partners.” (Ferwerda, et al, p. 7)

There is a precedent for national library involvement in the preservation and open archiving of open access scholarly content. In the United States, the Library of Congress initiated a pilot project involving their Digital Content Management Section (DCM) staff in collaboration with the LoC Collection Development Office (CDO). ([More Open eBooks: Routinizing Open Access eBook Workflows](#), Gonzalez-Fernandez, 2020). Using a selection process in which subject matter experts determined which works were in scope based on the collection policy statements, the teams identified matches in DOAB to print holdings of the Library of Congress. These were then downloaded from DOAB and processed for dissemination on [loc.gov](#), providing full and open access, as well as long-term preservation. The metadata records for the eBooks were cloned from original records and enhanced in a new record to reflect their open access, and digital, status.

The Library of Congress is working to refine and improve this process, but the workflows that have been created, codified, and documented during the pilot project are now being used to support a new routine for the processing of eBooks, in the DOAB and beyond. Though still in these initial stages, the process is a beneficial case study of how such a project might be undertaken.

While there is work to be done globally to support open access monograph publishers, the initial focus of COPIM has necessarily been primarily the UK and European context. Within this context is the inevitable impact that funder policy requirements will have. From 1 April 2022, publishers of in-scope research articles will be required by the UKRI's open access policy to have in place "long-term preservation [which] must be supported via a robust preservation programme such as CLOCKSS, Portico or an equivalent." (p. 10). While at present this is not required of long-form publications or monographs, which will be required to comply with open access mandates for the first time from 1 April 2024, the sector agrees it will be an inevitability at some point in the future. Will there be a central, shared solution for the entire UK prior to this point? And who will support this necessary infrastructure? And whose responsibility is it?

[cOAlition S](#), an international consortium of research performing and funding organisation, launched the initiative [Plan S](#) in 2018, which aims to make all funded research publications open access at the point of publication. Item 7 in the principles recognised that "the timeline to achieve Open Access for monographs and book chapters will be longer and requires a separate and due process"¹ and cOAlition S proposed to release a statement before the close of 2021 "they apply to monographs and book chapters, together with related implementation guidance."² A [statement was released by cOAlition S](#) in September 2021, expressing a "commitment is to make progress towards full open access for academic books as soon as possible" while also acknowledging that "standards and funding models

¹ 'Plan S Principles | Plan S'. Accessed 25 April 2023. https://www.coalition-s.org/plan_s_principles/.

² 'Guidance on the Implementation of Plan S | Plan S'. Accessed 25 April 2023. <https://www.coalition-s.org/guidance-on-the-implementation-of-plan-s/>.

may need more time to develop.”³ Instead of a uniform policy of direct principles, five recommendations were introduced, along with a promise to “collaborate with the OA books community to develop implementation guidelines that respect...bibliodiversity.” Additionally, it was acknowledged that these “guidelines will include a set of technical standards on OA books that mirror the technical requirements cOAlition S has set for OA journals and repositories.” The technical requirements presently set by cOAlitions S for journals and repositories, in terms of archiving and preservation, are the same as that of UKRI: “Deposition of content with a long-term digital preservation or archiving programme (such as CLOCKSS, Portico, or equivalent).”⁴ With both the anticipated UKRI requirements and the future guidelines expected from cOAlition S, there is indeed a precedent being set for open access monographs and technical requirements for preservation. And at the moment, this would automatically exclude the long tail of small presses without an existing preservation plan in place, which would in turn counter the necessary respect for bibliodiversity that cOAlition S has proposed. It is clear a larger solution is necessary, and soon.

At present, there is no central, shared preservation archive for open access monographs within the UK, and endeavours of this kind elsewhere are in their early stages. However, as further requirements emerge for longform research publications from governmental and research funding bodies, there will be an increasingly urgent need for effective and organised solutions. That multiple studies have highlighted this substantial gap in open access infrastructure cements the urgency of this need.

The UKRI gap analysis report cites COPIM’s work in this area as a potential pathway to developing solutions, and this is in part what we have been hoping to achieve with the Thoth Archiving Network (see **Chapter 7**). The Thoth Archiving Network is an initiative for the archiving of OA monographs from small and scholar-led presses

³ ‘COAlition S Statement on Open Access for Academic Books | Plan S’. Accessed 25 April 2023. <https://www.coalition-s.org/coalition-s-statement-on-open-access-for-academic-books/>.

⁴ ‘Technical Guidance and Requirements | Plan S’. Accessed 25 April 2023. <https://www.coalition-s.org/technical-guidance-and-requirements/>.

in a network of participating institutional repositories. The Network is still under development, with successful proof-of-concept already developed (see **Chapters 3 and 4**) and several subsets of publisher content archived at various locations (see **Chapter 5**). This is a ‘tier 1’ option for publishers who have no other external solutions in place to ensure their content does not entirely disappear should they cease to operate. Although there is a real necessity for a more comprehensive system, the goal for the Thoth Archiving Network is to attend to those OA monographs most at risk. We pursue the network with the understanding that archiving itself is not preservation, as many repositories do not have a preservation layer. As we develop the network, preservation options, along with repositories joining the network who have preservation in place, will enrich and improve the offering. In the building of this network, we will provide a service to the presses and publishers who would otherwise potentially disappear from the scholarly record should they cease to operate. Until there is a national solution, we forge ahead with community collaboration in mind.

Additionally, what we hope to provide with the “good, better, best” practice guidebook is a starting point for good practice in archiving and preservation for small and scholar-led presses that will also potentially benefit other types of presses, including new university presses. We will explore specific examples to this effect in **Chapter 2**, covering areas such as file formats; metadata; content selection and packaging; copyright, reuse and licensing; existing routes to digital preservation archives; and archiving and preservation workflows.

However, it is important to recognise, and emphasise, that there are different levels that will be achievable by the various presses. While we will call these “good, better, and best”, these subjective descriptors align to the level of assured benefit a press may achieve by taking certain actions: what certain actions will gain, and what may subsequently remain at risk. Similarly to [Project JASPER](#)’s offered tiers, these will also reflect levels of effort/action required on the part of the presses. Due to various resource deficiencies known to impact the small and scholar-led publisher, a press may only be able to engage the lowest level of effort, which should still afford them a “good” level of archiving/preservation practice.

The NDSA Levels of Digital Preservation

The [National Digital Stewardship Alliance \(NDSA\)](#) was created in 2010 as a membership organisation, part of an initiative of the National Digital Information Infrastructure and Preservation Program of the Library of Congress. The NDSA's [Levels of Digital Preservation](#) were first published in 2013, and updated in 2019 along with supporting documentation and additional resources.

The table below details levels 1 to 5, in each of the following categories: storage, integrity, control, metadata, and content.

Functional Area	Level			
	Level 1 (Know your content)	Level 2 (Protect your content)	Level 3 (Monitor your content)	Level 4 (Sustain your content)
Storage	<ul style="list-style-type: none"> Have two complete copies in separate locations Document all storage media where content is stored Put content into stable storage 	<ul style="list-style-type: none"> Have three complete copies with at least one copy in a separate geographic location Document storage and storage media indicating the resources and dependencies they require to function 	<ul style="list-style-type: none"> Have at least one copy in a geographic location with a different disaster threat than the other copies Have at least one copy on a different storage media type Track the obsolescence of storage and media 	<ul style="list-style-type: none"> Have at least three copies in geographic locations, each with a different disaster threat Maximize storage diversification to avoid single points of failure Have a plan and execute actions to address obsolescence of storage hardware, software, and media
Integrity	<ul style="list-style-type: none"> Verify integrity information if it has been provided with the content Generate integrity information if not provided with the content Virus check all content; isolate content for quarantine as needed 	<ul style="list-style-type: none"> Verify integrity information when moving or copying content Use write-blockers when working with original media Back up integrity information and store copy in a separate location from the content 	<ul style="list-style-type: none"> Verify integrity information of content at fixed intervals Document integrity information verification processes and outcomes Perform audit of integrity information on demand 	<ul style="list-style-type: none"> Verify integrity information in response to specific events or activities Replace or repair corrupted content as necessary
Control	<ul style="list-style-type: none"> Determine the human and software agents that should be authorized to read, write, move, and delete content 	<ul style="list-style-type: none"> Document the human and software agents authorized to read, write, move, and delete content and apply these 	<ul style="list-style-type: none"> Maintain logs and identify the human and software agents that performed actions on content 	<ul style="list-style-type: none"> Perform periodic review of actions/access logs
Metadata	<ul style="list-style-type: none"> Create inventory of content, also documenting current storage locations Backup inventory and store at least one copy separately from content 	<ul style="list-style-type: none"> Store enough metadata to know what the content is (this might include some combination of administrative, technical, descriptive, preservation, and structural) 	<ul style="list-style-type: none"> Determine what metadata standards to apply Find and fill gaps in your metadata to meet those standards 	<ul style="list-style-type: none"> Record preservation actions associated with content and when those actions occur Implement metadata standards chosen
Content	<ul style="list-style-type: none"> Document file formats and other essential content characteristics including how and when these were identified 	<ul style="list-style-type: none"> Verify file formats and other essential content characteristics Build relationships with content creators to encourage sustainable file choices 	<ul style="list-style-type: none"> Monitor for obsolescence, and changes in technologies on which content is dependent 	<ul style="list-style-type: none"> Perform migrations, normalizations, emulation, and similar activities that ensure content can be accessed

Figure 1 - NDSA Levels of Preservation (CC BY-SA 4.0)

Though based on version 1.0 of the NDSA Levels of Preservation, the Orbis Cascade [Digital Preservation Step By Step](#) guide provides an overview of practice in the five areas of File Fixity & Data Integrity, File Formats, Metadata, Information Security, and Storage & Geographical Location, with each area detailing practices at levels 1 through 4.

The categories have been slightly altered in the new NDSA levels (detailed by Jenny Mitcham in [this DPC blog post](#)⁵), but the Orbis Cascade guide provides a good overview of the levels that certain preservation practice falls into, and allows publishers and others involved in digital preservation a view towards advancing their practice.

Methods & Approaches

The work package has deliberately chosen a multi-angled approach to its investigations. We have liaised widely and worked with experts in a variety of fields to ensure that we got the best understanding of the preservation eco-system. The experts consulted include: digital preservation experts (including those working for digital preservation archives, those who work for institutions and organisations where preservation is one part of their work, and those involved in related projects); publishers (both those involved in COPIM and external partners); librarians; and technical experts.

Consultation has taken a variety of forms and includes workshops, interviews, presentations, surveys, and project to project meetings. In addition, [we conducted a literature review](#) to understand the present state of affairs and understanding.

The work package conducted a total of three external workshops and one internal workshops. The three external workshops were:

⁵ 'Introducing the New NDSA Levels of Preservation - Digital Preservation Coalition'. Accessed 25 April 2023. <https://www.dpconline.org/blog/introducing-the-new-nds-a-levels-of-preservation>.

- October 2020: [Scoping workshop](#) to understand the preservation arena and to begin to build relationships with interested parties.⁶
- November 2022: [UKCORR \(the United Kingdom Council of Open Research and Repositories\) workshop](#) to introduce the Thoth Archiving Network to UK Repository Managers and to get their thoughts on how we could progress the network.
- March 2023: [Copyright Workshop](#) to understand the copyright environment and the impact copyright legislation may have on the Thoth Archiving Network.

The internal workshop was focused on robust links and what options we could consider for preserving links to third party content within open access monographs. This provided a great deal of useful information for future work that will be further considered within the [Open Book Futures project](#), set to begin May 2023.

A survey of small, scholar-led or university, open access presses in Nov-Dec 2022 has also fed directly into this guide as it was used to identify the priority areas where presses would like guidance on preservation. This has contributed directly to the content in chapter five on this guide.

Finally, we have elicited feedback from attendees at conferences where work package members have presented our work.

Further details of our work in experimenting with repository workflows, proof-of-concept and establishing of the Thoth Archiving Network (including development of a dissemination tool in Thoth), and considerations for the archiving of complex and experimental monographs are included in the chapters following the “good, better, best” practices guidebook.

⁶ The October 2020 workshop and associated interviews formed the basis of the work package’s [Scoping Report](#) published in June 2022.

Chapter 2:

‘Good, Better, Best’: Practice Guidance for OA Monograph Publishers

A basic guidebook for the small and scholar-led press

By Miranda Barnes

Introduction

In many of our discussions with small and scholar-led presses, including a small survey distributed in December 2022, what distinctly emerged was the need for basic guidance around best practices, a “101 type guide” that included information around archiving and preservation. What we have also heard, from colleagues in digital preservation and scholarly communications organisations, is that perfect or “best” can be the enemy of good. While this cannot possibly be a comprehensive guide to everything that may possibly befall a small press when it comes to archiving and preservation, it is the first iteration of a resource that we hope will benefit this community.

When we employ the terms “good”, “better”, and “best” within this guide, it is with the understanding that they are rather abstract and subjective terms, but also terms that can be used to indicate levels of quality. “Good” practice is just that: good practice. This is the baseline to achieve and is always better than no effort at all. We know from our [Scoping Report](#) that small and scholar-led presses face the most challenges with the fewest resources. These presses may not be able to achieve the “better” or “best” levels of practice at first, but they should ideally be able to achieve “good” practice, if they do not do so already. By framing the advances possible in various areas within these terms, or levels, we discuss the reasons why some practices are “better” or “best”: more is achieved to protect and preserve the open access monographs involved.

This guide is not meant to be prescriptive, but a resource. Practices within digital publishing and digital preservation are evolving and changing every day, with research and development in these areas a key area of progress. We hope to update this guide with future versions. The guide covers several main areas that we found to be significant in our conversations with the community stakeholders: file formats, metadata, selecting content for preservation, existing routes to digital preservation archives, archiving and preservation workflows, and copyright, licensing, and third-party content. As there are many existing resources around digital preservation and open access monographs, where appropriate we signpost to these resources rather than duplicating the content.

File Formats

This section will consider the various file formats typically used in the publishing of open access monographs. Though there are many experimental modalities that are also used to create complex and nonstandard scholarly works, these will be further addressed in Chapter 6. Each subsection will examine the positives and drawbacks of the file format within the context of OA monograph preservation. A great deal of work has already been undertaken by organisations such as the DPC and the Library of Congress to provide detailed information about the preservability of various file formats and links to these guides and recommendations are provided below. Our aim here is to address what will be most of interest to small and scholar-led academic publishers who are considering how best to preserve their outputs going forward.

Key resources:

[DPC Digital Preservation Handbook: File formats and standards](#)

[Library of Congress Recommended Formats Statement 2022-2023](#)

PDF

The PDF format is by far the most ubiquitous format used for the publication and preservation of eBooks and monographs. A PDF most closely aligns to and mimics the form of a traditional printed text, and it is familiar and easy to produce. A

publisher or press does not need specialised technical knowledge to use PDFs. There are benefits and drawbacks to the PDF in terms of archiving and preservation, and these will depend largely on the content of the PDF file that represents the monograph in question. However, there is no question that if a publisher preserves this file type, it is highly likely that the PDF format will continue to exist and remain interoperable into the future, because it is so widely used. In this sense, the PDF is the baseline “good” requirement for a file format in terms of preserving the average OA eBook or monograph.

PDF benefits

Morrissey cites Malcolm Todd’s 2009 DPC Technology Watch report⁷ in which he synthesizes various sets of existing criteria at the time to a key criteria list, which can help assess the level of risk to the long-term viability of a file format.

File Format Assessment Criteria

- 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

As Morrissey details, the PDF as a file family fulfills these criteria well:

“Although it is a commercially developed format, it is widely seen as meeting many of the requirements deemed critical to reducing risk to the long- term viability of a format: it is in wide-spread use; there are many implementations, some of them open-source, of viewer applications that can operate on diverse platforms; there are provisions within the format for embedding metadata of various sorts; there is a publicly available specification of the format, control of which has been ceded to a public standards body (ISO).”⁸


⁷ Todd, Malcolm. ‘File Formats for Preservation’. DPC Technology Watch Reports, 2009.

<https://www.dpconline.org/docs/technology-watch-reports/375-file-formats-for-preservation/file>.

⁸ Sheila M. Morrissey, The Network is the Format: PDF and the Long-term Use of Digital Content, *Archiving 2012*, pg. 200-203 (2012).

While this was essentially the case in 2012, development of the PDF family of files has only continued to further fulfil the brief. The PDF was originally introduced as a file format in 1993, with the PDF ISO ([ISO 32000](#)) standard following in 2008, when Adobe released the format as an open standard. (Previous to this point, PDF was a proprietary format.)

For most publishers who are publishing standard text-based monograph publications, a PDF sufficiently suits the brief for an efficient and simple document type for their purposes. It is “self-contained, readily shareable and relatively hard to change.”⁹ The standard PDF allows for detailed metadata, embedded content, externally-linked dependencies, and a variety of other flexible features. Because PDF was designed with the purpose of preserving a page’s image across different devices, originating as part of print workflows, it lends itself readily to the process of publishing in digital form.¹⁰



- ✔ Widely-used and actively-sustained format is likely to continue to be supported into the future (ISO standard)
- ✔ Offers option to embed and link to external content
- ✔ Closely mimics the printed page
- ✘ Embedded content files will not be detected by preservation software

PDF/A

The PDF/A standard ([ISO 19005](#)) was created specifically for preservation and archiving and was introduced in 2005, and has been updated in 2011, 2012, and 2020.

⁹ Johnson, Duff, ‘The Only Archival Digital Document Format - Digital Preservation Coalition’. Accessed 31 March 2023. <https://www.dpconline.org/blog/wdpc/the-only-archival-digital-document-format>.

¹⁰ Kirchoff, Amy, and Sheila Morrissey. ‘Preserving eBooks’. DPC Technology Watch Reports, June 2014. <https://www.dpconline.org/docs/technology-watch-reports/1230-dpctw14-01/file>.

The initial PDF/A format specification restricted any dependencies external to the document, such as fonts not completely contained within the document, links to destinations outside the document, script, or the use of 3D images, audio and multimedia.¹¹ Alterations from PDF/A-1 to PDF/A-2 expanded capabilities, allowing embedding of files, but limited to those that were PDF/A-compliant. Further allowances were added in PDF/A-3, allowing any file format to be embedded. PDF/A-4, the present version as of publication, now supports interactive 3D models.

Library of Congress Recommended Format

Within the Library of Congress’s Recommended Formats Statement, under [ii. Textual Works – Digital](#), the file formats are divided into “Preferred” and “Acceptable”. File formats are then listed in order of preference. Second to XML, PDF is listed, with some specific specifications as to the file format versions and level of quality:



- PDF/UA (ISO 14289-1 compliant)
- PDF/A (ISO 19005-compliant)
- PDF (highest quality available, with features such as searchable text, embedded fonts, lossless compression, high resolution images, device-independent specification of colorspace, content tagging; includes document formats such as PDF/X)

Some drawbacks

Important to understand is that while PDF/A is the preferred file format for preservationists, “PDF/A is a restricted form of PDF intended to be suitable for long-

¹¹ Morrissey.

term preservation by *removing some features* [emphasis added] that pose preservation risks”¹² and “focuses on accurate preservation of the *static visual representation* [emphasis added] of page-based electronic documents over time.”¹³ After all, “PDF’s purpose is to be a document.”¹⁴ And generally, preservation’s purpose is to keep files safe and uncorrupted for as long as necessary, which the PDF/A format affords.

These restrictions, and the focus on “static” representation, can mean that for some more complex, enhanced, or experimental monographs, this will not be an ideal file choice for preservation. If a PDF file is necessary, an up-to-date version of a regular PDF format will prevent the restrictions and allow for external dependencies. While there are still challenges (and ongoing discussion) around how best to preserve external and supplementary content in these cases (see Chapter 6), if these are essential to the work, the PDF/A is not likely to be the best choice.

However, for eBooks and monographs that do not contain embedded audio or video, or necessary external dependencies, the PDF/A is recommended format for publishers to use. Publications that conform to the typical printed page, in that they are static and self-contained, will be best preserved in the archival PDF/A format, assuring they survive unchanged and uncorrupted for the longer term.

Embedded content

For monographs that contain embedded multimedia (i.e. other file formats, such as audio, video, 3-D model) content, or even a high number of images, the primary concern is that a PDF within an active preservation system will be seen only as a file, but not as a “container” of other files. So, while active preservation will assure the PDF file itself remains updated with continuous access, any different, additional

¹² Fanning, Betsy. ‘Preservation with PDF/A (2nd Edition)’. Second. Digital Preservation Coalition, 31 July 2017. <https://doi.org/10.7207/twr17-01>.

¹³ Johnson, Duff. ‘Glossary of PDF Terms’, 7 July 2021. <https://www.pdfa.org/glossary-of-pdf-terms/>.

¹⁴ ‘The Only Archival Digital Document Format - Digital Preservation Coalition’. Accessed 31 March 2023. <https://www.dpconline.org/blog/wdpd/the-only-archival-digital-document-format>.

file formats embedded *within* the PDF will not be updated individually. Whether or not this is a concern will depend on how essential the embedded content is to the scholarly work. In many fields, however, any embedded content is likely to be key to understanding the work in some way, particularly in the Humanities, Arts, and Social Sciences.

If a preserved file is opened years down the line but doesn't "render" (open and display properly), the content can't be read and viewed as intended at the time of creation. Depending on the extent, this can have a large impact on the usability of that file. While a preserved PDF may render some of the embedded content, some may not render properly, or at all, or this content could be degraded. The [US National Archives](#) has noted the particular challenge of video, audio, and other files shoehorned into PDFs in nonstandard ways, using bespoke open-source software rather than the standard software (Adobe).

Additional resources:

[PDF/A Family, PDF for Long-term Preservation](#) (Library of Congress)

[Preservation with PDF/A – 2nd Edition](#) (DPC, 2017)

[Glossary of PDF Terms](#) (pdfa.org)



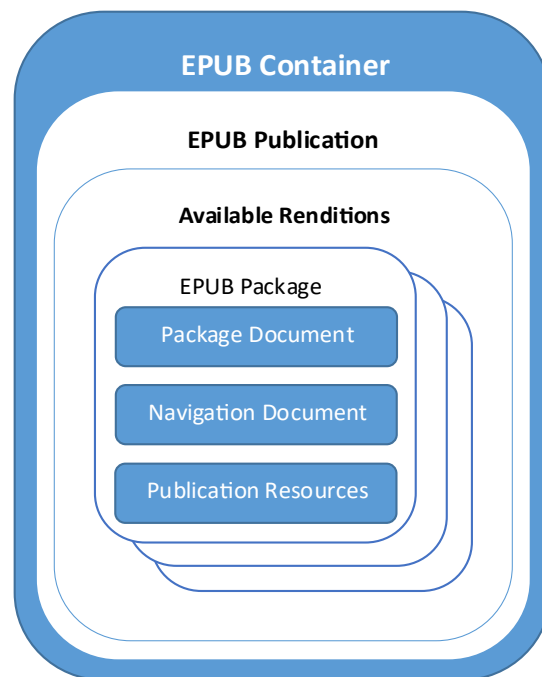
EPUB

An EPUB is an open eBook file format that has a flexibility of presentation allowing eBooks to be read on tablets, smartphones, and compatible eBook readers, as well as on computers, using compatible programs or online website services. EPUB became an [official technical standard](#) of the [International Digital Publishing Forum \(IDPF\)](#) in 2007. As the most widely used and supported XML-based eBook format, and because it is supported by nearly all hardware readers, EPUB is independent of proprietary delivery. 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 of an EPUB is based on the [ZIP](#) format and defined in the Open Container Format (OCF). From the Library of Congress’s entry on the EPUB file family:

“An EPUB Package consists of all the resources needed to render the content. The key file among these is the Package Document, an XML file that serves to centralize metadata, detail the individual resources that compose the Package and provide the reading order and other information necessary to render the Rendition.”¹⁵

The World Wide Web Consortium (W3C) provides the following conceptual diagram of the EPUB “container”¹⁶ and what may be included:



An EPUB can support the use of graphics, interactive elements, videos, audio, and linked content. One primary difference between the EPUB and other eBook formats is that EPUB is a “*reflowable*” document format: content (e.g. text) is presented in a

¹⁵ ‘EPUB (Electronic Publication) File Format Family’. Web page, 12 May 2020. <https://www.loc.gov/preservation/digital/formats/fdd/fdd000310.shtml>.

¹⁶ <https://www.w3.org/publishing/epub32/images/epub.svg>

way that fits the viewer device, the viewer software, or the user's preferences."¹⁷ This means that a change in text size will automatically re-flow the text in a dynamic fashion, allowing for ease of reading. This is counter to the PDF, which has each page fixed in size and positioning of content. (Note: some fixed layouts are possible with EPUB3).

EPUB Preservation

Any files provided to a CLOCKSS or LOCKSS instance will be bit preserved, meaning files will be kept safe and secure, exactly the same as they were originally provided. This would apply to EPUBs, PDFs, and all other file formats equally. Different preservation platforms have different methods of preservation, whether that be bit preservation or normalisation, or a combination of both, where the file formats provided by publishers present challenges to normalisation.

The [Guidelines for Preserving New Forms of Scholarship](#) from NYU Libraries examines several aspects of preserving EPUBs that are of interest. Though some of these may largely apply to enhanced and complex monographs (see more Chapter 6), inevitably there will be crossover with the more traditionally constructed digital monograph.

The preservation-specific EPUB:

Much like a PDF which has all content embedded within the document, an EPUB file that includes "large files, remote resources, or interactive features...can make the EPUB large and therefore impractical for general distribution."¹⁸ For readers with internet-connected devices, using the EPUB's functionality to link and retrieve external content means the disseminated file is not as big and therefore more appealing to the user. However, for preservation purposes, any externally linked or

¹⁷ van der Knijff, Johan. 'EPUB for Archival Preservation'. Open Preservation Foundation. KB/ National Library of the Netherlands, 20 July 2012.

<https://openpreservation.org/system/files/epubForArchivalPreservation20072012ExternalDistribution.pdf>.

¹⁸ <https://preservingnewforms.dlib.nyu.edu/guidelines/tag:EPUB>

located material will inevitably be at risk, whether it be due to link rot, obsolescence, or any number of other factors.

The guidelines therefore advise that where the “publishing platform has mechanisms for generating EPUBs, implementing a workflow for a preservation-specific EPUB3 that can be created alongside the public-facing ebook would be a boon to preservation services.”¹⁹ This EPUB version should:

- 1 abide by the official standard
- 2 keep to core media types
(as defined by the EPUB specification)
- 3 avoid encryption
- 4 encapsulate all required resources within the EPUB file

Other helpful guidelines include: use open, non-obfuscated, non-copyrighted fonts and embed them in the EPUB; avoid using iframes to embed in EPUBs; and when you must use remote resources or non-core media types in an EPUB, define a fallback. Read the full list of guidelines here: <https://doi.org/10.33682/221c-b2xi>

EPUB and normalisation

If an EPUB is provided to the [Portico digital preservation archive](#), for instance, with the accompanying extrinsic XML markup, the EPUB can be migrated, or converted, to Portico’s archival eBook XML format. Portico then preserves both the converted/transformed XML and the original eBook format artifacts. The benefit of normalising all content to the same file format and standard means it is then possible to turn on access to the content quickly if it is triggered.

¹⁹ Ibid.

Why is this “better”?

- ✓ Widely-used open standard format
- ✓ XML-based (an LoC preferred format)
- ✓ Additional metadata is not always required (as it would be with PDF) because metadata can be extracted from the EPUB itself
- ✓ Recognised as a “container” by digital preservation archive software (therefore other packaged files are updated)
- ✗ Requires publishing platform to generate EPUBs

Key resources:

[EPUB \(Electronic Publication\) File Format Family](#) (Library of Congress)

[Preserving eBooks](#) (DPC, 2014)

[EPUB for Archival Preservation](#) (OPF, 2012)

HTML

HTML is a markup language (a text-based code) that generates the appearance of content on the internet. HTML stands for hypertext markup language and is probably the most well-known and familiar of the markup languages. HTML is used to format webpages as well as tell web browsers how a document should look, whereas XML (below) is used to describe the content of a document, or how it is organised.



Some publishers, such as Open Book Publishers and Open Humanities Press, do create an HTML version of their open access monographs, which is produced to view online in a web browser, like a webpage.

HTML versions of open access monographs can be read online without downloading. Much like other webpages, these are “webcrawled” by the Internet Archive and preserved for future viewing.

XML

XML stands for Extensible Markup Language and is a markup language used to describe a document's content and data structure. XML is, therefore, 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). As previously mentioned, following the defined standard is necessary for successful long-term preservation and later rendering. As with PDF, if XML files are created in nonstandard ways, this could jeopardise future useability and prevent proper rendering.

Also, the data must be “well-formed”: “A well-formed XML file conforms to a set of very strict rules that govern XML. If a file doesn't conform to those rules, XML stops working... you can share XML data among programs and systems only if that data is well-formed.”²⁰ Precise, standard, and well-formed XML data is key to the language's interoperability, too. Also important to remember is that having the XML alone is not sufficient to create an *easily* readable eBook – the XML must be transformed into the readable eBook in a format such as EPUB. The benefit of a PDF is that it is independently readable immediately upon opening.

Because XML is used to describe the meaning and hierarchical order of the content, or data, this enables separation of content and structure. This in turn makes XML interoperable with various systems. XML is also a WC3 industry standard for delivering content on the internet, which helps to assure consistency as well as persistence. An XML version of an open access monograph has benefits for preservation, as well as dissemination: XML will allow the monograph to be as accessible and reusable as possible. Many among the scholarly publishing and digital preservation communities will be familiar with [the FAIR principles](#), which are growing in importance as digital scholarship progresses. These are: Findable, Accessible, Interoperable, and Reusable. XML, if accompanied by standard,

²⁰ ‘XML for the Uninitiated - Microsoft Support’. Accessed 28 April 2023. <https://support.microsoft.com/en-us/office/xml-for-the-uninitiated-a87d234d-4c2e-4409-9cbc-45e4eb857d44#bm2>.

necessary schema or DTD, therefore goes a fair way towards meeting these designations for open access monographs.

- **Findable:** XML contains usually rich, machine-readable metadata, which contributes to making it easily discoverable.²¹
- **Accessible:** The FAIR principles call for data to be both machine and human-readable to facilitate the retrieval and analysis of resources. XML is both a human and machine-readable format.²²
- **Interoperable:** XML is interoperable. The U.S. Department of Justice Office for Justice Programs states, “XML is the “glue” that promotes interoperability—it allows systems already in use and those being developed to communicate with each other.”²³
- **Reusable:** XML is flexible, transformable, and reusable.

As mentioned above, the Library of Congress lists the XML as the primary preferred file format²⁴ for the existing preservation of digital text documents. As XML is extensible, meaning new tags can be added, the Library of Congress specifies that preferred XML should be provided in recognized and standard formats, and the files should also come with schema, presentation stylesheets, and explicitly stated character encoding. These are to ensure the XML version of a document will display, or render, as intended. Publishers must be sure to follow the standard for the schema (XSD) or DTD, and not use bespoke, unique schema, or there could be challenges in rendering the work in the future.

²¹ https://ukdataservice.ac.uk/app/uploads/rdm_makingdatafair_2020-01-24.pdf

²² ‘Right to Data Portability’. ICO, 17 October 2022. <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-data-portability/>.

²³ ‘Extensible Markup Language (XML) and Its Role in Supporting the Global Justice XML Data Model’, 2004. https://bjaojp.gov/sites/g/files/xyckuh186/files/media/document/What_is_XML_article.pdf.

²⁴ <https://www.loc.gov/preservation/resources/rfs/text.html#digital>



Xml icon created by Freepik - Flaticon

1. XML-based markup formats, with included or accessible DTD/schema, XSD/XSL presentation stylesheet(s), and explicitly stated character encoding
 - a. EPUB3-compliant. (Other versions of EPUB are also preferred formats but EPUB3 is the most common.)
 - b. BITS (Book Interchange Tag Suite) version 2.0
 - c. Other widely-used book DTDs/schemas (e.g., TEI, DocBook, etc.)

While digital preservation archives need to be prepared to take whatever the publisher is able to provide to them for the purposes of preservation, XML is often a preferred format for many because of its flexibility and capacity for active preservation. Portico, for instance, can take apart any XML file and recombine the file in their standard Portico XML, as part of the normalisation process upon ingest. (Portico is more likely to receive JATS XML for journals, than BITS XML for books – in general the most important factor is the provider sending the content in the same format every time, whether it is BITS XML, ONIX XML, or a spreadsheet of metadata.) By normalising all content to the same file format and standard, it is then possible for Portico to turn on access to the content quickly (if content is triggered). Portico’s systems can also scan the XML for any referenced files to ensure they were packaged together for ingest, and present so the monograph can be properly rendered. This is highly complex work that requires a great deal of expertise and may not be the case for every digital preservation archive.

There are differing opinions in the world of digital preservation regarding whether or not preserving the XML alone (along with accompanying schema/DTD) is sufficient. Some digital preservation specialists recommend also preserving an immediately readable PDF or EPUB version alongside the XML to ensure the layout and intended structure are preserved accurately. As mentioned below in the **Packaging content** section, we do recommend all versions are preserved together if they are available from the publisher, as this covers all eventualities, including layout. While this section is solely about the benefits of formats, we would like to restate that multiple versions preserved are better than one in most instances.

Important discussions around archive copies vs. access copies, and the various file format benefits for both, are ongoing within the digital publishing and preservation communities and those same concerns apply here.



Why is this “best”?

- ✓ A preferred and recommended format for the preservation of digitally published books
- ✓ XML eBooks can be normalised within some digital preservation workflows, depending on the archive
- ✓ ***If packaged along with schema and stylesheets***, contains all necessary material for accurate rendition
- ✓ Complies with the FAIR Principles
- ✓ Longevity: XML is well-suited to long-term preservation
- ✗ Small and scholar-led publishers may lack the technological expertise, financial and staff resource to regularly produce and publish XML versions of their monographs

In terms of future-proofing for flexibility, automation, preservation, and scaling, XML as a document-centric publishing format does serve the purpose, as detailed by Jonathan McGlone, in his contribution to *The Library Publishing Toolkit* (IDS Project Press, 2013):

“XML workflows enable publishers to output content quickly and easily in several electronic formats (EPUB, HTML, PDF); repurpose content into other channels (catalogs, websites, databases, printers); automate processes; scale their services and publications; and preserve the digital content for the future.”²⁵

In a series of questions that follow, provided for the publisher in order to assess their requirements for XML, the qualities of the format offer clear benefits: searchable, offers content in multiple formats, ability to repurpose, and long-term viability in preservation. But as McGlone mentions, “the upfront costs to establishing an XML workflow can be quite considerable,”²⁶ and publishers will need to assess whether establishing this workflow is reasonably in scope for them,

²⁵ McGlone, Jonathan. ‘Preserving and Publishing Digital Content Using XML Workflows’. IDS Project Press, 2013. <http://deepblue.lib.umich.edu/handle/2027.42/99563>.

²⁶ Ibid.

and whether XML is necessary for their monograph output. However, it would be encouraging to see a community solution developed for small and scholar-led presses who, individually, could not reasonably implement a workflow of this type.

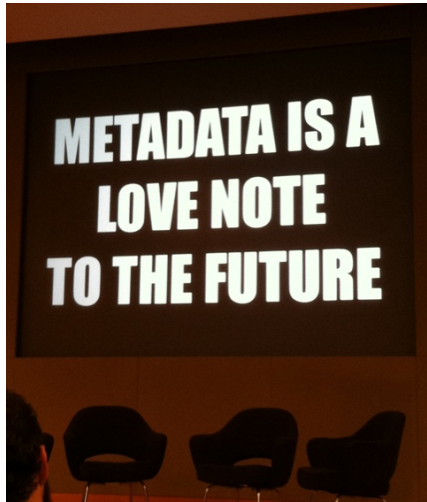
Key resources:

[XML Essentials](#) (W3.org)

[Preserving and Publishing Digital Content Using XML Workflows](#) (The Library Publishing Toolkit, IDS Project Press)

[XML \(Extensible Markup Language\)](#) (Library of Congress)

Metadata



Metadata is key to effective dissemination, discovery, and preservation. Without metadata that is as thorough and complete as possible, scholarly works and other published content can essentially disappear, even when “preserved”. As one digital preservation specialist said to us, with millions of records in a preservation archive, bad metadata means a record may never be found again. For preserved monographs to meet new readers in the future, quality metadata functions as the correct

address.

Jisc’s [New University Press Toolkit](#) has a good primer on metadata for presses, which can include:

- **Bibliographic information**
Such as author(s), title, abstract, publication date, ISBN/ISSN etc
- **Enriched data**
Such as cover images and chapter abstracts

- **Persistent identifiers (PIDs)**

Such as Open Researcher and Contributor Identifier (ORCID) and Digital Object Identifiers (DOI)²⁷

This Toolkit section also has helpful subsections on [Creating Metadata](#), [Metadata formats](#), and [Persistent identifiers \(PIDs\)](#). But as Jisc notes, and many others have found, there is still no minimum set of requirements agreed for metadata. This lack of a standard requirement has inevitably contributed to the gaps acknowledged by Gregg et al (2019)²⁸. And as Adema and Stone note in the 2017 Jisc landscape study, best practices need to be drawn up because of the inconsistency in metadata at New University Presses (NUP) and Academic-led publishers (ALP) that is often due to varying levels of maturity.²⁹ Jisc and OAPEN collaborated on a [metadata model for open access monographs](#) released in 2016, which was then adopted for OAPEN Library. Feedback from consultations with academics, institutional staff, funders and OA monographs publishers fed into the model. The main parts include:

- Book – a description of the monograph or chapter
- Creator – the person(s) responsible for the content of the book
- Funder – the organisation(s) supporting the research
- Format – a description of the digital format(s) that have been made available
- Collection – a description of the collection(s) the book is part of³⁰

[Work Package 5 of COPIM](#) is tasked with addressing many concerns around metadata and dissemination within their open metadata dissemination system

²⁷ Jisc. 'Dissemination', 24 March 2021. <https://www.jisc.ac.uk/guides/new-university-press-toolkit/dissemination>.

²⁸ Gregg, Will, Christopher Erdmann, Laura Paglione, Juliane Schneider, and Clare Dean. 'A Literature Review of Scholarly Communications Metadata'. *Research Ideas and Outcomes* 5 (5 August 2019): e38698. <https://doi.org/10.3897/rio.5.e38698>.

²⁹ Adema, Janneke, and Graham Stone. 'Changing Publishing Ecologies: A Landscape Study of New University Presses and Academic-Led Publishing'. Publication, 30 June 2017. <https://repository.jisc.ac.uk/6666/>.

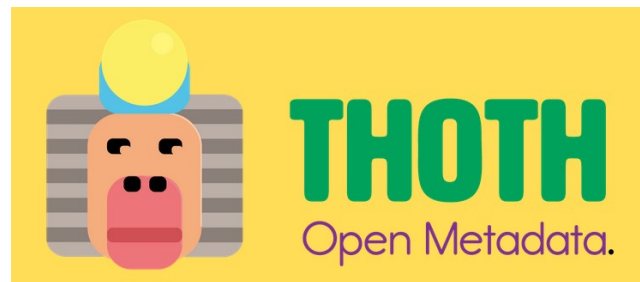
³⁰ Stone, Graham, Rupert Gatti, Vincent W. J. van Gerven Oei, Javier Arias, Tobias Steiner, and Eelco Ferwerda. 'WP5 Scoping Report: Building an Open Dissemination System'. *Community-Led Open Publication Infrastructures for Monographs (COPIM)*, 21 April 2021. <https://doi.org/10.21428/785a6451.939caeab>.

[Thoth](#). Their [Scoping Report](#)³¹ provides a thorough background to the importance of metadata, and the challenges to small and scholar-led presses around time and resource. This report also provides 61 recommendations built on the existing studies, models, and literature towards the improvement of OA scholarly monograph metadata and an agreed upon standard.

First among their recommendations is for COPIM to “consider developing two metadata requirements for OA monographs, a minimum set of metadata requirements and an enriched set.”³² A bare minimum standard would ensure consistency across publishers, but further standards for an enriched set would improve upon this minimum for a more complete set of metadata that could more effectively address the gaps acknowledged by the sector.

Thoth and open metadata

Work Package 5’s open metadata management system Thoth has been built with openness in mind, using open-source code, open APIs, and outputs released under a CC0 license.



Thoth’s main goals are as follows:

- To lower the entry barrier to good metadata management and practices for small/medium OA publishers who are currently struggling to produce their metadata to all the various different specifications that each distributing platform requires;
- To help distribute Open Access books, which have been systematically excluded from a book supply chain that was created for closed books;

³¹ Stone, Graham, John Rupert James Gatti, Vincent WJ van Gerven Oei, Javier Arias, Tobias Steiner, and Eelco Ferwerda. ‘Building an Open Dissemination System’. Report, 27 July 2020.

<https://www.repository.cam.ac.uk/handle/1810/310885>.

³² <https://copim.pubpub.org/pub/wp5-scoping-report-building-open-dissemination-system/release/2?from=104364&to=104488>

- To expose quality and first-hand metadata, using industry standards, publicly for anyone to consume.³³

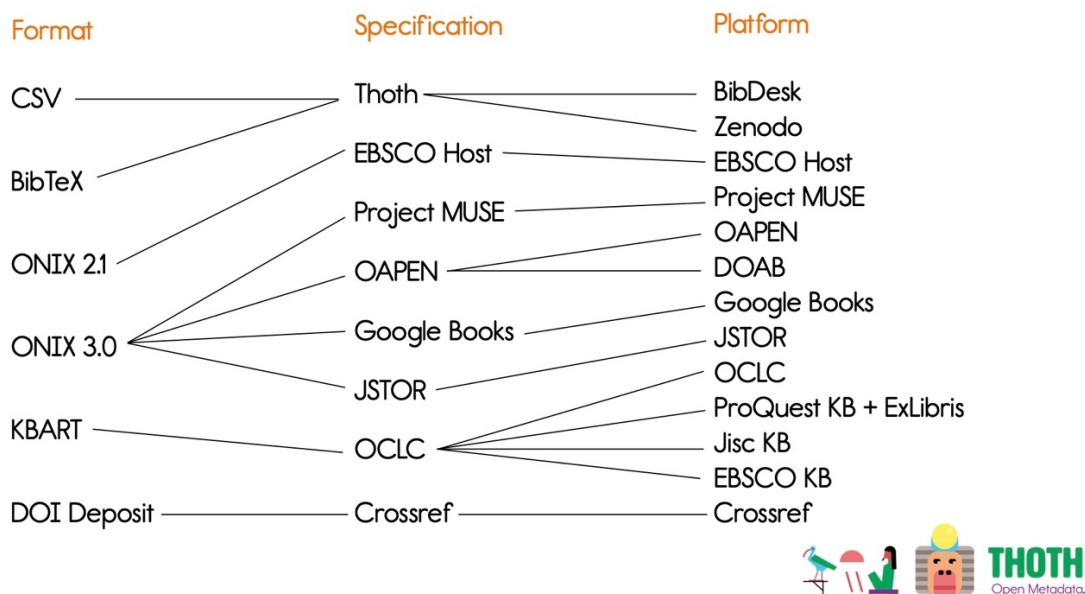
As Thoth developer Javier Arias comments in his interview with Lucy Barnes for Open Access Week 2021, metadata itself “cannot be owned, nor copyrighted. Databases, however, can.”³⁴ What this means is while metadata about a book cannot be closed, an ONIX file can be licensed and therefore closed. Arias continues, explaining that:

“Metadata aggregators collect data from various sources (mainly from the publisher), with which they produce a series of records (ONIX, MARC, etc.) that they then sell to libraries so that the libraries can include the books in their catalogues. This happens to both closed and open books, and while this process may make sense for closed books it really doesn’t for the OA ones — if a book is open, its metadata should be too.”

Thoth enables small and medium open access monograph presses to manage the metadata for their open access books within the system, and then use this metadata for export to various platforms, catalogues and other dissemination channels. A press can use Thoth for their day-to-day metadata management, and the Thoth Dissemination Service will allow for dissemination of metadata as well as content to platforms of their choice. Thoth offers the ability to generate metadata for any monograph in their catalogue in 7 different metadata formats, including multiple varieties of ONIX (for platforms such as JSTOR, EBSCO, and OAPEN) via Thoth’s export API.

³³ <https://copim.pubpub.org/pub/thoth-interview-oaweek2021/release/1?from=786&to=1271>

³⁴ Arias, Javier, and Lucy Barnes. ‘Thoth, Open Metadata and Building Structural Equity: An Interview for Open Access Week’. *Community-Led Open Publication Infrastructures for Monographs (COPIM)*, 27 October 2021. <https://doi.org/10.21428/785a6451.c7ddbe7d>.



Thoth’s basic metadata ingest and export services will remain free, as they have been intended since the service’s formation. As future features are added, there will be a paid level. More on this has been published in [“Developing Thoth’s “Software as a Service” Model](#), a PubPub by Gatti, van Gerven Oei, and Snyder (2022). The curated services offered in the Thoth Plus model will be provided to publishers according to their specific needs, with fees based sensibly on size of press and number/density of records.

While Thoth was created as part of the COPIIM Project, it will continue development as supported by the [Open Book Collective](#) and as part of the [Open Book Futures project](#), also funded by Arcadia and Research England.

Open access and preservation status

Laakso, Wise and Snijder (2022)³⁵ discuss in their iPres2022 conference paper the difficulty in determining open access and preservation status of OA books and how

³⁵ Laakso, Mikael, Alicia Wise, and Ronald Snijder. ‘Peering Into the Jungle: Challenges in determining preservation status of open access books.’ *Conference Proceedings - IPres 2022*, 1 December 2021. <https://ipres2022.scot/conference-proceedings/>. p. 388 – 391.

improved metadata can help. While Laakso et al's earlier study³⁶ of open access journals discovered that 174 OA journals "vanished from the web between 2000 and 2019, spanning all major research disciplines and geographic regions of the world," there so far existed no equivalent study for open access monographs. The work of "an ongoing study...to conduct a data-driven mapping of the current landscape of preservation within the content domain of OA books" is uncovering how insufficient metadata contributes to the overall challenges of obtaining dependable data from bibliometric databases.

- Limited classification of either "Book" or "Monograph" produced unreliable datasets containing theses, book chapters, and books that were not academic monographs (with no reliable, automated way to weed these out)
- There is no widely-used tag indicating a published work is "peer reviewed", which would help narrow the scope
- Ambiguity surrounding OA status: though some detailed granularity of OA metadata was present among the records, overall there was difficulty in determining OA status.
- Varying use of identifiers for books means that some sets will contain the ISBN but not the DOI, and vice versa.
- Reliable preservation data is sparse; only the largest service providers provide this, but their datasets need improvement nonetheless.
- Main digital preservation archives provide ISBNs but not DOIs, which is a challenge as for OA books, most major bibliometric services use DOIs.

Though the study is ongoing, the authors offered some observations and recommendations.

- Data sources that include book materials should strive to include both ISBNs and DOIs in the metadata. (This makes matching to preservation data more reliable.)

³⁶ Laakso, Mikael, Lisa Matthias, and Najko Jahn. 'Open Is Not Forever: A Study of Vanished Open Access Journals'. *Journal of the Association for Information Science and Technology* 72, no. 9 (2021): 1099–1112. <https://doi.org/10.1002/asi.24460>.

- OA status information for preservation: practices and data should be in place both to deposit OA monographs for preservation, and to verify the locations where various pieces of OA content are preserved.
- A service similar to the Keepers Registry for journals should be established for OA monographs.

While only some of these recommendations can be solved via metadata, with the hopes that a Keepers Registry for monographs might be on the horizon, there is a clear case for the importance of consistency, accurate metadata for all presses. Hopefully preservation, open access, and persistent identifier metadata will become standard entries in the future. For now, presses can make seek to make sure these fields are complete in their own.

Key resources:

[Dissemination \(incl. metadata\)](#) (Jisc New University Press Toolkit, 2021)

[Metadata – OAPEN OA Books Toolkit](#)

[COPIM WP5 Scoping Report](#) (2020, 2021)

[Open Metadata in Thoth](#) (COPIM WP5, 2020)

Selecting Content: What is essential?

A publisher should consistently offer the same package of files, which for most small publishers will be the minimum of a PDF file (the content) and a metadata file. As a baseline, this is effective, and assures preservation of the work as it was created. The metadata file submitted should be as complete and thorough of possible. For more about metadata, please see the previous section.



Figure 2 - Photo by [cottonbro studio](#) from Pexels ([License](#))

Packaging content

The key to good practice here is consistency. While there will always be variations in the approaches different publishers might take, for network repositories and digital preservation archives, a standard approach to packaging content is

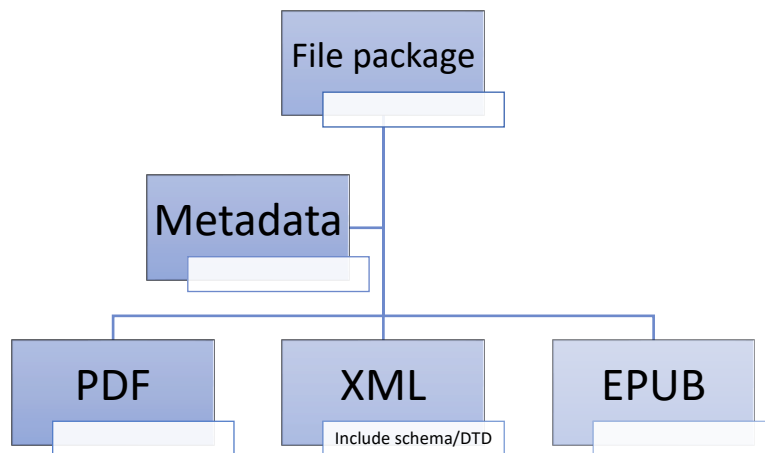
preferred. So, one thing that a press can do is determine a standard package and a workflow to prepare their content.

The minimum package should include a copy of the content file and a metadata file that is as comprehensive as possible. In many cases for the smaller publisher, this will be a PDF and a separate metadata file.

Why is this “good”?

- ✓ The PDF provides a baseline version of record to be preserved in a widely used and perpetually updated format
- ✓ Complete and thorough metadata means the file will remain discoverable
- ✗ Any additional or supplementary content is excluded
- ✗ Embedded multimedia may degrade over time

If there are multiple versions in different formats (PDF, XML, and EPUB, for instance), all versions may be packaged with a metadata file.



Why is this “better”?

- ✓ All versions are preserved
- ✓ Complete and thorough metadata means the file will remain discoverable
- ✓ XML can be “reassembled”; PDF can retain reliable layout images
- ✓ Content embedded in XML or EPUB can be actively preserved and detected by digital preservation software
- ✓ File packages are easily ingested by digital preservation archives (ZIP or file hierarchy)
- ✗ Any issues with the files may not be picked up until scanned in archive

Advanced: preparing content for ingest

The following section is based on, and draws from, the [UK National Archives Digital preservation workflows](#) resource, specifically the Introduction³⁷ and Ingest pages.³⁸ Most steps here will advance beyond the basics and may not be feasible for all publishers, and it is important to understand that the UK National Archives preservation workflows are aimed at archives rather than publishers. Most of these workflow details will be for information only, or a guide to what is often done to prepare content for ingest into a digital preservation archive. However, there are a few steps here that could be built into publisher workflows to prepare for digital preservation that may be adapted as an advanced level of practice, and more broadly could assist anyone dealing with important files.

Begin by organising your files:

- Create a dedicated folder in your preferred storage location.
 - Create a folder and use consistent naming conventions across all your content files. You may wish to create subfolders – one for the content (e.g. called “content”) and one for any documentation about the content (e.g. called “metadata”).
 - Jisc has useful tips on creating a folder hierarchy here: [File management and formats](#).
 - Whatever your system of file naming and file hierarchy, be consistent. As covered elsewhere, consistency is key for digital preservation archives.*
- Understand what you have: Create a list of the content that is being transferred (archived/preserved)
 - You can use software, such as [DROID](#) (which is free) to identify what you have and create a list of the content. Always include

³⁷ <https://www.nationalarchives.gov.uk/archives-sector/projects-and-programmes/plugged-in-powered-up/digital-preservation-workflows/introduction/>

³⁸ <https://www.nationalarchives.gov.uk/projects-and-programmes/plugged-in-powered-up/digital-preservation-workflows/2-ingest/>.

as much information as possible: file names, file paths, sizes, file formats (especially important), last modified date etc.

- Prepare a “metadata” folder and save the list in an open format (e.g. CSV or XML, rather than a proprietary format, like .xls).

Understand what you have, part 2: File manifest and checksums

A primary step recommended in preparing your catalogue, or files, as a publisher for preservation is this: to create a “verifiable file manifest” that lists all files in each file package along with a checksum for each file. Package this manifest with your library files to provide additional authenticity and validity of the content.

verifiable file manifest – a file manifest is a metadata file that accompanies and describes a group of files that are part of a set or coherent unit, such as the library of a publisher. From Sharon Meekin of the DPC: “At a minimum this should be a list of the files, their storage locations, and a **checksum** for each file (this is an alphanumeric string of characters that is generated by a software tool to represent a file’s structure, more on this in a moment). There are many free “characterization” tools that can generate this information for you, including DROID from The National Archives (UK).”³⁹

DROID: <https://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/>

checksum - A unique numerical signature derived from a file, used to compare copies. A checksum on a file is a ‘digital fingerprint’ whereby even the smallest change to the file will cause the checksum to change completely. (Source: [Digital Preservation Coalition](#), [Fixity and Checksums – DPC Handbook](#))

³⁹ ‘Getting Started in Digital Preservation: Taking Your First Steps’, 10 June 2020. <https://blog-ica.org/2020/06/10/getting-started-in-digital-preservation-taking-your-first-steps/>.

This verifiable file manifest, including checksums, can then be validated on receipt, proving no files were corrupted in transit, no files are missing, and no extraneous files have been introduced. This may then be shared with archives where the content is ultimately deposited, proving chain of authenticity from publisher to archive.

Validation

Validation of files is advanced and complex, and while there are multiple options for free and open-source software available for anyone to use, comprehending the error reports from them require a high degree of expertise. So, it isn't recommended that validation software is used by the publisher, particularly to try and "fix" any reported errors that come back on their files, as some errors may not actually cause problems in rendering, and without that expertise, "fixing" may cause more danger of damaging files in other (unnecessary) ways. As Paul Wheatley of the DPC cautions, "interpreting the relevance of those reports to any sensible digital preservation intervention is usually very difficult... a validator may provide a variety of superficially impressive reports, whilst completely failing to check on issues that might be of great interest to a digital preservationist."⁴⁰

However, for information, the following is a list of such software, which can be used to run validation checks on content to determine whether the content conforms to file format specifications.

Validation software (all open/free to use)

[JHOVE](#) (validates certain file formats and also carries out identification)

[Jpylyzer](#) (validates JPEG 2000 Part 1)

[veraPDF](#) (validates PDF/A)

[MediaConch](#) (validates audiovisual files)

⁴⁰ 'A Valediction for Validation? - Digital Preservation Coalition'. Accessed 28 April 2023. <https://www.dpconline.org/blog/a-valediction-for-validation>.

Additional resource: [Bodleian Libraries: Introduction to Digital Preservation: Validation](#)

Important to acknowledge is that practice in this area is evolving and will continue to evolve. Digital preservation as a practice is a relatively new field in the world of publishing, and many publishers are presently gaining a crash course due to impending funder guidelines. Begin with the basics and work up from there.

Existing routes to digital preservation archives

While there are many benefits to online hosting or aggregation, it is important for presses to understand that the presence of content online does not constitute preservation. As Randy Kiefer, Executive Director of the CLOCKSS archive, states in his 2015 UKSG paper:

“Commercial hosting is not preservation. This includes aggregation databases, journal-hosting platforms and distribution platforms for e-books. These are not preservation modes, and they are not archives. Commercial hosting that publishers have with a number of entities, and the relationship with its rights to content, end when the publisher no longer pays for the service. Aggregators are not preservation archivists.”⁴¹

At the moment, there are two primary routes for the open access monograph publisher to archive their publications in an established digital preservation archive. These are either membership directly with the digital preservation archive, or doing so via a third party. In the larger digital preservation landscape, this falls under programmatic preservation. As colleagues at Ithaka S+R (strategies and research) explain:

“There are fundamentally two different types of approaches being taken to preservation: One is *programmatic preservation*, a series of cross-institutional efforts to curate and preserve specific content types or collections usually based on the establishment of trusted repositories. Examples of providers in this category that provide programmatic preservation include CLOCKSS, Internet Archive, HathiTrust, and Portico. In addition, there are *third-party preservation platforms*, which are utilized by

⁴¹ Kiefer, Randy. ‘Digital Preservation of Scholarly Content, Focusing on the Example of the CLOCKSS Archive’. *Insights the UKSG Journal* 28 (5 March 2015): 91–96. <https://doi.org/10.1629/uksg.215>.

individual heritage organizations that undertake their own discrete efforts to provide curation, discovery, and long-term management of their institutional digital content and collections.”⁴²

Ithaka’s focus is heritage organisations, while COPIM is looking more specifically at the small and scholar-led open access monograph publisher. Ithaka’s scoping research and report also set out to specifically examine the third-party preservation platforms, assessing their ability to serve their mission and purpose. What we aim to provide here is instead a brief, informational summary of digital preservation options available to the small and scholar-led press.

We will focus in this guide on the programmatic preservation side, specifically the digital preservation archives already mentioned: CLOCKSS (and the LOCKSS network), Portico, the Internet Archive, and HathiTrust. We will also introduce the Thoth Archiving Network and how this will contribute to the archiving and preservation of open access monographs. A selection of third-party preservation platforms will be listed in this document’s Appendix I: Resources.

An important clarification is non-profit vs. commercial. While a preservation service may be third-party, they may also be non-profit. The reverse may also apply. For instance, Figshare, a commercial institutional and data repository provider, partners with Arkivum, also a commercial provider, for digital preservation. Worth noting here, however, is that the four main archiving and preservation bodies we will be discussing here (CLOCKSS, Portico, HathiTrust, and the Internet Archive) are all non-profits or not-for-profits.

Rieger, Schonfeld, and Sweeney (2022) acknowledge that “in practice this is a false dichotomy as there are hybrid deployment approaches combining tools developed by vendors and not-for-profit entities that operate within the same market (sometimes competing for the same clients).”

⁴² Rieger, Oya Y., Roger C. Schonfeld, and Liam Sweeney. ‘The Effectiveness and Durability of Digital Preservation and Curation Systems’. Research Report. Ithaka S+R, 19 July 2022. <https://sr.ithaka.org/publications/the-effectiveness-and-durability-of-digital-preservation-and-curation-systems/>.

Not-for-profits



Commercial



Figure 3 - Rieger, Schonfeld, and Sweeney, 'The Effectiveness and Durability of Digital Preservation and Curation Systems'. Ithaka R+S.

Programmatic Preservation



Internet Archive - <https://archive.org/>

The Internet Archive is a non-profit digital library founded in the United States in 1996 and began with an aim to begin archiving the internet's websites, at risk from obsolescence. The Internet Archive now holds:

- 735 billion [web pages](#)
- 41 million [books and texts](#)
- 14.7 million [audio recordings](#) (including 240,000 [live concerts](#))
- 8.4 million [videos](#) (including 2.4 million [Television News programs](#))
- 4.4 million [images](#)
- 890,000 [software programs](#)

The Internet Archive allows anyone, including members of the public, to both download and upload digital content to their data cluster (to upload you first must register for a free account). However, the majority of content on the Internet

Archive is collected automatically by its web crawlers. These crawlers harvest any available content online, and for scholarly content, this includes journal articles, reports, conferences, and monographs.

[Project JASPER](#) is working closely with the Internet Archive to provide an effort-minimal preservation option, via the web-crawling route, for scholarly journal publishers that have no other preservation method in place. COPIM's current work towards the Thoth Archiving Network has completed a pilot test, performing a bulk upload of 600 works already in Thoth into the Internet Archive via a registered account. This was successful and the Thoth Archiving Network collection can be viewed here: <https://archive.org/details/thoth-archiving-network> More on this process, which provides a fully-open, base-level archiving option to the Network's participating publishers, is detailed in Chapter 5.

Wayback Machine

Webpages and internet sites are preserved using [the Wayback Machine](#). Researchers can also use the Wayback Machine to archive any website as it exists at the time of research, which then provides an archived link that can be reliably referred to in the future. Weblink archiving may become a critical tool as a remedy against link rot, and perhaps as part of future citations where online references are critical to the validity or context of the research.

Scholar IA

Scholar IA (<https://scholar.archive.org/>) is a beta service still in development. It offers a full-text searchable index of 25 million research articles and "other scholarly documents" (including open access conference proceedings) that are currently preserved within the Internet Archive. At time of publication, Scholar IA does not appear to include monographs.

Open Library

The Internet Archive's [Open Library](#) allows registered users to borrow and read digital copies of millions of library books via [Controlled Digital Lending](#). Books can be read online or borrowed in other formats using eBook software with validation.



CLOCKSS - <https://clockss.org/>

CLOCKSS (Controlled LOCKSS) is a dark archive, preserving digital content via bit preservation at twelve different “nodes”, or mirror repository sites, at academic institutions in diverse geographical locations. Content is migrated to the latest format when it is triggered to assure it remains useable. CLOCKSS is built with LOCKSS (Lots Of Copies Keeps Stuff Safe)* open-source technology.

CLOCKSS permits ingestion via FTP (file transfer protocol) or [via web harvesting](#), so long as the necessary conditions are met. There are no definitive file-type requirements for publishers, and all file types are permitted via either file transfer or web harvesting. Files are kept preserved unchanged. A technical overview of the CLOCKSS preservation workflow is found on the CLOCKSS website: [How CLOCKSS works](#).

Publisher participation costs can also be found [on the CLOCKSS website](#). Fees include an annual fee based on the total publishing revenue of a participating publisher, and to get started, there is a one-time setup fee. Additional transactional fees may apply based on the amount of preserved material per publisher.

**LOCKSS is the software built by Stanford University that is used to power CLOCKSS, but the two are not interchangeable. LOCKSS is an open-source technology that can be adapted and employed by other preservation services (such as the Public Knowledge Project Preservation Network (PKP PLN) and Michigan Digital Preservation Network (MDPN)). The term has also been known to refer to the Global LOCKSS network.*



Portico - <https://www.portico.org/>

Portico is a community-supported digital preservation archive that safeguards access to e-journals, e-books, and digital collections. Portico is a dark archive overall, triggering access to preserved digital content when there is a [triggering event](#), such as the cessation of a publisher’s operations. However, Portico

differentiates between proprietary content and that which has been published open access and works to make any clearly indicated open access content “triggered open” by default, making it openly accessible to anyone.

Portico’s digital preservation service is part of [ITHAKA](#), a not-for-profit organization helping the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. Portico is a centralized and replicated repository and uses migration as its primary long-term archival approach, as part of a managed preservation strategy. Portico defines fully managed digital preservation as the series of policies and activities necessary to ensure the usability, authenticity, discoverability, and accessibility of content over the very long-term.

Active preservation is performed within Portico’s archive, monitoring for threats of technology and format obsolescence, with migration of file formats to the most up-to-date versions wherever possible. Portico performs a “normalisation” process where possible on ingested content to their internal standard XML. Transformed content is also kept preserved alongside the original ingested files and any related information.

An overview-level summary of Portico’s preservation workflow process is available on their website: [Preservation step by step](#). See also full documentation of their [Preservation policies](#) and [how to join](#). Content policies for eBooks (including monographs) are included within the Preservation policies page.

Portico is also working to examine preservation issues surrounding complex and experimental monographs with NYU Libraries as part of the [Embedding Preservability project](#), which follows on from the [Preserving New Forms of Scholarship](#) project on which they were also a partner.



HathiTrust - <https://www.hathitrust.org/>

HathiTrust is a not-for-profit collaborative of academic and research libraries preserving 17+ million digitized items. HathiTrust offers reading

access to the fullest extent allowable by U.S. copyright law, computational access to the entire corpus for scholarly research, and other emerging services based on the combined collection. HathiTrust members steward the collection — the largest set of digitized books managed by academic and research libraries — under the aims of scholarly, not corporate, interests. (Text from hathitrust.org/about

The **HathiTrust Digital Library** is based at the University of Michigan. The digital preservation repository includes library content scanned by Google in its Google Books Library Project, content from the Internet Archive, and member institution contributions, as well as additional contributions from external organisations.

Other programmes that are part of HathiTrust:

- [Emergency Temporary Access Service](#), which permits temporary, emergency access to the collection for member libraries during service disruptions.
- [HathiTrust Research Center](#) offers services that support use of the HathiTrust corpus as a dataset for analysis via text and data mining research.
- [Shared Print Program](#) develops a distributed, shared network of print collections with collective print retention.
- [U.S. Federal Documents Program](#) expands access to and preserve U.S. federal publications.
- [Copyright Review Program](#) review team finds and opens public domain materials in the U.S. and around the world.

HathiTrust currently considers membership applications from academic, research, and university libraries, but not publishers. For more information on how to join, see the information page on their website: <https://www.hathitrust.org/how-to-join>.

Pathways into preservation archives

Aside from direct membership of publishers and libraries to digital preservation archives such as CLOCKSS and Portico, there are some pathways via third party dissemination and aggregation platforms into these preservation archives. These include:



OAPEN - <https://www.oapen.org/>

OAPEN (Open Access Publishing in European Networks) is an online library and publication platform, based in the Netherlands, with its main office at the National Library in The Hague. OAPEN is dedicated to open access, peer-reviewed books, and operates three platforms:

- [OAPEN Library](#) - central repository for hosting and disseminating OA books
- [OAPEN Open Access Books Toolkit](#) - toolkit on OA book publishing for authors
- [Directory of Open Access Books](#) - a discovery service indexing OA books, in partnership with [OpenEdition](#)

OAPEN provides a guide to their typical publisher workflow, which is included below under a CC-BY 4.0 license. This shows the possible ways that publishers' content may arrive within the OAPEN Library, and the relationship between OAPEN and [DOAB \(the Directory of Open Access Books\)](#).

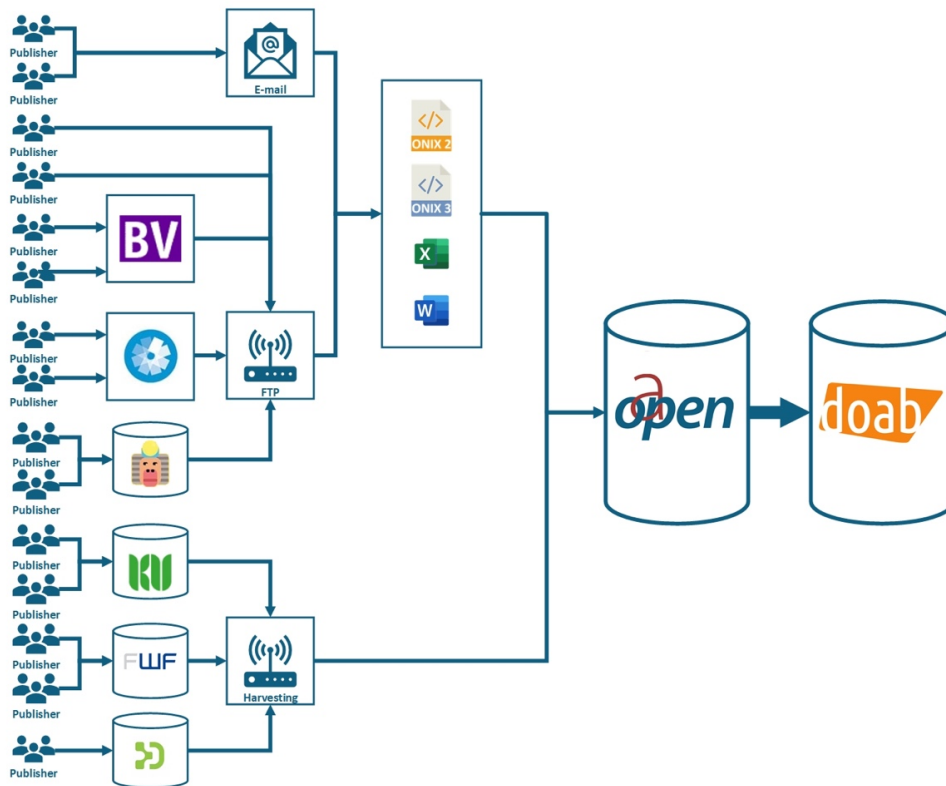


Figure 4 - Source oapen.org (CC-BY 4.0)

OAPEN allows member publishers to [upload books and book chapters](#) to the Library database. Publication file formats that are accepted are PDF and EPUB. Metadata must also be provided, packaged with the book file, in either of two formats (structured textual format, for small publishers under 50 titles, or ONIX 3.0 XML). The OAPEN Library is built using DSpace repository infrastructure and collaborates with Portico for digital preservation. Pricing can be found [on their website](#).



JSTOR - <https://www.jstor.org/>

JSTOR is a digital library originally founded in 1994 to provide access to digitised back issues of academic journals. JSTOR now includes books, primary material, and current issues of journals in the humanities and social sciences. Most content on JSTOR is subscription-only access (often via institutional libraries), but some content is public domain, and any open access content in the library is available free to anyone.

Book publishers can find out more about their [content and metadata standards](#), as well as [how to participate](#), via the JSTOR website.

JSTOR is also part of ITHAKA and archives and preserves all content in Portico.



Project MUSE - <https://muse.jhu.edu/>

Project MUSE is a digital distribution platform based at Johns Hopkins University, providing scholarly access to digital humanities and social science content. Established in 1995, Project MUSE is a not-for-profit collaboration between publishers and libraries, acting as a third-party aggregator similarly to EBSCO, ProQuest, or JSTOR.

Following a grant from the Andrew W. Mellon Foundation in 2016, Project MUSE developed the MUSE Open platform, with the aim of distributing open access monographs from within Project MUSE more broadly, with increased discoverability.

Project MUSE archives and preserves all digital content in Portico.

Archiving & Preservation Workflows

Below is a general archiving and preservation workflow diagram, detailing the component parts of several main steps that are a part of this lifecycle. Digital preservation archives are going to be responsible for most steps beyond the creation of the content.

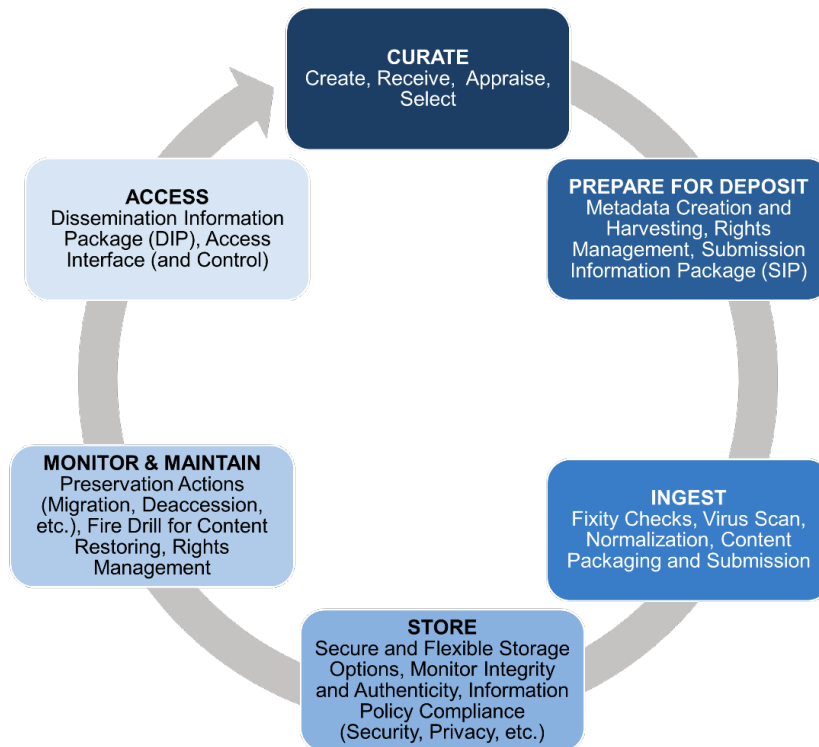


Figure 5 - Rieger, Schonfeld, and Sweeney, 'The Effectiveness and Durability of Digital Preservation and Curation Systems'. Ithaka R+S.

The step in this workflow that will most apply to publishers and presses is 1.1, Create (and, hopefully, 2.1, metadata creation). The DPC has a section within their [Digital Preservation Handbook](#) entitled '[Creating digital materials.](#)' This section opens with the following quote: "The first line of defense against loss of valuable digital information rests with the creators, providers and owners of digital information." ([Waters and Garrett, 1996](#))⁴³

If publishers create and publish their works without consideration of future preservation, this can often be the first thing that could put the work in jeopardy. By using preferred, supported formats in conventional, intended ways, there is a higher likelihood of work surviving for the long term. Less conventional or outright experimental monographs and other scholarly works will inevitably face challenges in this area, particularly because their intention is often to innovate and disrupt. The [Guidelines to Preserving New Forms of Scholarship](#) addresses this in several of

⁴³ 'Creating Digital Materials - Digital Preservation Handbook'. Accessed 13 April 2023. <https://www.dpconline.org/handbook/organisational-activities/creating-digital-materials>.

the guidelines, which can be filtered by the keyword ‘Planning’: <https://preservingnewforms.dlib.nyu.edu/guidelines/tag:planning>. These address ways in which publishers can anticipate future issues and address them event while innovating, as well as advise on where in the process of planning concerns are best addressed. More on complex and experimental works is covered in Chapter 6.

As Colin Post notes in [Educopia’s OSSArcFlow Guide to Documenting Born-Digital Archival Workflows](#), at institutions and organisations, “[w]orkflows can be completely informal and ad hoc or very formalized” or, in some cases, these workflows are “significantly formalized” in parts, but only tacitly known by current staff.⁴⁴ The guide provides advice and procedures to assist organisations in clearly documenting their archiving and preservation workflows. But this is a case in point of the evolving nature of archiving and preservation workflows, and how various technical and organizational factors⁴⁵ will shape how a workflow is constructed. For presses within institutions looking at how to construct a workflow, the guide provides a helpful resource. However, this also reaffirms the evolving nature of the archiving and preservation of born-digital objects (including open access monographs). No doubt this is an area to watch as further technical developments and funder policies impact the world of digital publishing.

Copyright, Reuse & Licensing

The implications of licensing and inclusion of third-party content for archiving and preservation can get quite complicated, as proprietary material may only have permission granted for the publication element, and not for material preserved afterwards. However, open access monographs, by their very nature, should in theory cause no permissions issues for archiving and preservation, as all

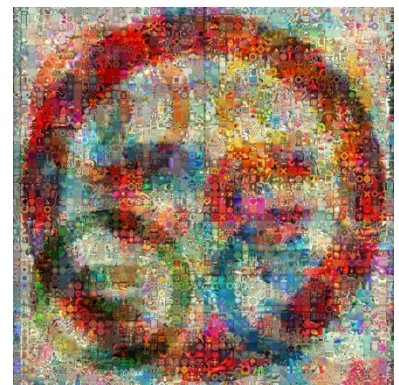


Figure 6 - 2500 Creative Commons Licenses by qthomasbower (flickr.com) CC-BY-SA 2.0

⁴⁴ Chassanoff, Alexandra, and Colin Post. ‘OSSArcFlow Guide to Documenting Born-Digital Archival Workflows | Educopia Institute’. Educopia Institute, 23 June 2020. https://educopia.org/wp-content/uploads/2020/06/OSSArcFlow_Guide_FINAL-1.pdf. p. 30.

⁴⁵ Ibid.

usage permissions for free and open dissemination should have been cleared by the publisher prior to publication.

Any properly licensed open access monograph should be able to be archived, preserved, and disseminated in any location or on any platform without issue. Authors and publishers must work together during the creation of the work to assure any necessary permissions are obtained, and that any licenses for third-party content that differ from the overall monograph license are clearly labelled. A statement in the front matter of the monograph should also communicate the monograph license and state that third-party content within is excluded, directing the reader to refer to the licensing for instances of third-party content individually. This means it also falls to the author and publisher to assure any necessary permissions and clearances are obtained, and third-party content is appropriately labelled with source and credit as well as license.

There are excellent resources already in existence covering the basics of copyright, third party content, and licensing. OAPEN's [OA Books Toolkit](#), for instance, provides guidance to authors surrounding the creation of open access books, but the content provided is useful for anyone involved in the process.



Relevant sections include:
[Contracting and Copyright](#)
[Third-Party Permissions](#)
[Choosing a Licence](#)

Also relevant is [How will researchers use, re-use and build upon my research?](#), which contains a helpful breakdown of the language behind CC-BY licensing.

Jisc's [New University Press Toolkit](#), while primarily aimed at those involved in, or interested in establishing, a new university press, contains relevant information for publishers of all stripes, including the small and scholar-led monograph press. Additional staff within UK institutions will also find this guide useful, such as senior university staff who function as decision-makers, librarians and library directors, scholarly communication managers, and the academic staff member, who may be an author, editor, or publisher. Within this guide, there is a valuable section on [Using third-party copyright](#), including subsections [Creative Commons \(CC\) and rights for reuse](#) and [Third-party rights holders and the cost of licences](#).



Though separate and part of the Jisc [Research Data Management Toolkit](#), the section on [Intellectual property and copyright](#) provides key information about copyright, including explanations of [fair use](#) and [fair dealing](#).

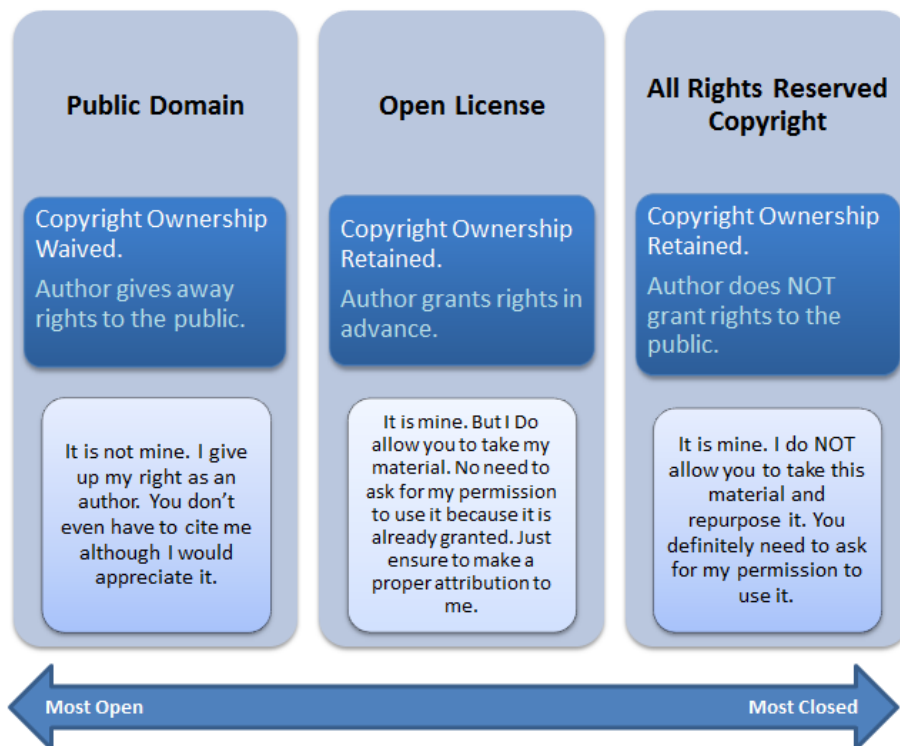


Figure 7 - 'Difference between open license, public domain and all rights reserved copyright' - Wikimedia Commons CC-BY 4.0 ([Boyoungc](#))

Licensing

As the above image conveys, copyright is assigned to the owner of the work, while licensing is the owner's granted permissions for using the work (or not). An author granting permission via, for example, a CC-BY license, still retains ownership and must be attributed, but grants future users permission in advance to use the material as the license indicates.

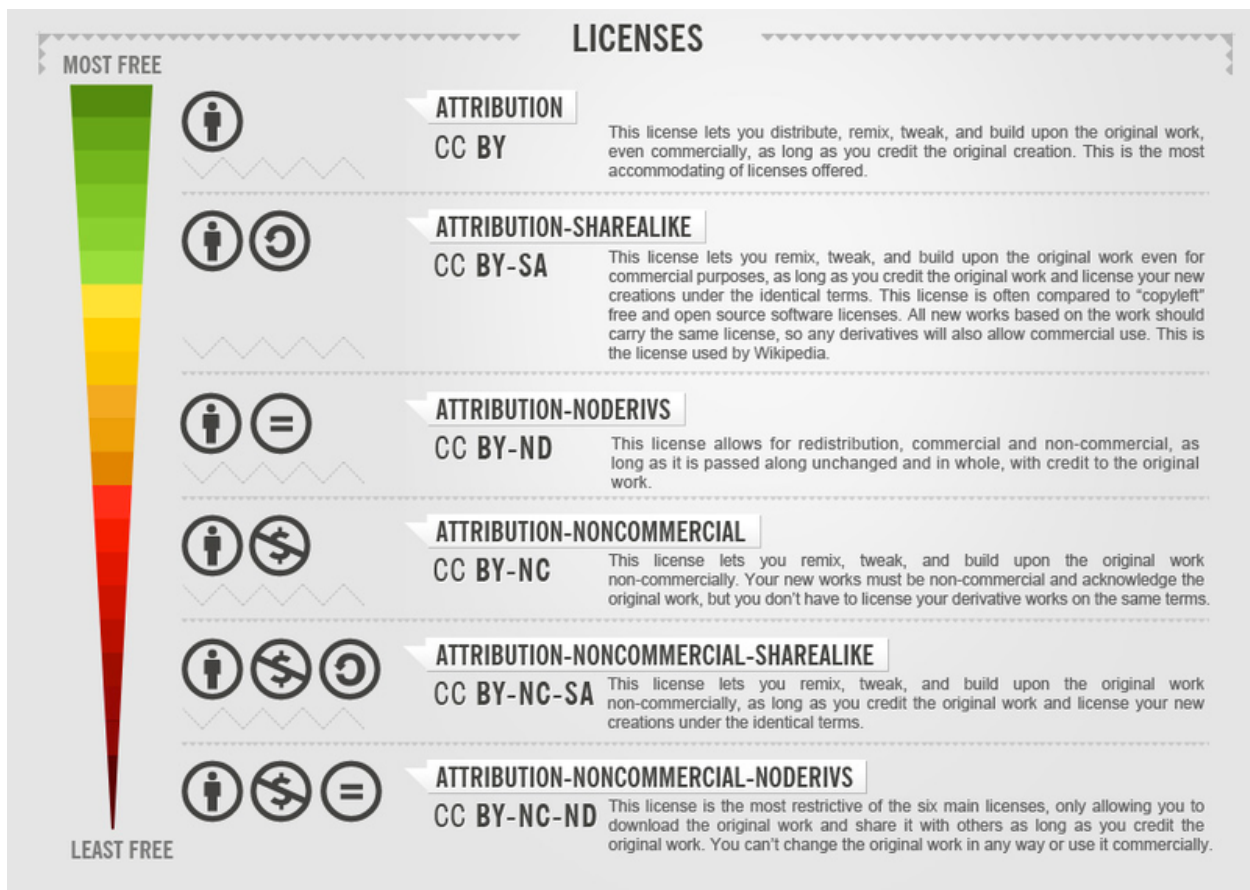


Figure 8 - CC License Freedom Scale Chart - Wikimedia Commons - [Romaine](#)

There are varying levels of permission assigned via each CC-BY license, from full and complete permission to change and build upon the work, including commercial applications (CC-BY), to CC-BY-NC-ND, which only permits downloading and dissemination, but without changing the work in any way or using it commercially. This license is the closest to All Rights Reserved copyright, except it allows for dissemination and download without requesting permission.

As Jisc explain in their Toolkit, “A Creative Commons (CC) licence only covers a new piece of scholarship, e.g. the work that the press is publishing. An author can only license their *own* [emphasis added] work, not that of others. Third party content is therefore excluded from the scope of the Creative Commons licence attached to the new work.”

Third-party content

Jisc’s [New University Press Toolkit](#) section on [third-party content](#) begins by addressing a common myth around third-party content and licensing for open access monographs:

“One of the most important points to make about third-party copyright in open access publications is that they do not need to have the same licence as the published work.”⁴⁶

As the guide elaborates, this is one of the most frequent misconceptions, but it simply isn’t true. If there is clear indication of what license and reuse permissions the third-party content applies, or the content is used within the allowed boundaries of fair use where applicable, the author and publisher have performed their responsibilities in good faith. More detail on good practices surrounding this are covered further down in this section, where workflows are examined.

Within a work, content from a third party that is licensed under a different license from the rest of the work will need to have an indication as to what permissions are given by the third party. This indication should accompany the content in the form of, ideally, a caption within the text where the content is located, but could also be referenced in the form of a footnote or detailed elsewhere in the monograph’s material. However, a footnote is preferred, because there is less chance the license information might be missed or ignored.

⁴⁶ <https://www.jisc.ac.uk/guides/new-university-press-toolkit/production#third-party>

The passage below from Jisc makes clear the reasons for this, as well as emphasizing the importance of seeking permissions where necessary.

If an image, graph or diagram that an author wants to use in a publication does have its own CC licence, the rights holder has made the permissions on reuse very clear and the image can be used without seeking further permission within the terms of that licence.

If this is not the case, or the proposed use is different than the terms of the CC licence allows, then the author needs to seek permission to reuse from the rights holder – often another publisher.

The press must ensure that, for any third-party material where the rights for reuse have been obtained, the copyright of this material is clearly stated. This will then ensure that the rights holder for this material is clear and that anybody who reuses the third party content without permission from the original rights holder would be violating the third party's copyright, even if they found the content in an open access publication.

Figure 9 - Creative Commons (CC) and rights for reuse (Jisc NUP Toolkit, CC-BY 4.0)

As Jisc notes, clearly labelling third party content with copyright and licensing information within an open access monograph may not be enough for some third-party rights holders. They may be concerned that in a digital context, the proliferation of their content online may undermine their business model or revenue streams or fear a higher risk of illegal misuse.

In cases such as this, it is worthwhile, as the publisher, to have a conversation with the third-party rights holder to allay any concerns and try to come to an agreement. If this is not possible, approach your author for an alternative image or other content component to replace the one that is not permitted (or cost-prohibitive) to use. This follows reasonable practice in print publication.⁴⁷

⁴⁷ Jisc. 'Production', 24 March 2021. <https://www.jisc.ac.uk/guides/new-university-press-toolkit/production>.

Third-party licensing workflows for authors and publishers

Within our [copyright workshop](#), Work Package 7 asked, “What are the responsibilities of authors, publishers, and archiving platforms regarding copyright?” The responses guided us to a proposed workflow that may be of use to both parties. The basic/essential version of the workflow for the permissions and archival submission of the final published work should be as follows:

Third-party licensing workflow: “Good”

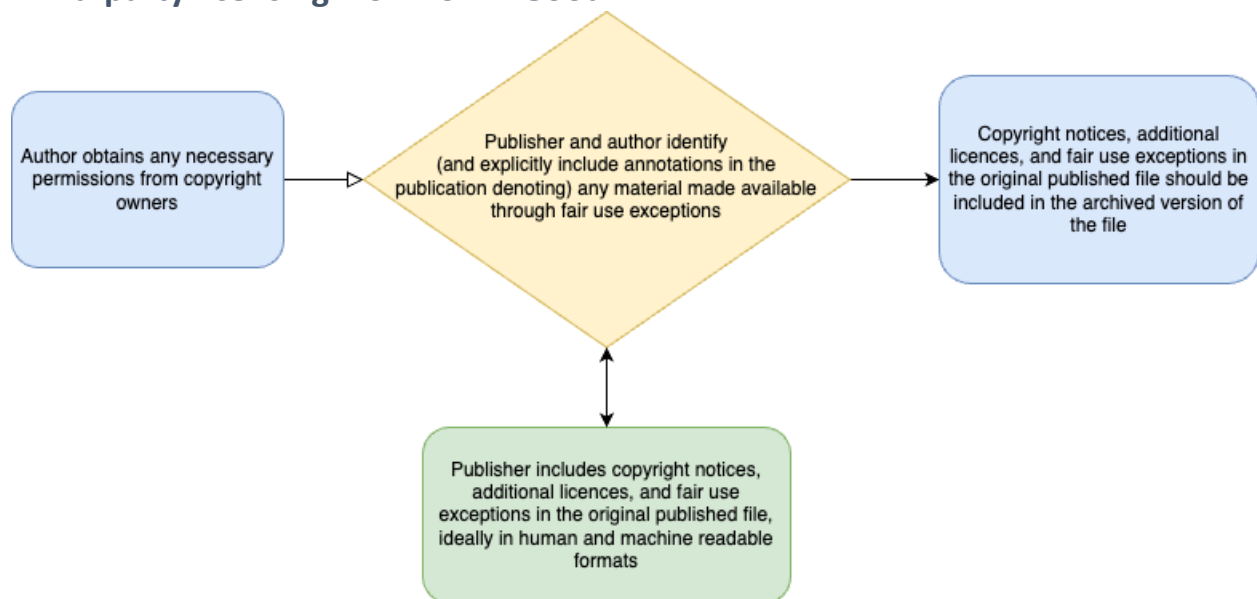
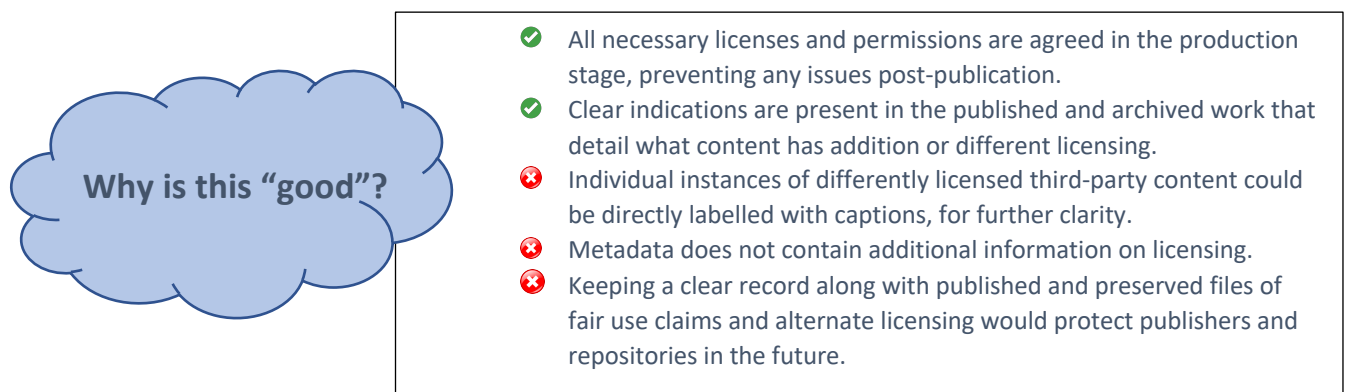


Figure 10 - Third-Party Content Workflow – “Good”



This workflow includes all necessary permissions obtained and documented, but also that this information is included with any and all archived material from the

monograph and its content. Consider future use as well as use for the immediate term when documenting all permissions and licenses.

What could be done better is the direct labelling, as previously mentioned, where instances of third-party content occur within the text. This prevents future users missing or ignoring the differing license.

Third-party licensing workflow: “Better”

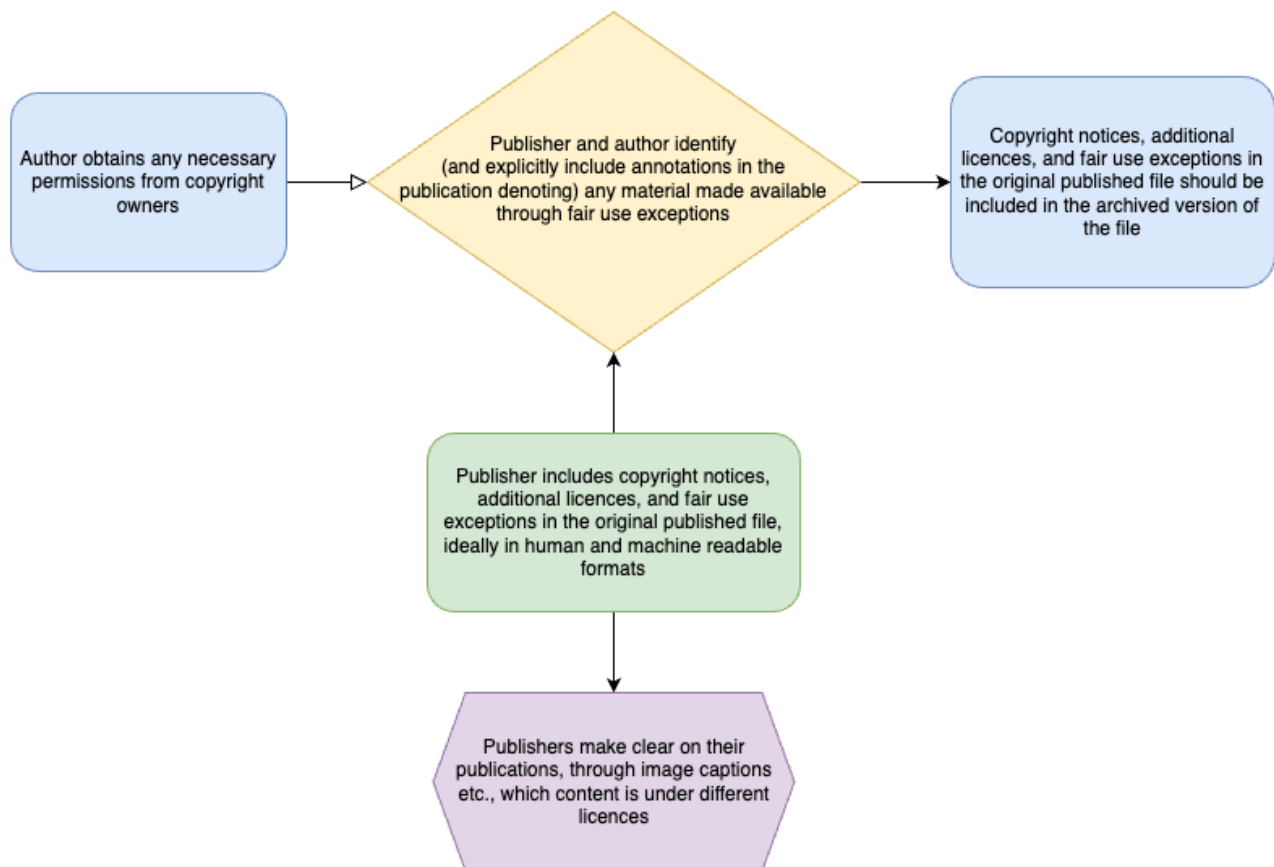


Figure 11- Third-Party Content Workflow – “Better”

Why is this “better”?

- ✔ All necessary licenses and permissions are agreed in the production stage, preventing any issues post-publication.
- ✔ Clear indications are present in the published and archived work that detail what content has addition or different licensing.
- ✔ Captions provide in situ detail about specific third-party content items.
- ✘ Metadata does not contain additional information on licensing.
- ✘ Keeping a clear record along with published and preserved files of fair use claims and alternate licensing would protect publishers and repositories in the future.

While the addition of captioning is “better” practice, there are still further steps that could improve on this workflow. Enriching the metadata for the monograph would mean all licenses have an added layer of documentation. As well, a private documentation record kept with managed and preserved files would also protect publishers and repositories against take-down requests in the future.

Third-party licensing workflow: “Best”

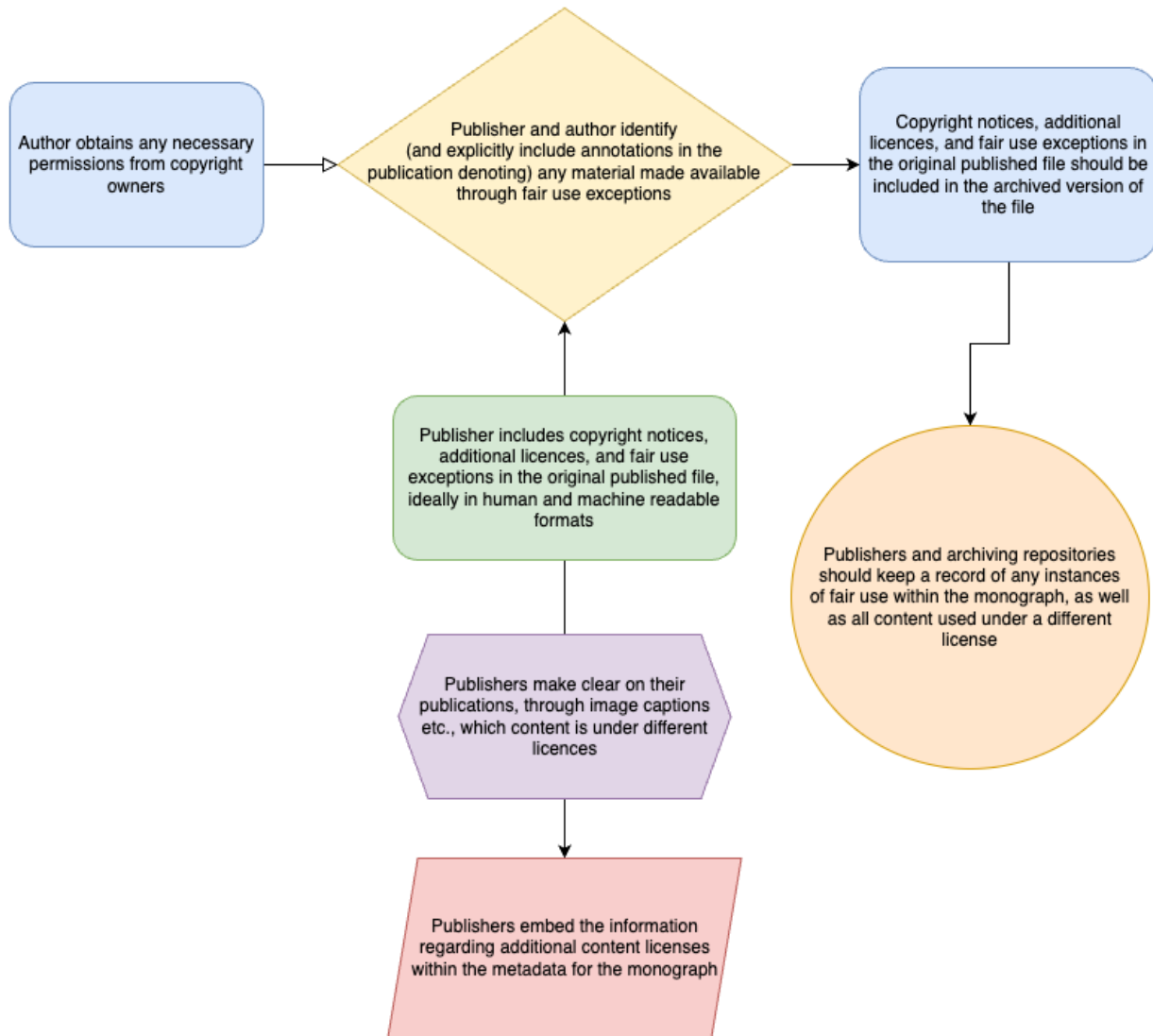


Figure 12 - Third-Party Content Workflow – “Best”



Why is this “best”?

- ✓ All necessary licenses and permissions are agreed in the production stage, preventing any issues post-publication.
- ✓ Clear indications are present in the published and archived work that detail what content has addition or different licensing.
- ✓ Captions provide in situ detail about specific third-party content items.
- ✓ Metadata is enriched and complete with this information added.
- ✓ Both publishers and archiving repositories are “covered” for future challenges and takedown requests because a record (either private, if needed, or open) is kept of all third-party fair use instances and alternate licensing.

As with all suggested improvements, the progress from good, to better, to best may be iterative, if at all. Some publishers will not be able to do all the recommended steps due to deficiencies in resource. The advice here it meant to show the advancement gradient of all opportunities, and not to be prescriptive. While best is always the “gold standard”, any action taken is better than none, and meeting the “good” standard is the ideal, if that is what is possible for an individual publisher.

Key resources:

[Jisc OA mythbusting webinar on copyright](#)

[Code of Best Practices in Fair Use for Scholarly Research in Communication](#) (CMSI)

[Copyright Literacy](#) (CopyrightLiteracy.org)

[OA Books Toolkit](#)

- [Contracting and Copyright](#)
- [Third-Party Permissions](#)
- [Choosing a Licence](#)
- [How will researchers use, re-use and build upon my research?](#)

Jisc’s [New University Press Toolkit](#)

- [Using third-party copyright](#)
 - [Creative Commons \(CC\) and rights for reuse](#)
 - [Third-party rights holders and the cost of licences.](#)

[About CC Licenses](#) (Creative Commons)

[Copyright](#) (UK Data Service)

Chapter 3:

Case Study 1 – Manual Ingestion

Experimenting with repository workflows for archiving

A [version of this chapter](#) first appeared on COPIM's open documentation site, copim.pubpub.org.

Over the course of the 2021-2022, colleagues in [COPIM's archiving and preservation team](#) considered ways to solve the issues surrounding the archiving and preservation of open access scholarly monographs. Most large publishers and many University presses have existing digital preservation relationships with digital preservation archives, but small and scholar-led publishers lag behind due to lack of resource.

One of the potential solutions we considered is the university repository as open access archive for some of these presses. COPIM includes a number of scholar-led presses, such as [Mattering Press](#), [meson press](#), [Open Humanities Press](#), [Open Book Publishers](#) and [punctum books](#). Partners on the project also include University of California Santa Barbara (UCSB) Library and Loughborough University Library. In cooperation with Loughborough University Library, we began to run some preliminary repository workflow experimentations to see what might be possible, using books from one of the partner publishers. Loughborough University employs Figshare as their primary institutional repository, so we began with this as a test bed for our experimentations.

A tale of two monographs: Open Book Publishers

The first volume from Open Book Publishers (OBP) that was employed in these workflow experiments was *Denis Diderot 'Rameau's Nephew' – 'Le Neveu de Rameau': A Multi-Media Bilingual Edition* (<https://doi.org/10.11647/OBP.0098>). The reason for the selection of this book is the relative complexity of the content.

There are images, audio files, and additional texts, as well as several different file format versions of the main text. The 13 audio files, offered in both .wav and .mp3, are essential components to the text, important to the understanding of the work as a whole. Because part of our work investigating the archiving and preservation of digital monographs is the varying level of complexity they possess, we felt this would be a good selection for our exercises.

Internal deposit

The first manual workflow experiment approached the deposit of Rameau's Nephew and all corresponding materials as if the book had been published by a theoretical press, which we called "Loughborough University Press (LUP)." The theoretical author in this instance is an academic depositing internally to Loughborough University, where the press would be positioned. An internal publishing author will already have login access to the university's Figshare repository, so the element of access is simplified.

As there are multiple files of varying types that make up this digital book, the initial premise was to choose between the Figshare repository functions of "project" or "collection", both of which are groups of "items". An item is the deposit of one or more files, possessing a single set of associated metadata, onto a single record. Items can be gathered into either a project or collection, which also have their own metadata for the grouped items. There are benefits and drawbacks for our purposes when it comes to either project or collection, but in order to determine what these were, we needed to create one of each for both selected monographs.

Rameau's Nephew

For *Rameau's Nephew*, the following items were created:

- An item for each of the file formats for the central text files: PDF, XML, EPUB, and MOBI (4 individual items total)
- Items for each of the 13 musical compositions (13 items total, each including a WAV and an MP3 file)
- Items for each of the 5 supplementary texts (PDFs, 5 items total)

The items were created individually, and then grouped together into a Collection, or created within a Project container as a separate workflow.

This was done in order to represent full metadata for the component parts of the book and its essential and supplementary material. One of the findings we have determined in our research so far is that frequently when a digital monograph is preserved, it is often only the main text file, usually a PDF or XML, and not any of the supplementary material, regardless of how “essential” this material may be to understanding the work. This process also allowed us to clearly indicate the connections between all of the materials deposited.

Here some screenshots of the deposited material:

The screenshot shows a digital collection page on the Loughborough University website. The page title is "Denis Diderot 'Rameau's Nephew' – 'Le Neveu de Rameau': A Multi-Media Bilingual Edition". It was posted on 05.10.2021 at 15:22 and is the first published by Open Book Publishers. The page includes a description of the collection, which contains PDF and XML formats of the monograph, along with five textual and 12 musical resources. The authors listed are Marian Hobson, Kate E. Tunstall, Caroline Warman, and Pascal Duc. The page also features a list of authors with their initials in circles: DD (Denis Diderot), MH (Marian Hobson), KT (Kate Tunstall), CW (Caroline Warman), and PD (Pascal Duc). There are buttons for "Follow", "DataCite", and "Read the published paper". The page is part of a collection from Loughborough University.

Denis Diderot 'Rameau's Nephew' – 'Le Neveu de Rameau': A Multi-Media Bilingual Edition

+ Follow Posted on 05.10.2021 - 15:22
First published by Open Book Publishers.

USAGE METRICS [↗](#)
5 views | 0 citations

This collection includes the PDF and XML formats of the monograph, with additional resources including five textual resources and 12 musical resources.

Marian Hobson, Kate E. Tunstall, Caroline Warman, Pascal Duc, Denis Diderot, Rameau's Nephew — Le Neveu de Rameau: A Multi-Media Bilingual Edition.
Cambridge, UK: Open Book Publishers, 2016. <http://dx.doi.org/10.11647/OBP.0098>

Edited by Marian Hobson. Translated by Kate E. Tunstall and Caroline Warman. Music researched and played by the Conservatoire national supérieur de musique et de danse de Paris under the direction of Pascal Duc.

Text © 2016 Marian Hobson, Kate E. Tunstall, Caroline Warman, Pascal Duc
Music © 2016 Conservatoire national supérieur de musique et de danse de Paris

CITE THIS COLLECTION

Diderot, Denis; Hobson, Marian; Tunstall, Kate; Warman, Caroline; Duc, Pascal (2021): Denis Diderot 'Rameau's Nephew' – 'Le Neveu de Rameau': A Multi-Media Bilingual Edition. Loughborough University. Collection.
<https://doi.org/10.0166/FK2.stagefigshare.c.2859034.v1> [Copy citation](#)

<https://doi.org/10.0166/FK2.stagefigshare.c.2859034.v1> [Copy DOI](#)

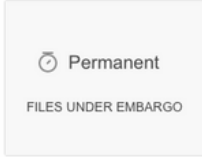
Loughborough University

AUTHORS (5)
DD Denis Diderot
MH Marian Hobson
KT Kate Tunstall
CW Caroline Warman
PD Pascal Duc

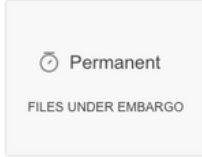
CATEGORIES

21 results found

sort by: Date added



Denis Diderot 'Rameau's Nephew' – 'Le Neveu de ... Monograph posted on 05.10.2021 Denis Diderot



Denis Diderot 'Rameau's Nephew' – 'Le Neveu de ... Monograph posted on 05.10.2021 Denis Diderot



Denis Diderot 'Rameau's Nephew' – 'Le Neveu de ... Monograph posted on 29.09.2021 Denis Diderot



Denis Diderot 'Rameau's Nephew' – 'Le Neveu de ... Dataset posted on 29.09.2021 Denis Diderot



Textual Resources: 1. Goethe's translation of Denis Diderot's ... Dataset posted on 29.09.2021 Marian Hobson



Textual Resources: 2. Denis Diderot's Rameaus Neffe With... Dataset posted on 29.09.2021 Joanna Raisbeck



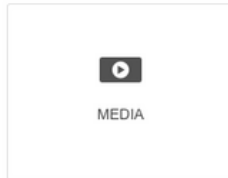
Textual Resources: 3. Anmerkungen Über Personen und ... Online resource posted on 29.09.2021



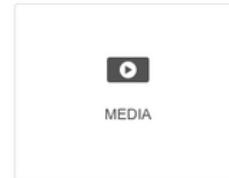
Textual Resources: 4. Nachträgliches Zu „Rameaus Neffe“ Online resource posted on 29.09.2021



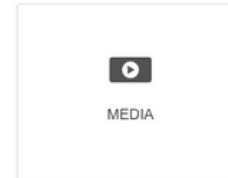
Textual Resources: 5. Goethe's and Schiller's ... Online resource posted on 29.09.2021 Denis Diderot



Musical Resources: 1. François-André Danican Philidor, L'... Media posted on 29.09.2021 François-André Danican ...



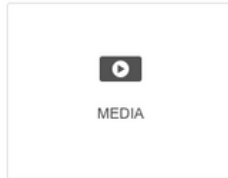
Musical Resources: 2. Jean-Philippe Rameau, Fêtes de ... Media posted on 29.09.2021 Jean-Philippe Rameau



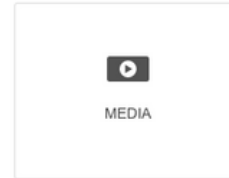
Musical Resources: 3. Jean-Philippe Rameau, Fêtes de ... Media posted on 29.09.2021 Jean-Philippe Rameau



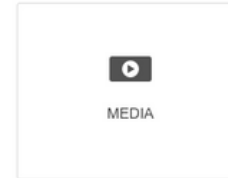
Musical Resources: 4. Jean-Philippe Rameau, Fêtes de ... Media posted on 29.09.2021 Jean-Philippe Rameau



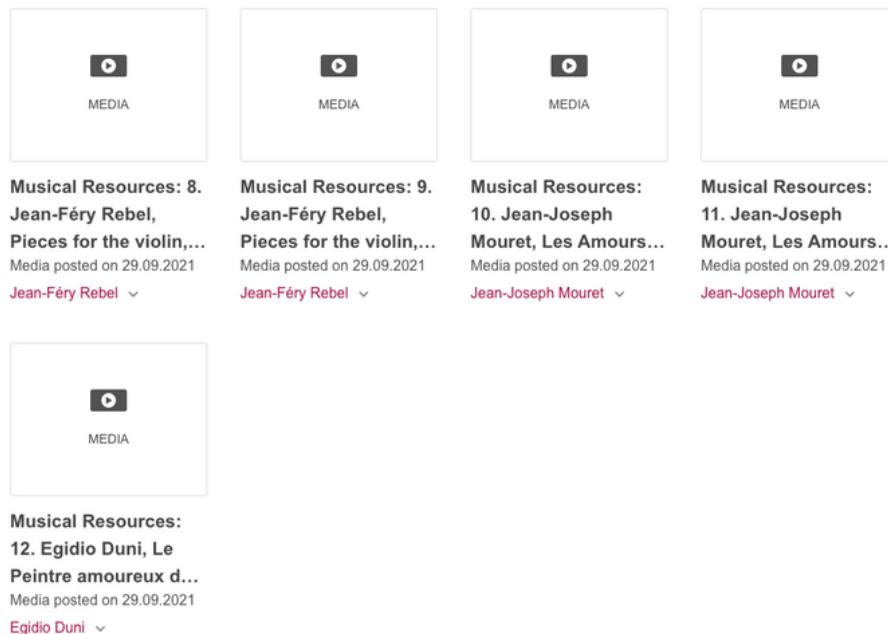
Musical Resources: 5. Pietro Locatelli, Sonata op. VI no. 5, ... Media posted on 29.09.2021 Pietro Locatelli



Musical Resources: 6. Domenico Alberti, Sonata for the ... Media posted on 29.09.2021 Domenico Alberti



Musical Resources: 7. Giovanni Battista Pergolesi, Stabat ... Media posted on 29.09.2021 Giovanni Battista Pergolesi



Overall, for *Rameau’s Nephew*, there were 22 items in each project/collection, with each item having its own unique set of metadata, as well as a set of metadata accompanying the full project or collection.

Image, Knife, and Gluepot

The second volume used in the workflows was Kathryn M. Rudy’s *Image, Knife, and Gluepot: Early Assemblage in Manuscript and Print* (<https://doi.org/10.11647/OBP.0145>). This monograph has fewer supplementary materials but contains a high number of images. Like *Rameau’s Nephew*, there are four different file formats available for the book: PDF, XML, EPUB, and MOBI. OBP also makes HTML versions available on their website for these books, which allows visitors to read the books on the web without downloading. However, the HTML versions were not included in our archiving workflow experimentations, as HTML is not a downloadable “file” in the same sense as a PDF or EPUB⁴⁸, for instance, and hence not “archivable” in the same way either of these would be within an institutional repository. HTML and webpages are often better handled by

⁴⁸ The EPUB format is an archive file of XHTML, but also includes images and other supplementary files, and is supported in e-readers and other software to be opened and read as a book.

webcrawling services such as the Internet Archive’s Wayback Machine, or similar services, and so they are outside the remit of our current work.

For *Image, Knife, and Gluepot*, we made items for each of the four file formats of the main monograph text (PDF, XML, EPUB, and MOBI). Each item was published individually with its relevant metadata, and the four items were gathered into the “project” and “collection” containers, respectively. The below images are from the “project” created for *Image, Knife, and Gluepot*.

The screenshot shows a digital repository page for a project. At the top, there is a navigation bar with the Loughborough University logo, a search bar containing 'Search on Loughborough U...', and user options for 'Upload', 'My data', and a notification bell. The main content area features a folder icon and the title 'Image, Knife, and Gluepot: Early Assemblage in Manuscript and Print'. Below the title, there is a '+ Follow' button and the text 'Published on 07.10.2021 - 11:21 by Miranda Barnes'. To the right, 'USAGE METRICS' shows '5 views'. The page also includes the text 'Originally published by Open Book Publishers.' and 'This collection in includes the PDF and XML formats of the monograph.' The author information is 'Kathryn M. Rudy. Image, Knife, and Gluepot: Early Assemblage in Manuscript and Print. Cambridge, UK: Open Book Publishers, 2019, https://doi.org/10.11647/OBP.0145'. Under 'Chapter Abstracts:', there are two entries: '1. Cut, Pasted, and Cut Again: The Fate of 140 German and Netherlandish Single-Leaf Prints at the Hands of a Limburg Franciscan and a Modern Connoisseur' and '2. A Novel Function for the Calendar in Add. Ms. 24332'. The first entry has a detailed paragraph describing the reconstruction of a manuscript by beguards in Maastricht. The second entry has a short paragraph discussing the significance of an unusual calendar. On the right side, there is a 'MEMBERS (1)' section with a profile for 'Miranda Barnes'.

[cite all items](#)





SHARE

[facebook](#) [twitter](#) [linkedin](#) [email](#)

search content on this project...

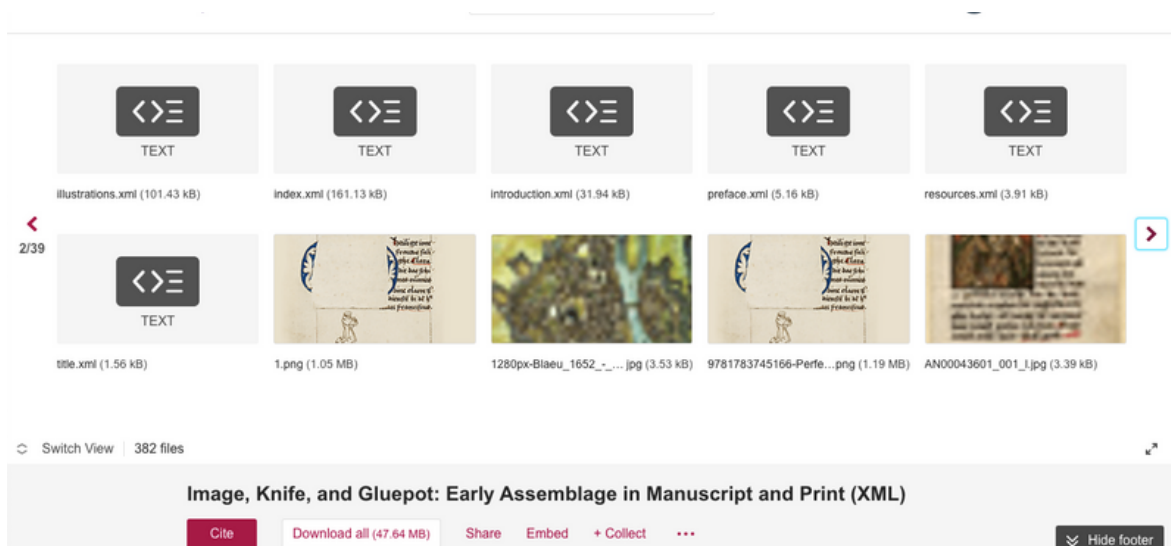
need help? [+ Follow this search](#)

4 results found sort by: Date added

		 Permanent FILES UNDER EMBARGO	 Permanent FILES UNDER EMBARGO
Image, Knife, and Gluepot: Early Assemblage in ... Monograph posted on 07.10.2021 in Loughborough University Kathryn M. Rudy	Image, Knife, and Gluepot: Early Assemblage in ... Monograph posted on 07.10.2021 in Loughborough University Kathryn M. Rudy	Image, Knife, and Gluepot: Early Assemblage in ... Monograph posted on 07.10.2021 in Loughborough University Kathryn M. Rudy	Image, Knife, and Gluepot: Early Assemblage in ... Monograph posted on 07.10.2021 in Loughborough University Kathryn M. Rudy

The XML Item: Preview issues

Because Figshare has the function allowing files to preview in-browser on any live record, the previews of each image file were available to view in the XML item. However, so were icons or preview images all of the other file types within the XML item, which meant some did not preview well due to their size or type, and this led to a slightly confusing presentation of the content. See screenshot below.



The screenshot shows a file list interface with the following items:

- illustrations.xml (101.43 kB) - TEXT icon
- index.xml (161.13 kB) - TEXT icon
- introduction.xml (31.94 kB) - TEXT icon
- preface.xml (5.16 kB) - TEXT icon
- resources.xml (3.91 kB) - TEXT icon
- title.xml (1.56 kB) - TEXT icon
- 1.png (1.05 MB) - Image thumbnail
- 1280px-Blaeu_1652_... .jpg (3.53 kB) - Image thumbnail
- 9781783745166-Perfe... .png (1.19 MB) - Image thumbnail
- AN00043601_001_1.jpg (3.39 kB) - Image thumbnail

At the bottom, there is a summary bar for the XML item: "Image, Knife, and Gluepot: Early Assemblage in Manuscript and Print (XML)". Below this bar are buttons for "Cite", "Download all (47.64 MB)", "Share", "Embed", "+ Collect", and "Hide footer".

As one can see, some of the images, due to their small size, are blurry, and these are also mixed in with other text files, which generally means the contents of the item are difficult to parse. While previewing in-browser as a function is excellent for more straightforward items (such as a single PDF or EPUB, or a small set of high-quality images), for XML the preview function is less a help than a hindrance.

When reviewing the manual workflow experimentations, we considered depositing the Zip file that contained the XML. This option would render as a “file-folder view” of the content, which could be easier to follow: instead of the content being “previewed”, there would just be a list of the contents and how they are organised within the folder. The files wouldn’t be viewable unless downloaded, but they would be present and decipherable, as the following image shows:



The Zip file option was not revisited fully at this point but could be a possible option for use in automated workflows in the future.

External deposit

The second manual workflow experiment was approached from the angle of an academic author publishing with “Loughborough University Press” from an

external position to the university. Theoretically, external users can be invited to a project on Figshare by an internal member of staff, and that external user can deposit material into that Project, once they accept the invitation and create a Figshare account (if they do not already have one). This function works well in the actual Figshare instance. However, as we have been using the “sandbox”/test area to complete these workflow experimentations, the extra layer of security meant that an actual “external” email address couldn’t be invited. (We attempted this, and the expected invitation email was never received despite multiple attempts to various email addresses.)

We worked around this for the sake of completing the manual workflow experimentation by inviting another member of Loughborough University staff into the project. The COPIM colleague performing the workflow experimentations was then given administrative access (with permission) to the Loughborough staff member’s account, allowing them to complete the manual deposit on the Loughborough staff member’s behalf. While this wasn’t exactly “external”, the process was useful, because we came to realise that while items created individually and put into a “collection” could not be added to a “project”, the reverse is possible: if items are created within a project, they can then be added to a collection.

Pros and cons

When weighing up the “project” and “collection” functions for the sole purpose of archiving a monograph, the project function won out, because when a collection is created in Figshare and subsequently published, a DOI is automatically created. While for other purposes, such as creating an online collection of authored/created content, this is ideal, for an already-published monograph it is not. This is because in most cases the original DOI minted by the publisher should be the only DOI for a monograph.

Multiple DOIs will lead to confusion and multiple citations, as well as usage data being obscured. The project function allows for gathering and connecting of

monograph materials, making the archived content available in an open access fashion, while not creating an extraneous and unnecessary DOI. The project function also allows for potential collaboration with external members of small and scholar-led presses, or external authors to a university press.

The reality is, however, that Figshare is only one of several main players in terms of repository software used by universities and libraries. The manual deposit workflow option has not been applied to DSpace or Eprints as of yet, due to access issues. But also, the other reality is that manual input itself has some very recognisable pros and cons.

The benefit of manual deposit and manual metadata input for repository-archived monographs and their supplementary components is the ability to create very specific and thorough metadata for the files, as well as to assure clearly articulated connections between the files, both monograph text and supplementary content. However, there are glaring cons to this pathway, as well.

One primary issue with the entire process is this: manual deposit takes a lot more time (individual/staff resource) as well as requiring technological resource, or expertise. Another major finding from earlier research which contributed to our first [Scoping Report](#) is that small and scholar-led presses have major deficits of resource: financial, staffing, and at times technological expertise. While the COPIM staff member completing the manual deposit workflows was expertly familiar with how to use Figshare repository software, this wouldn't necessarily be the case for every press staff member.

Despite this expert familiarity, the process of depositing both volumes from OBP took approximately three days (though this was creating both the project and collection versions). However, the reality is clear: the process of depositing digital monographs into willing repositories for archiving needs to be automated. Small and scholar-led presses won't be able to spare the staff time to complete a manual deposit process for every monograph, particularly if training is needed beforehand.

This would mean less nuance in some ways. Functions like collections or projects, and elaborate individual metadata for all component parts, wouldn't be possible in precisely the same way via an automated process. Metadata would not be able to be finessed and enhanced as often. But in reality, the time expense of manual deposit would simply be prohibitive to most small presses, meaning archiving in this way simply wouldn't happen. Because many small and scholar-led presses do not yet have any active preservation policy in place, at least one of the options we present must be as simple, straightforward, and as quick as possible.

Summary & concluding thoughts

While ultimately our findings from these manual workflow experiments led us to appreciate the need for automation, there were still some important insights that arose from the work. Grasping the differences in metadata fields (default/existing and custom) within repository systems was one of several. However, many of these understandings have led to more questions, particularly whether the access vs. the archive copy of a complex/enhanced digital monograph should be the same, how connections between monographs and their associated content might be indicated if they are in different online locations, and how to approach the archiving and preservation of linked content. Many of these are still being examined and will be discussed more in future archiving and preservation work package outcomes.

The key finding is the need for some sort of automated option for basic archiving of outputs from small presses that publish open access monographs. In order to figure out what might be possible, the next step was to bring in the help of one of our developer team to perform workflow experiments with automated API deposit. We started with Figshare, as we had access already to the sandbox side of Loughborough University's repository. Our next chapter will address some of the findings from this test workflow experiment.

Chapter 4:

Case Study 2 – Automated Ingestion

Options for computer-assisted repository archiving for small and scholar-led presses publishing open access monographs

By Ross Higman

A [version of this chapter](#) first appeared on COPIM's open documentation site, copim.pubpub.org.

Following on from the manual ingest experiment detailed in Chapter 3, we concluded that a purely manual archiving workflow would be prohibitively time- and resource-intensive, particularly for small and scholar-led presses who are often stretched in these respects. Fortunately, many institutional repositories provide routes for uploading files and metadata which allow for the process to be automated, as an alternative to the standard web browser user interface. Different repositories offer different routes, but a large proportion of them are based on the same technologies. By experimenting with a handful of repositories, we were therefore able to investigate workflows which should also be applicable to a much broader spread of institutions.

The basics of automated ingest

Many websites which allow users to store and view data, such as repositories, offer access to that data via an API (Application Programming Interface). ([COPIM's own Thoth system](#) has two: [a GraphQL API](#), for accessing raw metadata in its database format, and [an Export API](#), for downloading that metadata in the specific formats required by various platforms.) An API defines a standard set of instructions which another computer program can send to it in order to perform certain actions, such as reading existing data from the database, or adding new data to it. Once we know what instructions an API expects, we can develop code to generate and send them, and run that same code quickly and easily to trigger multiple similar actions.

Our desired workflow would therefore be to obtain a book’s metadata from the Thoth API, retrieve all required content files from the URLs specified within it, convert the metadata into the structure expected by the repository API, and then send the files and metadata to the API with an instruction to upload them into the database. The work should then be viewable within the repository in exactly the same way as if it had been manually ingested.

Limitations

The manual ingest experiment deliberately selected books which posed particular challenges for archiving, such as *Denis Diderot 'Rameau's Nephew' – 'Le Neveu de Rameau': A Multi-Media Bilingual Edition* (<https://doi.org/10.11647/OBP.0098>), a complex work with several “additional resources” in both text and audio formats. These resources, external to the main text, are not currently represented within Thoth’s metadata framework, so details about them (such as links to the content) cannot be stored in the Thoth database. Any automated upload at this stage would therefore miss out these files. However, as we are continuously developing Thoth to better fit users’ needs, this is something which can be improved – and this experiment allowed us to identify this desired new feature and raise [an issue to track its development](#).

Another limitation is that while URLs linking to all published digital versions of a book can be stored in Thoth, some of these content files may not be freely retrievable. As discussed in the manual ingest chapter, both of the books selected are available in PDF, XML, EPUB and MOBI formats, as is standard for Open Book Publishers. However, only the PDF and XML are free to access; the EPUB and MOBI must be purchased, and therefore their URLs lead to a paywall. These formats would therefore need to be manually uploaded by a representative of the publisher who had access to the original files (or, alternatively, the publisher would have to reconfigure their paywall to allow access by the automated ingest program).

Figshare

As the manual ingest experiment was conducted via the Loughborough University Figshare repository, this was also where we began when investigating automated ingest. Although Figshare is proprietary software and its code is unfortunately not open-source, it has [a well-documented open API](#), which can be fully accessed by submitting the credentials of an existing Figshare user account. Instructions can be sent to the API over the internet via HTTP requests, with additional data included in JSON format. The API will return information (such as data from the database, or confirmation of successful uploads) in a similar format. We found that most of the actions required for manual ingest could be performed using the API. These included creating “items”, creating “projects” and “collections”, adding items to projects/collections, setting the appropriate metadata on an item, adding files to an item, and making draft items public. Corrections can also be made to records which have already been created, as the API allows updates to metadata, deletion of files, changing of links between items and collections, and so on.

The next task was therefore to write some code which would send the necessary instructions to the API. Figshare provide some example code for interacting with the API in various different programming languages, and this can be found at the top of each subsection in the documentation (e.g. the “Curl”/“Java”/“C#”/etc tabs in the [Public articles subsection](#); the full set of example code can be downloaded as a ZIP file from the “Other” tab). However, this code is automatically generated from the API specification, and needs to be augmented by the programmer in order to be usable. There is also no example code for the programming language Rust, which is what is used by Thoth. The experimentation therefore involved writing [some Rust code from scratch](#) which would check whether a Thoth book record had an equivalent record on Figshare, and then either create a new Figshare record for it or update the existing one, submitting up-to-date metadata from the Thoth database and uploading a sample data file.

Using this code, we were able to successfully interact with the Figshare API, creating and updating basic records which could be viewed in the Figshare user

interface. The two main challenges during development were the complexity of the API, and the specifics of the Figshare metadata format. The API does not make use of any standard frameworks, and instructions which are similar in nature (e.g. adding new metadata vs updating existing metadata) often have slight differences or return inconsistent responses, all of which have to be individually taken into account when writing the code. Many actions also require a multi-step process rather than a single instruction; for example, large files have to be uploaded by sending multiple file parts separately, requiring monitoring the API's response each time before continuing.

Meanwhile, there was the question of how best to represent the broad set of metadata available from Thoth within the Figshare format. This is also relevant to manual ingest, but becomes more apparent when working directly with the two databases. Some correlations are straightforward, e.g. the "title" of a book within Thoth equates to the "title" of a book within Figshare. However, whereas Thoth can store many different types of contributor to a work, such as "Author", "Editor", "Translator", and "Music Editor" (all of which apply to *Rameau's Nephew*), Figshare only accepts "Authors" – and cannot store many of the details available for them in Thoth, such as biographies, or institutional affiliations. Figshare does allow users to create "custom fields" for storing additional metadata, but it is not clear whether these are easily searchable, so while they would add to the completeness of an archive record, they might not enhance its discoverability.

EPrints

The next repository system we investigated was EPrints, having been given access to a test account by the Library at Bath Spa University. EPrints is another popular option for institutional repositories, but unlike Figshare, it is free and open-source. While this is a closer fit with the [ethos of the COPIM project](#), it also increases the likelihood that institutions will be slow to upgrade the software when new versions come out. The EPrints software is also highly customisable, so one institution's API might be very different from another's. Nevertheless, if we focus on developing a workflow for the standard API, and bear in mind that we may need to support

legacy versions, we should be able to cover the majority of repositories that use EPrints.

SWORD

One major advantage of EPrints is that it uses the SWORD protocol (Simple Web-service Offering Repository Deposit). As the name suggests, this is a standardised format for interacting with APIs which was designed with institutional repositories in mind. Since it is a technical standard, there is a lot of support for institutions and developers who want to use it. In particular, the SWORD team offer a wide range of [open-source software libraries](#): bundles of pre-written code which can be integrated into new programs, making development much simpler. The set of programming languages covered by these libraries does not include Rust but does include Python. Python libraries are relatively easy to connect to Rust code; Thoth itself also has [a Python library](#) for accessing its API, so it would not take much work to write a small Python program to transfer data from Thoth to EPrints.

The current version of EPrints, version 3.4, uses SWORD version 2.0 by default. In version 2.0, instructions are sent via HTTP requests as for Figshare, but additional details are formatted as XML (in a variation on the standard Atom Publishing Protocol format). The latest version of SWORD, version 3.0, replaces the XML formatting with a JSON-based format. Future versions of EPrints, as well as other repository systems, might move from using SWORD version 2.0 to version 3.0, so we ideally need to support them both. However, using the SWORD libraries would make this very simple: they do the work of creating instructions for the API in the correct format, and dealing with the responses it sends back, so we would not need to worry about the switch from XML to JSON.

We carried out some basic experiments with the SWORD version 2.0 Python library, and successfully used it to connect to the Bath Spa EPrints API and upload some metadata and a small sample file. These became visible within the EPrints user interface in the same way as records which had been created manually. Attempts to upload the full PDF file of *Rameau's Nephew* unfortunately failed, seemingly due

to its size (around 50MB, whereas the sample file was smaller than 1MB); this would need to be resolved before we could finalise a workflow for automated deposit of standard-sized book files. Although we did not have access to any repository using SWORD version 3.0, we briefly tested its Python library and were easily able to create (but not send) API instructions containing appropriately-formatted metadata.

DSpace

DSpace is another commonly-used repository system, and, again, it is free and open-source. We were able to carry out some tests on the Cambridge University Library “Apollo” repository, which currently uses DSpace version 5, although they are working towards moving to version 7. Both of these DSpace versions offer API access via a number of different methods, one of which is SWORD version 2.0. Having already successfully tested the SWORD version 2.0 Python library on the Bath Spa EPrints repository, we quickly confirmed that the library also worked for connecting to Apollo and uploading basic metadata. Cambridge University Library have agreed to contact us when they have a test site set up for DSpace version 7, so that we can check if this will require any changes.

Internet Archive

Finally, we investigated the possibility of using the Internet Archive as an automated archiving solution. The Internet Archive is not an institutional repository system, but a publicly accessible online archive of all kinds of digital and digitised material, registered as a non-profit in the USA. It is well-known, well-funded and widely used, and allows anyone to create an account and add materials. This could make it a useful alternative for small and scholar-led presses who do not have close institutional connections.

Similar to Figshare, the Internet Archive has developed its own bespoke API, using HTTP requests and JSON content in a custom format rather than adopting standards such as SWORD. However, they also provide [their own Python library](#)

allowing users to easily interact with it. After signing up for a test account, it was simple to use this library to upload sample files and metadata, all of which can be freely accessed and redownloaded by anyone on the open internet (and even some people not on it).

Conclusions

Thanks to the provision of API access and software libraries, it is very easy to set up automated deposit to a number of institutional repositories. While more complex works may need some level of manual intervention to ensure they are correctly represented within an archive, the bulk of small and scholar-led presses' output is low-hanging fruit in this respect. Thoth data is well-structured and simple to systematically convert into the formats required by different APIs, so once set up, an automated workflow can be reliably repeated for large numbers of works. The main question, as for manual ingest, is how best to enter the rich metadata available into a system which may not be well-designed for storing it. Nevertheless, when many small publishers have few resources to devote to archiving and preservation, an imperfect but frictionless workflow is better than no workflow at all.

Furthermore, when viewed through the lens of automated deposit, archiving is not actually so different from dissemination of a work on publication. Both processes require, via some means, the submission of metadata and/or content files to an online platform; [COPIM's dissemination and distribution team](#) is already investigating ways to ease the burden on small and scholar-led presses by automating aspects of the dissemination process. Where distribution platforms offer API access, the workflow might be almost identical to the archiving workflow. These experimentations therefore formed the basis of further work where archiving is centred, alongside distribution, within the publishing process – not left as an afterthought.

The following chapter (Chapter 4) elaborates upon the further workflow development and bulk upload of punctum and Open Book Publishers monographs

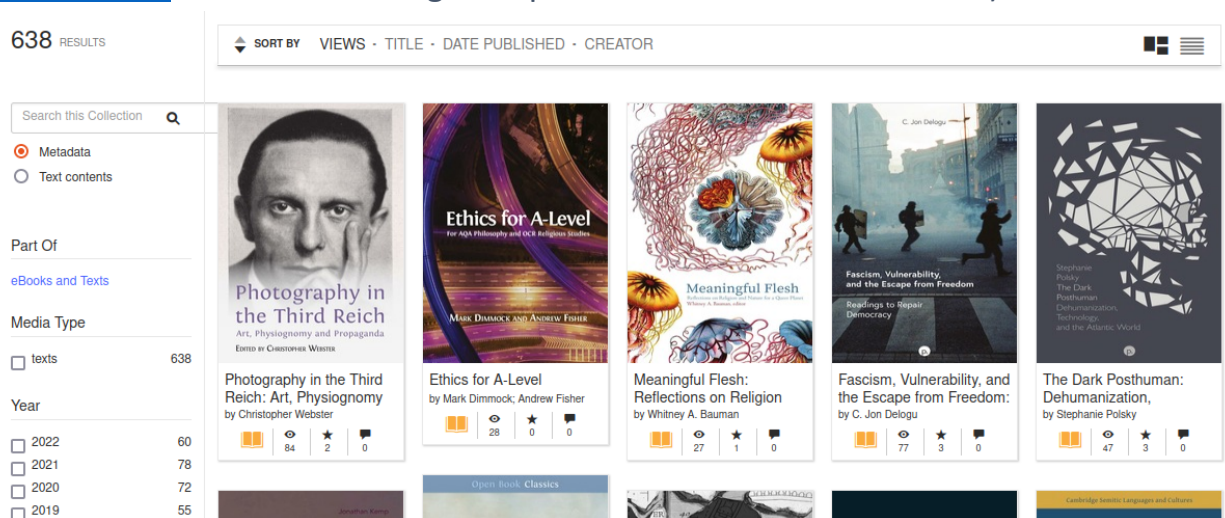
to the Internet Archive, and touches briefly on what is next for the dissemination tool under development in Thoth.

Chapter 5: Development of Dissemination Tool in Thoth

Case studies of content deposit in the Internet Archive and Figshare By Ross Higman

A [version of this chapter](#) first appeared on COPIM's open documentation site, copim.pubpub.org.

Having experimented with integrating [Thoth](#) with various archiving platforms, we concluded that a basic level of automated ingest would be both worthwhile and eminently achievable. In particular, it would tie in with existing plans to develop the [Thoth Dissemination Service](#), a tool for retrieving publishers' files and metadata via Thoth records and submitting them in the required formats to distribution platforms. Our next step was to obtain proof of concept for this proposed service and its use in archiving, which we did with [a bulk upload of over 600 Thoth works to the Internet Archive](#). (The Internet Archive has also [blogged about Thoth integrations](#) and collaborating with punctum books on their site.⁴⁹)



Just some of the uploaded works now present at the [Thoth Archiving Network collection](#), as displayed within the Internet Archive interface

⁴⁹ 'Punctum Books Helps Build Streamlined System for Archiving Open Access Monographs | Internet Archive Blogs', 22 February 2023. <https://blog.archive.org/2023/02/22/punctum-books-helps-build-streamlined-system-for-archiving-open-access-monographs/>.

This chapter will explore the steps taken to accomplish this, providing pointers for anyone looking into implementing a similar system themselves, as well as giving some background for publishers interested in joining the Thoth programme to take advantage of this feature. All of the Thoth Dissemination Service code is available on GitHub under an open-source licence, as is standard for the COPIM project. The chapter will also outline our plans for building on this initial work as we start to develop the [Thoth Archiving Network](#).

Previous work

During initial investigations, we had successfully uploaded temporary test files to the Internet Archive (IA) using the same method which would form the basis of our proof-of-concept workflow. As briefly discussed in the previous chapter, both Thoth and the Internet Archive offer APIs (Application Programming Interfaces) as a simple, standardised way for software programs to interact with their databases. They also both offer open-source software libraries in the Python programming language, packages of “canned code” for performing common tasks which can be utilised when developing new programs instead of writing everything from scratch.

This meant we could quickly write a piece of Python software which would do the following:

- Given the Thoth ID of a work, obtain its full metadata in an easily-digestible format (using [the Thoth Python library](#))
- From the metadata, extract the URL where the PDF of the work’s content can be publicly accessed online
- Use this URL to download a copy of the PDF content file
- Rearrange the work metadata into the format used by the Internet Archive
- Log in to an appropriate IA user account, and send the PDF and the formatted metadata to the Archive to create a new openly-accessible copy of the work there (using [the Internet Archive Python library](#)).

Publisher input

For the proof-of-concept workflow, we decided to perform a one-time upload of real-world files and metadata from publishers' full back catalogues⁵⁰. Open Book Publishers (OBP) and punctum, as key COPIM partners, elected to participate in the upload, and the development team consulted with them throughout in determining their approach. The first decision was the choice of platform. As mentioned above, the original investigations focused mainly on institutional repositories as archiving platforms; similarly, the Thoth Archiving Network aims to bring together a group of institutions who are willing to host open-access works from smaller publishers in their repositories. The Internet Archive (IA) is not an institutional repository, but the workflows are similar, and the publishers involved agreed that it would be a good place to have accessible, discoverable archive copies of their works hosted as a first step while waiting for agreements with individual institutions to be finalised.

Next, we discussed exactly which files should be included in the upload, and which IA metadata fields should be filled out with which Thoth metadata elements. We decided to prioritise simplicity in our approach, only uploading the PDF version of a work (even though OBP standardly produces editions in additional digital formats, some of which are closed-access and therefore require a degree of consideration when archiving), and bypassing concerns about metadata loss by simply uploading a full Thoth metadata file alongside the content file. While we also did our best to ensure that major IA metadata fields were appropriately filled out to improve discoverability, we were mindful that third-party interfaces such as IA's could well change over time, and we should therefore not put disproportionate effort into conforming to them at the expense of making progress elsewhere.

This resulted in [the creation of a new Thoth Export API output option](#), using JSON

⁵⁰ These back catalogues are openly available via Thoth and their contents can be viewed in varying levels of detail at the [GraphQL API Explorer](#). Try starting with the query `{books(publishers: ["85fd969a-a16c-480b-b641-cb9adf979c3b"])}` to explore the OBP catalogue. [More examples can be found here](#).

format to be easily readable by both humans and computers, and added another step to the workflow described above:

- Using the work’s Thoth ID, download its JSON metadata file from the Thoth Export API, to be included in the eventual upload to IA (this is easy to achieve using basic Python, as the Thoth Export API is uncomplicated).

Platform-specific considerations

Contents

PREFACE			
1. Exam Specification Details		1	
2. Book Structure		1	
References		2	
INTRODUCTION		3	
1. Philosophy, Ethics and Thinking		3	
2. Respecting Ethics		3	
3. The A-Level Student		4	
4. Doing Ethics Well: Legality versus Morality		5	
5. Doing Ethics Well: Prudential Reasons versus Moral Reasons		5	
6. Doing Ethics Well: Prescriptive versus Descriptive Claims		6	
7. Doing Ethics Well: Thought-Experiments		6	
8. Doing Ethics Well: Understanding Disagreement		7	
Summary		7	
Questions and Tasks		8	
References		8	
PART I NORMATIVE ETHICS			
CHAPTER 1 UTILITARIANISM		11	
1. Utilitarianism: An Introduction		11	
2. Hedonism		11	
3. Nozick's Experience Machine		12	
4. The Foundations of Bentham's Utilitarianism		13	
5. The Structure of Bentham's Utilitarianism		14	
6. Hedonic Calculus		15	
7. Problems with Bentham's Utilitarianism		16	
8. Mill's Utilitarian Proof		20	
9. Mill's Qualitative Utilitarianism		21	
10. Mill's Rule Utilitarianism versus Bentham's Act Utilitarianism		22	
11. Strong versus Weak Rule Utilitarianism		23	
12. Comparing the Classical Utilitarians		24	
13. Non-Hedonistic Contemporary Utilitarianism: Peter Singer and Preference Utilitarianism		24	
Summary		26	
Common Student Mistakes		26	
Issues to Consider		26	
Key Terminology		27	
References		28	
CHAPTER 2 KANTIAN ETHICS		31	
1. An Introduction to Kantian Ethics		31	
2. Some Key Ideas		32	
3. Acting for the Sake of Duty and Acting in Accordance with Duty		33	
4. Categorical and Hypothetical Imperatives		34	
5. The First Formulation of the Categorical Imperative		36	
6. Perfect and Imperfect Duties		37	
7. Second Formulation of the Categorical Imperative		38	
8. The Third Formulation of the Categorical Imperative and Summary		38	
9. Kant on Suicide		39	
10. Problems and Responses: Conflicting Duties		42	
11. Problems and Responses: The Role of Intuitions		43	
12. Problem and Responses: Categorical Imperatives and Etiquette		43	
13. Problems and Responses: The Domain of Morality		44	
Summary		45	
Common Student Mistakes		45	
Issues to Consider		45	
Key Terminology		46	
References		47	

(6 of 264)

Ethics for A-Level
by Mark Dimmock; Andrew Fisher

Favorite Share Flag

One of the uploaded works open in the feature-rich Internet Archive BookReader

One notable feature of the Internet Archive is that, as a very large platform geared towards ease of access to content, it has many automatic processes in place for enhancing uploaded files. When a simple PDF is submitted, by default the Archive

[derives](#) multiple additional files from it, such as a thumbnail image to represent it across the site, a version which can be read in the web browser using the Archive's own [BookReader](#), and basic text versions enabling screen-reading for visually impaired readers as well as full-text searching (created using [OCR](#)). The publishers agreed that these derived formats were beneficial for making works more discoverable and accessible to users, although there were some unexpected effects.

Firstly, one of the derived formats is an EPUB version, a potential concern for publishers such as OBP who produce and sell their own EPUB versions of published works as alternatives to the free PDF versions. However, on inspection, the IA-created EPUB is a very utilitarian document based on the OCR text (with all its inevitable mis-scanning) and acknowledges throughout that it has been automatically generated and may contain errors. It is clearly aimed mainly at users who prefer EPUB readers over PDF viewers for reasons that outweigh the reduction in quality (such as smaller file size), and those who want a well-formatted publication thoughtfully tailored to the EPUB standard will still opt for the official publisher's version.

A more intriguing issue was that when the Archive recognised an uploaded PDF, with its publisher-provided metadata, as representing a published book with an entry in a catalogue such as [WorldCat](#), it would attempt to enhance it by pulling in metadata from said catalogue. While this could sometimes be useful, correctly identifying and adding details such as OCLC numbers which had been omitted from the Thoth record, it also sometimes overwrote accurate, detailed metadata with poorer-quality information – replacing a full publication date with just a year, appending an “[Author]” tag to an author's name, or dropping some keywords. This could be avoided by turning off the “derive” option altogether when submitting the work, but this meant we lost the other benefits of derived files as discussed above. When we contacted the Archive to ask if there was a way to continue creating derived files but prevent source metadata being overwritten, they were responsive and helpful. They acknowledged that this was an issue, explained that it could be resolved by enabling an advanced feature, and suggested that we set up a

[collection](#) where this feature could be enabled by default for all submissions. We agreed and they created the [Thoth Archiving Network collection](#) for us, which worked exactly as planned, while also providing a convenient presence for Thoth on the Archive.

Scaling up

Once we had finalised the appropriate process for submitting a single work to IA, including the fine details of source files, metadata mapping and post-upload processing, it was time to extend this process to handle large numbers of works. At a basic level, this would simply require taking the original Python program and running it multiple times, each with a different Thoth work ID; the logic would be identical on each run, so the submissions would be uniform. We just needed to obtain the appropriate set of work IDs to input to the program. Fortunately, the Thoth API is very flexible, so it was easy to write a supplementary Python script which would ask it for a list of all work IDs:

- by the opted-in publishers
- marked as Active (i.e. complete and published)
- excluding book chapters (i.e. only standalone parent works)
- sorted from least to most recently published (for convenience, to give the collection some coherence and help us to track the progress of the bulk upload)
- separated into two sections, one for each publisher (as above).

However, another consideration was that as a task gets bigger, it becomes increasingly important to make the program robust. We would be gathering and submitting a large amount of data for a large set of works; on each submission, there were many points at which the attempt might fail. For example, we might have incorrectly set the login credentials for the IA user account; we might start trying to upload a work then discover that necessary information (such as the URL of its PDF) was missing from the Thoth record; we might simply have bad luck and attempt to submit a work at a time when the Archive was already trying to process

vast numbers of other submissions, leading it to ask us to try again later. It was therefore important to identify all of these possible points of failure and tell the program how to deal with them (e.g. if asked to try again later, it would understand the request and do just that, rather than giving up and producing an error message).

The final step was to ensure that the program would clearly communicate the results of every upload attempt in a way that could be easily read and referenced by the human running it – because every “automated” process requires at least a small amount of manual handling. In our case, if any upload failed, we wanted to know that this had happened and what had caused it, so that we could investigate the problem and potentially try again. This required writing clear and detailed error messages for the program to write out to a log file at each point where a failure might occur.

```
6 INFO:2022-12-13 16:55:02,415: Beginning upload of 7fbc96cf-4c88-4e70-b1fe-d4e69324184a to InternetArchive
7 ERROR:2022-12-13 16:55:08,166: No PDF Full Text URL found for Work
```

Log messages show the progress of an upload and explain why it failed

The results

When we actually ran the finished program and attempted the bulk upload, only seven works encountered failures, out of a total of 640 works identified as eligible for archiving from the two publishers’ back catalogues, dating from as far back as 2008. Of these, one was a temporary error due to the Archive being overloaded, which was automatically retried and then succeeded; the rest failed due to lacking PDF URLs, or having PDF URLs listed which did not actually link to a PDF. On discussion with the publishers, two of these were found to be legacy print-only publications, therefore exempt from our digital archiving attempt, and the rest just needed quick corrections to their Thoth records before they could be resubmitted, this time successfully. The full upload process took less than eight hours.

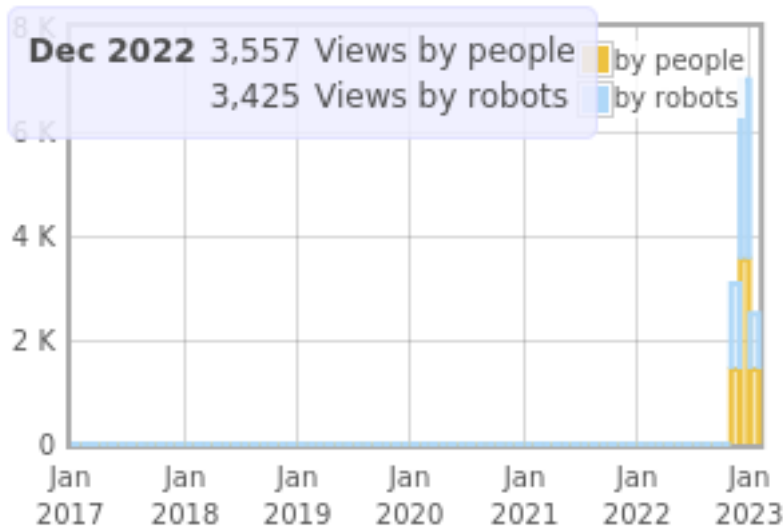
At a glance, and based on some spot checks, the bulk of the works’ files and metadata appear to have been uploaded correctly, and they are well-presented in

terms of derived images and searchable/filterable details. No checks were performed on the completeness or accuracy of the metadata prior to upload, as this is a workflow in which the publisher assumes full responsibility for the correctness of the Thoth record. As discussed in the previous chapter, any “additional resources” which are considered part of the work as a whole but not included in the PDF (such as accompanying videos hosted on YouTube) will not have been archived by the process, as it is “one size fits all” rather than considering the curation needs of individual works.

While these are limitations from an archiving perspective, they help to make the workflow almost entirely automatable, a boon for the resource-strapped smaller publisher, providing an acceptable “first line of defence” for those with few or no other archiving or preservation solutions in place. There is also the option for publishers to examine specific works in the Archive and make manual enhancements to them at any point following upload, so even for a work which is known to be particularly complex, the basic automated upload is a helpful first step.

VIEWS

Total Views 12,557 (Older Stats)



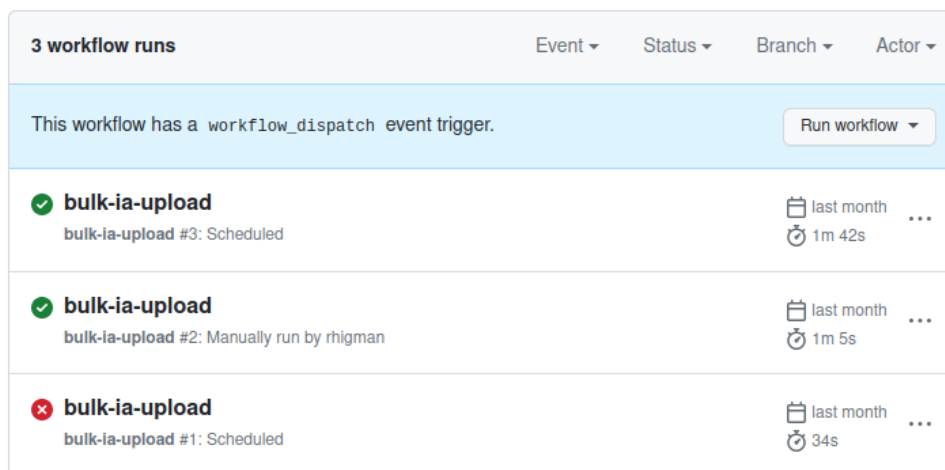
IA stats chart for the Thoth Archiving Network collection

Two months after the upload, the 638 works in the collection had also amassed

over 12,000 “views” between them (half by automated web crawlers, and half by real people), proving that the Internet Archive is a valuable platform not only for archiving, but also for dissemination.

From manual to automated

Following this successful manually-triggered one-time upload, we then worked to set up recurring automated uploads of newly-published Thoth works. The ideal workflow would be to automatically submit a work to IA as soon as the publisher marked it as Active within the Thoth system; a similar process could be particularly useful for submissions to distribution platforms, which are more time-sensitive. As an interim step, we implemented recurring periodic “catch-up” uploads, where a modified version of the one-time process finds all works published since the last upload (i.e. in the past month), and submits them in a much smaller “bulk” upload. The main additional component in this workflow was a [GitHub Action](#). All of the code used in the one-time upload is already publicly hosted on the GitHub platform, and GitHub Actions provide a way to run programs via the GitHub system (rather than on a personal computer), either manually or on a set schedule. The results of Actions are also clearly displayed on the GitHub dashboard and given as email notifications. This allows the Thoth project manager to quickly identify any failures and take appropriate action, just as was done during the bulk upload.



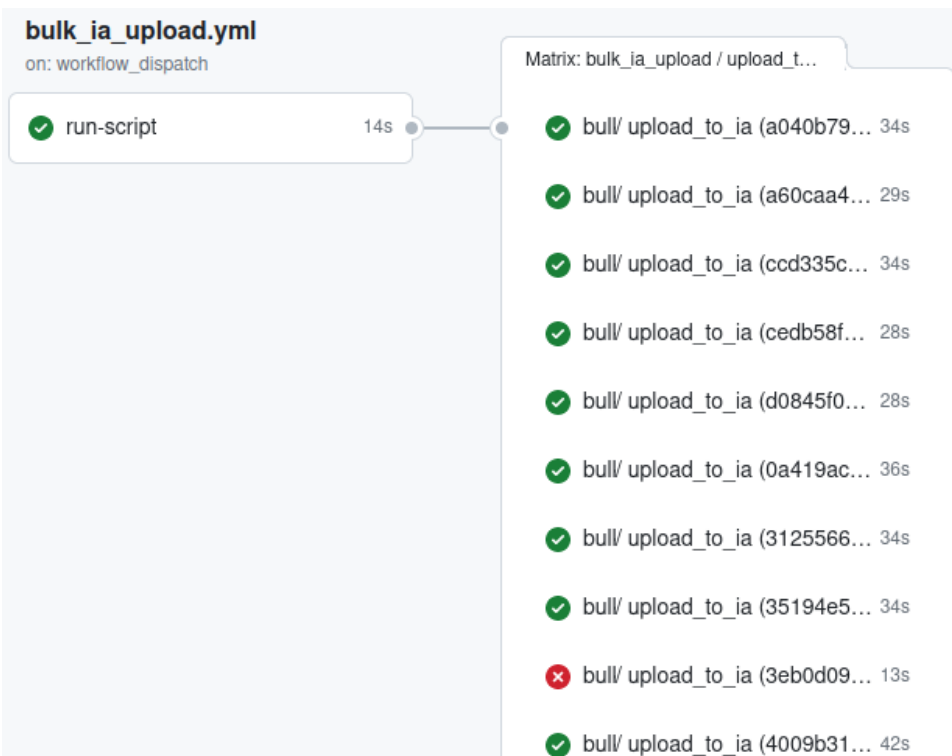
The screenshot shows a GitHub Actions dashboard for a workflow named 'bulk-ia-upload'. At the top, it indicates '3 workflow runs' and provides filters for Event, Status, Branch, and Actor. Below this, a light blue banner states 'This workflow has a workflow_dispatch event trigger.' with a 'Run workflow' button. The main content is a list of three workflow runs:

Run ID	Status	Event	Time
bulk-ia-upload #3	Scheduled	Workflow Dispatch	1m 42s
bulk-ia-upload #2	Manually run by rhigman	Workflow Dispatch	1m 5s
bulk-ia-upload #1	Scheduled	Workflow Dispatch	34s

[Dashboard](#) for the recurring “bulk” IA upload GitHub Action

To create the required GitHub Action, we wrote [a YAML file](#) which acts as instructions to GitHub to perform the required tasks on a set schedule. GitHub Actions have a healthy ecosystem of pre-built tools to help with standard tasks (similar to the software libraries discussed above), such as building and running programs from the raw GitHub code, therefore reducing the development effort needed. The final piece of the puzzle was [a short Python script](#), expanded from the one used for the initial bulk upload, which obtains the IDs of all books from the opted-in publishers marked in Thoth as Active, then compares them to the set of IDs already present in the Internet Archive collection, to determine which have been newly published and need to be uploaded. The GitHub Action can then pass this set of new work IDs to the main dissemination program, seamlessly triggering the necessary uploads.

This process is currently scheduled to run shortly after midnight on the first day of every month, and at the time of writing, it has run twice. In the first month, it failed due to an error in setting up the IA login credentials (as can be seen from the red cross symbol in the dashboard screenshot above!). However, once this was rectified, it was easy to manually re-run the process from the dashboard, and this time, all except one of the uploads completed successfully. As the Action is set up in a modular way, with separate steps for running the Python script and then performing each of the uploads, it was easy to see where the process had failed. We could even click directly into the log for the failed step, immediately showing us that the issue was a user-entered PDF URL which did not link to a PDF. Again, after the publisher had corrected this, it was easy to manually re-run just the failed step and successfully trigger the last remaining upload. In the second month, the whole process completed successfully, with no manual intervention needed.



More detailed dashboard display of the initial failed run, where all steps except one succeeded

Building on the foundations

Now that the infrastructure is in place for automated upload to a single platform, it is very easy to extend the code to upload to additional platforms which use similar processes. We have started on this by reworking the test code previously used to connect to Loughborough University’s Figshare repository to fit in with the new Thoth Dissemination Service structure. By uploading a representative sample of books to this repository for public access, we verified that the new Figshare code module was ready to be integrated into the existing automated workflow.

As they had been used in all of the previous workflow experimentations within Figshare, we continued using *Rameau’s Nephew* and *Image, Knife, and Gluepot* from OBP for this proof of concept. Firstly, projects were created for the two books by sending instructions to the Figshare API, targeting Loughborough University’s public instance this time, rather than the test instance. Items were then created within the projects for both the PDF and XML versions of both books, including the

full metadata file from Thoth in JSON format along with the book file in each item, to ensure metadata was not lost where it was not supported by Figshare's data structure.

Metadata sent directly to the repository via API was kept fairly simple, using only metadata that would match existing fields within the repository. This reduces the burden on repository managers by avoiding the need for them to create custom fields (which could also lead to difficulties and metadata loss when upgrading the repository software). The presence of the full JSON metadata file could allow automated completion of new metadata fields if the repository structure changes in future. Figshare also hope to implement full-text searching in due course, which would support discoverability, as relevant terms would be found in the JSON text despite not appearing in the standard metadata fields.

The projects and items were all created successfully, and once they were complete, a publishing request was sent for each of the four items in the two collections. Once the items were published, the project could be published and finally the two books were archived within Loughborough University's Figshare.

The two projects can be viewed as public records at the following links:

[Denis Diderot's 'Rameau's Nephew' \(Ed. Marian Hobson\)](#)
[Image, Knife, and Gluepot \(Kathryn M. Rudy\)](#)

As projects do not automatically generate a DOI, as collections do in Figshare, handles were created instead in order to provide some type of persistent identifier for the project records.

While COPIIM draws to a close, we present these instances of archiving subsets of open access monographs from two publishers as success of our workflow experimentations and development. However, there is more to be done going forward within the Open Book Futures project. Next we will be implementing recurring automated uploads to Figshare as we have put in place for the Internet

Archive, starting by depositing the entire Open Book Publishers back catalogue, and possibly that of other publishers within Thoth. Reports of our progress will be forthcoming in due course.

Chapter 6:

Enhanced vs. Complex Digital Monographs: Implications for Archiving & Preservation

Considerations for complex & experimental monographs

As the use of new digital technologies to publish scholarly works has rocketed forward in recent years, with new potentialities introduced on a perpetual basis, the preservation of these works has lagged behind. The scholarly opportunities presented by internet-enabled interactivity are significant, and it is important for the digital preservation world to examine options for these works, so they are not lost or left behind. The first step is considering the variety of these works and what they involve.

Worth noting in this chapter is that there does not yet appear to be a standard definition for terms for the varieties of enhanced and experimental monographs, much like many other terms related to digital publishing, preservation, and open access. We will address some of the different terms and their various contexts and consider those that may most closely fit our use.

Enhanced, Complex, Experimental

Though the page is no longer live, Emory University (Atlanta, Georgia, USA) had a useful differentiation between “enhanced” digital monographs and “complex” digital monographs on their guides to digital and open access publishing. (An archived version of the page can be found on [The Wayback Machine](#), from the Internet Archive.)



Figure 13 - Photo by [FLY:D](#) on Unsplash
([License](#))

Emory defines an “enhanced” digital monograph as similar to traditional OA monographs, in that they are structured like conventional books,

but they “take advantage of the online environment to extend the functionality of the digital edition.” Enhanced monographs will embed hyperlinks and “also integrate audio and video clips, dynamic maps, or interactive data visualizations that cannot be included in the print edition.”⁵¹ This would place the books we considered as part of our workflow experimentations in earlier chapters, from Open Book Publishers, as enhanced OA monographs, due to the way they include and incorporate external content. Emory also mentions that enhanced OA monographs often have a print book counterpart, which OBP does offer, with QR codes for separate online access to embedded or linked content.

“Complex” digital monographs, according to these pages, are “born-digital publications that cannot be replicated in print form.” This is because complex monographs “rely on multimodal content and the web’s interactive nature to create a distinctive reading experience that often deviates from the linear structure of a printed book.”⁵² While there is a basic set of software typically used to create and publish enhanced monographs, complex monographs use a wide array of digital tools and platforms. Both authors and publishers are actively experimenting with possibilities in this fairly new publishing modality, which means there are no set standards or workflows, though research is being done to address this (see the [Embedding Preservability](#) project, more below).

The complex monograph most closely aligns with what is being called experimental publishing within the COPIM Project, specifically [Work Package 6](#). This work package has among its aims to “align existing open source software, tools, workflows and infrastructures for experimental publishing with the workflow of open access book publishers.” Colleagues in WP6 have co-developed a set of experimental academic books with the scholar-led presses Open Humanities Press, Mattering Press and Open Book Publishers, who are also partners on the project.

⁵¹ ‘What is a digital monograph?’ Emory University. Originally accessed at: <http://www.fchi.emory.edu/digitalpublishing/tome/monographs.html>. (2021) Archived at: <https://web.archive.org/web/20221013143330/http://fchi.emory.edu/digitalpublishing/tome/monographs.html> (2023)

⁵² Ibid.

These experimental books fall into the following package-defined categories: combinatorial books, data books, and computational books.

Combinatorial books

Combinatorial books are part of the Gathering Flowers pilot within Work Package 6, and these were created by “revisiting and rewriting...books within the OHP catalogue as a means of generating radical new responses to them.”⁵³ These open access books from the Open Humanities Press are licensed for open distribution and reuse via CC-BY licenses. The book re-created from those chosen and remixed for the project borrowed from both contemporary cut-up and remix practices and those of medieval “commonplace books”. This was done with an aim to examine and critique academic publishing and scholarly writing practice. The combinatorial book would also fall under the larger category of “remixed books”, defined by the typology of books also published by the work package as: “Books that consist of previously published materials that are remixed, reused or rewritten into a new publication (which often itself is open for remix again too).”⁵⁴

Computational books

A computational book, as defined by WP6, “combines human-readable text with computational functionality”⁵⁵ and “can contain audio and video objects, 3D models (tested using [.obj files](#) and via embedded [.stl viewers](#)), datasets from linked open data repositories (tested using [SPARQL queries](#) against [Wikibase](#) instances), and media like images pulled in via linked open data queries. This allows for

⁵³ Adema, Janneke, Gary Hall, and Gabriela Méndez Cota. ‘Combinatorial Books - Gathering Flowers - Part I’. *Community-Led Open Publication Infrastructures for Monographs (COPIM)*, 28 April 2021. <https://doi.org/10.21428/785a6451.d3ecc6cc>.

⁵⁴ Adema, Janneke, Tobias Steiner, Simon Bowie, Marcell Mars, and Tobias Steiner. ‘Part 2: A Typology of Experimental Books’. *Community-Led Open Publication Infrastructures for Monographs (COPIM)*, 29 January 2021. <https://doi.org/10.21428/785a6451.cd58a48e>.

⁵⁵ Bowie, Simon. ‘A New Model for Computational Book Publishing’. *Community-Led Open Publication Infrastructures for Monographs (COPIM)*, 14 March 2023. <https://copim.pubpub.org/pub/computational-book-model/release/1>.

publications that link directly with linked open data repositories such as Wikidata and pull data directly from those sources.”⁵⁶

One of the collaborative pilots, X-Sketchbook, is for example using [Jupyter Notebooks](#). The Jupyter Notebook app contains computer code, such as Python, as well as traditional rich text elements, such as paragraphs, links, and figures. The documents created are both human-readable and machine-executable. This software is traditionally used more in STEM fields, but in this project, commands are sent to databases of artwork to automatically retrieve images and metadata. For more information about the work in Experimental Publishing within COPIM, please see links to the reports published by Work Package 6 below.

Reports:

[Books Contain Multitudes: Exploring Experimental Publishing](#) (A three-part scoping report, 2021)

[A workflow for Combinatorial Books](#) (2022)

[Implementing a Workflow for Combinatorial Books](#) (2022)

[Computational Publishing Pilot Project. Introducing Our Partners and Communities](#) (2022)

[A New Model for Computational Book Publishing](#) (2023)

Preservation challenges

The very nature of experimental publishing presents various challenges to the concept of digital preservation. Quite often the aim of experimental publishing is in fact to reconsider what a book is, and to break new ground using innovative digital programs and capabilities. There can be infinite revisions and recreations that result from ongoing collaboration or the effects of reader interactions, and the experimental book as a “live” document is notoriously difficult to preserve. Questions around essential versioning vs. infinite changes have been discussed within our WP6 and WP7 meetings, as has the prickly query of what to document if an experimental work is designed to disappear.

⁵⁶ Ibid.

The reality is that digital preservation, as it currently exists, requires a “fixed” version of any eBook or monograph to preserve it. The typical file package for a traditional digital OA monograph would include the book (in either one file format, such as PDF, or several, if these are produced) along with a metadata file. Where external content is essential to the work, this could be packaged for inclusion, or web-archived, but there are challenges around the external content depending on licensing (e.g. open datasets vs. images pulled from online museum collections; see more in copyright discussion in Chapter 2). There are also metadata concerns around supplementary or additional content, though there is work towards some solutions taking place.

Conversations between Work Packages 6 and 7 have raised some possibilities for experimental works and preservation. Firstly, the process will have to be nuanced and include dialogue early on between author and publisher, deciding what is important and essential to preserve. Secondly, as standard publishing and preservation is built around an “output”, consider what the output is, and how one might also preserve documentation of the methods or processes alongside the output. Thirdly, an alternative option could be to archive/preserve not the output, but the input files and documentation of method. These are, however, just possibilities, and need further testing and discussion. One other point raised is that the scholarly argument of the work may define in individual circumstances what is essential to preserve, and this may look different in a variety of scenarios.

Embedding Preservability & Preserving New Forms of Scholarship

Preserving New Forms of Scholarship

The Preserving New Forms of Scholarship project was an Andrew W. Mellon Foundation-funded project at New York University Libraries, the first of two projects led by the Digital Library Technology Services (DLTS) unit investigating the preservation of complex scholarly outputs. The PNF team included a group of digital preservation institutions, libraries, and university presses collaborating in the study of dynamic outputs. Twenty complex works were examined and tested in preservation workflows to determine a set of guidelines to enhance the

preservability of these complex works of scholarship, which were [published in September 2021](#).⁵⁷ A full report on the creation of these guidelines and the work of the project was also [published the same year](#).⁵⁸

Embedding Preservability

The Embedding Preservability project, funded in 2021, is moving forward with the progress made in the Preserving New Forms project to examine ways these foundational guidelines can be incorporated earlier in the process of creating new forms and experimental publications. In this project, the aim is to support publishers who are involved in the creation of complex, experimental works to make choices in their design of the work that will facilitate effective digital preservation at scale. The project also aims to do so without the sacrifice of functionality, often a major challenge with innovative scholarly works. A team of preservation experts will be “embedded” within a subset of publishers to learn and assist with publisher workflows and technology decisions within the publication process, from the beginning.

Key links:

[Guidelines for Preserving New Forms of Scholarship](#) (NYU, 2022)

[Embedding Preservability Project](#) (NYU)

Portico involvement: [Link](#)

Emulation

Emulation is a familiar requirement for those trying to preserve software but is a relatively new concept in the sphere of digital monograph preservation. Emulation is defined as follows: “a component of a digital preservation strategy in which obsolete file formats are rendered accessible by replicating their original digital or hardware environment.”⁵⁹ The DPC cites emulation as “particularly useful for

⁵⁷ Greenberg, Jonathan, Karen Hanson, and Deb Verhoff. ‘Guidelines for Preserving New Forms of Scholarship’, September 2021. <https://doi.org/10.33682/221c-b2xj>.

⁵⁸ Greenberg, Jonathan, Karen Hanson, and Deb Verhoff. ‘Report on Enhancing Services to Preserve New Forms of Scholarship’, December 2021. <https://doi.org/10.33682/0dvh-dvr2>.

⁵⁹ ‘SAA Dictionary: Emulation’. Accessed 29 March 2023. <https://dictionary.archivists.org/entry/emulation.html>.

complex objects with multiple interdependencies,” such as games, but as we have already explored, experimental academic works are often constructed in this way.

Emulation typically allows for the mimicking of older, obsolete, or no longer supported software within a generated environment. This is often used to access older versions of files, so they look the way they were intended (which is especially important concerning data). Emulation could also be a potential solution for the preservation of certain types of experimental books where the functionality, look, and feel of the original work remain essentially important to future use and access. Though the various software and platforms that are used in experimental works may be new or current when the work is created, likely at some point these could become obsolete or unsupported. While COPIM’s exploration of this remains nascent, the Embedding Preservability project has engaged with EaaS, or the Emulation-as-a-Service-Infrastructure project, who has helped consider some of these questions alongside publishers and preservation archives.⁶⁰

The EaaS ([Emulation as a Service Infrastructure](#)) offers a scalable model which allows organisations to access provided emulated environments, and also allows the organisation to deliver these environments to users. This simplifies access to digital objects that are archived or preserved by the organisation while not requiring the users to attend a workstation in person. There is a flexible web service API provided so digital preservation workflows can be tailored for purpose. COPIM has met with EaaS colleagues with the hope of testing some potential environments and plan to continue this conversation into future endeavours.

Within the Preserving New Forms project, Portico worked to successfully preserve two books within the EaaS platform.⁶¹ The PNF team performed three “sprints” with monographs of varying complexity, working from least to most complex. The

⁶⁰ Verhoff, Deb, Hanson, Karen, and Jonathan Greenberg. ‘Preservation strategies for new forms of scholarship.’ *Conference Proceedings - IPres 2022*, 1 December 2021. <https://ipres2022.scot/conference-proceedings/>. p. 81 – 88.

⁶¹ Greenberg, Jonathan, Karen Hanson, and Deb Verhoff. ‘Report on Enhancing Services to Preserve New Forms of Scholarship’, December 2021. <https://doi.org/10.33682/0dvh-dvr2>. p. 31-32.

two selected books that were preserved in this way, via encapsulation on a virtual machine, involved multiple external dependencies, and when preserved using a web-harvested version did not meet the publisher’s minimum requirements for preservation where the dynamic experience is essential to retain. While this option provided a number of challenges for preserving at scale for the digital preservation archive, it was acknowledged that “If successful, this may be the most efficient approach to preserving the experience of some of the most complex works from publisher platforms.”⁶²

While there are no set standards or workflows, and the pathway to scalable preservation solutions for experimental books are in flux at this time, there is promising and ongoing work that is sure to advance the possibilities for effectively archiving and preserving these creative and nonstandard works of scholarship.

⁶² Verhoff, Deb, Hanson, Karen, and Jonathan Greenberg. ‘Preservation strategies for new forms of scholarship.’ *Conference Proceedings - IPres 2022*, 1 December 2021. <https://ipres2022.scot/conference-proceedings/>. p. 81 – 88.

Chapter 7:

A Brief Introduction to the Thoth Archiving Network

A community solution for open access monograph archiving

A [version of this chapter](#) first appeared on COPIM's open documentation site, copim.pubpub.org.

As discussed in several of the previous chapters, the Thoth Archiving Network has been created in response to the needs of small and scholar-led presses who make up much of the “long tail” of publishers without an active preservation policy in place, putting their significant contributions to the scholarly record at risk of disappearing should they cease to operate or fall prey to technical failure. While large-scale publishers have existing agreements with digital preservation archives, such as [CLOCKSS](#) and [Portico](#), the small press often languishes without financial or institutional support, [alongside challenges in technical expertise and staff resource](#). The Thoth Archiving Network would not solve every issue, but it would be an initial step towards essential community infrastructure, allowing for presses to use a push-button deposit option to archive their publications in multiple repository locations. This would create an opportunity to safeguard against the complete loss of their catalogue should they cease to operate.

The Thoth Archiving Network is being developed as part of the Thoth Dissemination System, which has in turn grown from the Thoth Metadata Management System created as part of COPIM's Discoverability Work Package 5. As detailed in Chapter 4, the archiving deposit component of Thoth will be built as part of their larger dissemination service and will allow automated deposit of book files and metadata via pushbutton functionality into multiple repositories on the network. Presently there are three institutional repositories who have agreed to join the pilot phase of the Thoth Archiving Network, which will initiate at the start of the Open Book

Futures project. Two of these repositories are located within the UK, and one is located on the west coast of the United States.






So that we could begin to scope out both interest and potential barriers to joining the Thoth Archiving Network for institutional repositories within the UK, we held an online workshop, inviting attendees from the UKCORR (United Kingdom Council of Open Research and Repositories). Our Thoth Archiving Network workshop was held virtually on Tuesday, 2nd November 2022. Around 30 participants attended from various institutions across the United Kingdom. The video of the first half of the workshop (the presentation portion) can be found here, with many thanks to the DPC for hosting: <https://www.youtube.com/watch?v=tHgg1KWzgL4>.

Work Package 7 Lead Gareth Cole began the workshop with a presentation, updating attendees on the activities of the [COPIM Project](#), including [Opening the Future \(Work Package 3\)](#), the [Open Book Collective \(Work Package 4\)](#), and the [Thoth metadata management system \(Work Package 5\)](#), [Experimental Publishing \(Work Package 6\)](#), and of course, [Archiving & Preservation \(Work Package 7\)](#).

Gareth explained the overall values and goals of the COPIM Project and introduced the core objectives and activities of each work package. This led into the important discussion of the proposed Thoth Archiving Network, a collaboration between Work Packages 5 and 7, to create a simple dissemination system for small publishers to archive their monographs in a network of participating institutional repositories. As examined in Chapter 4 and Chapter 5, [proof-of-concept has been developed and tested](#), and several universities have already agreed to take part.

For the second half of the workshop session, the attendees and COPIM colleagues were divided into three breakout rooms. The same two questions were posed for each group: ‘Would you be interested in joining the Thoth Archiving Network?’ and ‘What are the potential barriers for you joining the Thoth Archiving Network?’.

Earlier in the presentation, Gareth had posed a set of acknowledged challenges, and these were carried forward for discussion within the breakout groups, alongside the two main questions. These challenges were:

-  How many repositories have a preservation policy?
-  Is there metadata consistency across the repositories? If not, how to approach to assure fullest metadata record is preserved alongside the content?
-  Collection Policies
-  Different versions of software and how this might impact archiving
-  Proprietary systems

At the forefront of the discussions was the typical Content Policy at universities and institutions. For many, the policy is to only include content or research created by the institution's own academics and researchers. This potential barrier had already been identified by WP7, but discussion among the attendees provided a useful confirmation. While this type of content policy is indeed widespread, there is growing momentum within some institutions towards supporting open access infrastructure and initiatives as part of the university library's investment in research. As the role of institutions, their libraries, and research support bodies evolves within a changing open research landscape, there is little reason why the role of institutional repositories should not also evolve.

The practice of archiving with institutions is not unprecedented: CLOCKSS employs 12 mirror repositories within academic institutions across the world, which are called [“archive nodes.”](#) Though the structure and purpose are different, the premise is similar. Content policies at interested institutions will certainly be a primary barrier to the Thoth Archiving Network, but we do not see this as insurmountable. One significant shift emerging is the recognition within

institutions of the need for their committed support as the academic publishing model evolves. As mentioned within the workshop by one of our participants, when comparing the cost of BPCs (and APCs) against potential investment in evolving the open research infrastructure, the balance is largely tipped toward the future of truly open scholarship: a sustainable move in the right direction, rather than perpetuating a failing business model.

One interesting suggestion considered the packaging of publishers or collections of open access monographs on a particular theme or area of research. These could relate to research specialisms at institutions participating in the Thoth Archiving Network, which could provide an angle for institutional investment, as funding choices and external involvement require justification. An institution specialising in the Arts and Humanities, for instance, could support one or a number of small and scholar-led humanities publishers by creating a “special collection” of archived monographs in their repository as part of the Thoth Archiving Network. The use and promotion of materials archived within the repositories was also raised, highlighting the importance of knowledge contribution to university research culture.

There is also a precedent for special collections within university repositories established for the purpose of presented papers, presentation material, and conference proceedings from conferences held at the institution. The paper authors may be from many different institutions separate to the host institution, and though the institution will have hosted the conference, these external authors will still have their work made available through the host institutions repository.

Another point raised was the level of expected involvement between the institutional repositories and the publishers, for instance if there would be expected communication on a regular basis. Based on what is envisioned, this additional communication is unlikely to be required, and the main contact would be between Thoth/COPIIM and the repositories. As the Thoth Archiving Network is intended to be an easy, quick solution for under-resourced small and scholar-led publishers, the anticipated interaction would be either minimal or non-existent.

The aim would not be to add significantly to the workloads of repository managers, though there would need to be some initial onboarding support at the institution. The importance of engaging the correct decision-makers at institutions was also raised, as there could be a different combination of roles involved depending on the university. The key individuals could be the Head of the Library or Head of Research Office, or decisions may be made by representative groups, such as the Open Research Group or Research Committee. This ties into consideration of governance within the involved or potentially involved organisations, and how this will impact each step of the process as the network is implemented.

Further discussion around the capabilities of different repository software centred around the content of monographs, their formats, and potential additional requirements. For instance, what would happen if the monographs in question were complex monographs with additional content and audio/visual files, or experimental monographs in an unconventional format? Would certain repositories be unable to accommodate the archiving of this material? Our response to this insightful question was that the Thoth Archiving Network would have participating repositories of various types (EPrints, DSpace, Figshare, HAPLO, Samvera, etc.), and would have a push-button functionality allowing deposit in appropriately selected repositories. There would be user guidelines indicating the appropriate archiving location for a spectrum of open access monograph types, assuring the content would be deposited in a repository that could effectively contain the monograph content.

The landscape of open access monographs in this context is only beginning to emerge⁶³. Not all institutional repositories have a preservation layer, which is something that COPIM's WP7 colleagues recognise, and this was a point of conversation in most of the breakout rooms during the workshop.

⁶³ Laakso, M, Wise, A & Snijder, R 2022, Peering into the jungle: Challenges in determining preservation status of open access books. in *Proceedings iPres 2022 Glasgow 12—16 September 2022*. pp. 388-391, 18th International Conference on Digital Preservation, Glasgow, United Kingdom, 12.09.2022.
<<https://web.archive.org/web/20221104090209/https://ipres2022.scot/conference-proceedings/>>

The Thoth Archiving Network is therefore named because the initially envisioned solution here is that there is at least one if not several additional online locations where the open access monographs would continue to exist if the publisher ceased to operate and disappeared – they are archived. This does not always guarantee “preservation”, whether bit preservation or active preservation, as a fair few institutional repositories do not have a preservation layer. While many larger publishers, as well as some small-to-medium sized publishers, do pay for their publications to be preserved in a digital preservation archive or have them preserved via a third party (OAPEN, etc.), there are still many small and scholar-led publishers that do not have any relationship with a preservation service. We do hope to involve some repositories that do have preservation offered as part of their archiving; however, this is not envisioned as a prerequisite. While it would be ideal to have a “perfect” solution, such as a central, national repository for all open access monographs for publishers of any size⁶⁴, this is not yet possible. In the meantime, the importance of protecting the “long tail” of small publishers without sufficient resource to engage preservation players means that creative, community thinking is required.

University IT systems, and in particular, layers of network security, could pose a potential barrier to institutions participating in the Thoth Archiving Network. For most institutional systems, access is dependent on user identity, which could therefore complicate automated deposit via API depending on the credentials needed (or allowed). One workshop participant raised a particular query about Symplectic Elements and their authentication process.

The potential cost of hosting material on the institutional repository, particularly for those who have a preservation layer and hence additional operating costs, was raised as a key point. The small presses the Thoth Archiving Network would largely serve are likely not to be prolific publishers. Some publish as few as 5-10

⁶⁴ Adema, J & Stone, G 2017, *Changing Publishing Ecologies. A Landscape Study of New University Presses and Academic-led Publishing*. Joint Information Systems Committee, p. 72.
<<http://repository.jisc.ac.uk/6666/1/Changing-publishing-ecologies-report.pdf>>

monographs per year. Therefore, the storage burden would not likely be very sizeable if a repository only wished to support one or two publishers. However, this is an important consideration, one which is tied to the possible future business models for the Thoth Archiving Network, and which is already under consideration within the team.

Additional discussion and questions considered what might happen to existing deposited monograph records if the institution migrated to a new platform (and how this scenario might be handled by the Thoth Archiving Network); Library workflows and how content will be managed in the repositories; and questions around necessary rights and copyright that would impact the participating institution. Work Package 7 of COPIM subsequently held a workshop surrounding copyright, and a summary of this can be found [on COPIM's PubPub](#) open documentation site.

While those of us in Work Package 7 have been aware of most of the potential barriers to implementing the Thoth Archiving Network, and confirmation of these from the workshop participants was an important milestone, there were some useful and unexpected questions and challenges raised by the workshop participants that are deeply helpful. These will benefit the next steps in development now that we have nearly completed the proof-of-concept stages.

In the end, while there were certainly understandable reservations and questions about the Network, the workshop participants were generally quite supportive of the concept presented and some were certainly interested in knowing more about the Network and keen to be involved. We are hopeful that current development work will continue to progress the functionality and look forward to pilot testing with the Universities already beginning to participate.

Chapter 8:

Looking ahead to COPIM's Open Book Futures

A new grant to significantly expand and accelerate COPIM's open access infrastructures

A [version of this chapter](#) first appeared on COPIM's open documentation site, copim.pubpub.org.

Open Book Futures: announcing the new project

The [Community-led Open Publication Infrastructures for Monographs project \(COPIM\)](#) is delighted that [Arcadia](#) and the [Research England Development \(RED\) Fund](#) are supporting a new initiative that will build on the pioneering work of the COPIM project.

The Open Book Futures project (OBF), led by [Lancaster University](#), will significantly expand key infrastructures created by COPIM to achieve a step change in how community-owned Open Access (OA) book publishing is delivered.

Open Book Futures will follow the principles of [‘Scaling Small’](#) that guided the work of the COPIM project, further developing the infrastructures, business models, networks and resources that are needed to deliver a future for Open Access books led not by large commercial operations, but by communities of scholars, small-to-medium-sized publishers, not-for-profit infrastructure providers, and scholarly libraries.

Among its activities, OBF will deepen and accelerate the work of:

- the recently launched [Open Book Collective](#), which makes it easier for academic libraries to provide direct financial support to small- and medium-sized OA publishing initiatives;

- the [Thoth](#) metadata management and dissemination platform;
- the [Opening the Future](#) revenue model;
- the [Experimental Publishing Compendium](#);
- the forthcoming [Thoth Archiving Network](#).

Open Book Futures, which will run from 1 May 2023 to 30 April 2026, will increase COPIM's long-term impact and ensure that a wide range of voices have the opportunity to shape the future of open access book publishing. In order to amplify bibliodiverse and equitable community-led approaches to OA book publishing, OBF aims not just to strengthen existing networks in the UK and North America, but also to engage further with publishers, universities, and infrastructure providers in a diverse set of national and linguistic contexts, including Africa, Australasia, Continental Europe, and Latin America.

With that in mind, OBF will reunite many of the COPIM project partners, including [Birkbeck, University of London](#), [Coventry University](#), [Directory of Open Access Books \(DOAB\)](#), [Jisc](#), [Loughborough University](#), [Open Book Collective \(OBC\)](#), [Open Book Publishers \(OBP\)](#), [punctum books](#), [Thoth](#), and [Trinity College, Cambridge University](#), and they will also be joined by a wide range of new partners including [Continental Platform/University of Cape Town](#), the [Curtin Open Knowledge Initiative \(COKI\)](#), the [Digital Preservation Coalition](#), the [Educopia Institute](#), [Knowledge Futures](#), [Lyrasis](#), [OPERAS](#), [Public Knowledge Project \(PKP\)](#), [Research Libraries UK \(RLUK\)](#), [SciELO Books](#), [Scottish Universities Press/SCURL](#), and [SPARC Europe](#). The project is also supported by [Lancaster University Library](#).

[COPIM](#), a strategic international partnership led by Coventry University, was also jointly funded by Arcadia⁶⁵ and the RED Fund⁶⁶, and the COPIM project partners are delighted that OBF will carry the torch to significantly increase and improve the

⁶⁵ [Arcadia](#) is a charitable foundation that works to protect nature, preserve cultural heritage and promote open access to knowledge. Since 2002 Arcadia has awarded more than \$1 billion to organizations around the world.

⁶⁶ [RED](#) supports innovation in research and knowledge exchange in higher education that offers significant public benefits.

quantity, discoverability, preservation and accessibility of academic content freely and easily available to all.

COPIM co-Principal Investigator, Dr Janneke Adema of Coventry University, said:

“We are really thankful to the funders of the Open Book Futures project, the RED Fund and Arcadia, for their ongoing support of our work and proud to be able to continue what we started as part of COPIM, including the development of the [Open Book Collective](#), the [Thoth](#) metadata management platform, the [Opening the Future](#) revenue model and the Experimental Publishing Compendium. This grant will support the long-term sustainability of these community-led and community-owned book infrastructures, while building further international connections and networks with other partners and projects working towards an open knowledge commons for books.”

OBF Principal Investigator Dr Joe Deville, of Lancaster University, who is also a Co-Investigator on the COPIM project, has said:

“It is exciting to be able to contribute to a project that promises to profoundly reshape the very mechanisms through which academic knowledge circulates, in a context in which far too much high-quality book-length scholarship remains widely inaccessible.”

Archiving and Preservation within Open Book Futures

One of the main goals for Work Package 7 in COPIM’s new Open Book Futures (OBF) project is to expand not only to number of repositories on the Thoth Archiving Network, but also the geographical reach. A major part of OBF is to expand the global involvement and impact of our work, moving beyond the UK, USA, and Europe. These areas will of course continue to include key partners and collaborations, but a central ethos to OBF is supporting bibliodiversity, both in terms of the sizes and knowledge areas of presses, and the countries, nationalities, and languages involved.

Another key endeavour will be to establish an informal National Libraries Network, with the involvement of the British Library and other partners, whereby discussion might be undertaken about what requirements for the creation of an OA books archiving/preservation network at this level would involve. The outcome will be a report on the funding, technical and administrative needs for a National Libraries archiving/preservation network, should a network be created.

While COPIM's main focus has been on the published open access monograph, OPF will expand with a view to examine current and potential future practices for the archiving and preservation of PhD theses. Within the UK context, there has been a great deal of discussion amongst the university library and research communities regarding effective practices for digital PhD theses, which are increasingly becoming the norm. As there is a large amount of overlap between the monograph and the thesis, both in terms of general format, as well as requirements around access, preservation, and supplementary materials, one strand of this research will consider the digital thesis. A scoping report on current practices and further educational materials will be released.

We will also continue research and development towards further useful tools to benefit the community, including a toolkit with recommendations and software for small presses; Open Educational Resources (OER), including training materials for authors, publishers and libraries on archiving & preservation practices around PhD theses; and a report of recommendations for national and international policy makers.

We look forward to working closely with consortium project partners and network partners alike over the course of the next three years within Open Book Futures and, we hope, in the years beyond.



Appendix I: Tools and Resources

Open Access Books Toolkits & Guides

Toolkits

[Jisc New University Press Toolkit](#) – A toolkit created to “support and give guidance to new university presses and library-led publishing ventures as well as those with a hybrid model, who publish open access and non-open access material.”

[OAPEN Open Access Books Toolkit](#) – “aims to help book authors to better understand open access book publishing and to increase trust in open access books.”

Guides

[Open Book Publishers Authors Guide](#) – author-directed guide from OBP, including sections with guidance on licensing, audio and video material, and OA content resources.

OAPEN-recommended guides

(Source: [OAPEN Open Monograph Publisher Guides](#))

The OA effect: How does open access affect the usage of scholarly books? – Springer Nature Whitepaper (2017): “It is frequently claimed that open access (OA) has the potential to increase usage and citations. This report substantiates such claims for books in particular, through benchmarking the performance of Springer Nature books made OA through the immediate (gold) route against that of equivalent non-OA books. The report includes findings from both quantitative analysis of internal book data (chapter downloads, citations and online mentions) and external interviews conducted with authors and funders. This enables the comparison of actual performance with perceptions of performance for OA books.”

Download: [The OA effect: How does open access affect the usage of scholarly books? – Springer Nature Whitepaper \(2017\)](#)

Jisc and OAPEN: Publisher information on open access monographs (2016)

The guide ‘Publisher information on open access monographs’ presents recommendations for information that OA monograph publishers should make available on their websites to make their service clear to end users. The recommendations were created as part of the project [‘Investigating OA monograph services’](#), conducted by [Jisc](#) and OAPEN.

Download: [Jisc and OAPEN: Publisher information on open access monographs \(2016\)](#)

Jisc and OAPEN: Metadata for open access monographs (2016)

The guide ‘Metadata for open access monographs’ presents a metadata model for OA monographs. The model was created as part of the project ‘Investigating OA monograph services’, conducted by [Jisc](#) and OAPEN.

Download: [Jisc and OAPEN: Metadata for open access monographs \(2016\)](#)

OAPEN-UK: ‘Guide to open access monograph publishing for arts, humanities and social science researchers’ (2015): “The ‘Guide to open access monograph publishing for arts, humanities and social science researchers’ informs researchers about making their work available in open access. It provides a very useful overview of OA for books and is also relevant for other interested parties. The guide provides many helpful links to relevant projects and organisations. By providing an overview of possible business models, funders’ requirements, and a fair list of the benefits but also the many concerns involved, it helps the researcher to make a well-considered decision on publishing in open access. The guide is a result of Jisc Collections’ OAPEN-UK project.”

Download: [OAPEN-UK: 'Guide to open access monograph publishing for arts, humanities and social science researchers' \(2015\)](#)

Wellcome Trust: 'Open Access Monographs and Book Chapters: A practical guide for publishers' (2015): "'Open Access Monographs and Book Chapters: A practical guide for publishers', gives publishers information and recommendations on publishing open access books. The guide gives a clear answer on pressing questions that publishers might have before or while publishing in open access, such as which information should be available on their website, or how to make readers aware of the open access version of the book. The guide is developed by the Wellcome Trust and is of course indispensable for a publisher of open access monographs funded by WT, but also highly recommended for anyone interested in the area of OA book publishing."

Download: [Wellcome Trust: 'Open Access Monographs and Book Chapters: A practical guide for publishers' \(2015\)](#)

OAPEN-UK: 'Guide to Creative Commons for Humanities and Social Science Monograph Authors' (2013): "This guide explores concerns expressed in public evidence given by researchers, learned societies and publishers to inquiries in the UK House of Commons and the House of Lords, and also concerns expressed by researchers working with the OAPEN-UK project. The guide has been edited by active researchers, to make sure that it is relevant and useful to academics faced with making decisions about publishing. The 'Guide to Creative Commons for Humanities and Social Science Monograph Authors' is a result of Jisc Collections' OAPEN-UK project."

Download: [OAPEN-UK: 'Guide to Creative Commons for Humanities and Social Science Monograph Authors' \(2013\)](#)

Copyright, Reuse Licenses & Third-Party Content

Center for Media & Social Impact (CMSI) [Code of Best Practices in Fair Use for Scholarly Research in Communication](#)

College Art Association Code of Best Practices in Fair Use for the Visual Arts

- [Online Version](#)
- [Downloadable PDF](#)

Copyright Literacy (<https://copyrightliteracy.org/>)

- [Copyright Anxiety Scale](#)

[The Future of Copyright: Achieving Sustainable Universal Open Access through Copyright Reform \(A Debate\)](#) (Oxford University)

Publishing

Fulcrum – “Fulcrum is a community-based, open-source publishing platform that helps publishers present the full richness of their authors' research outputs in a durable, discoverable, accessible and flexible form.”

(<https://www.fulcrum.org/about/>)

Manifold – Manifold is a publishing software used to create different “projects”, which can include monographs, as well as textbooks, journal articles, among others. The software is flexible, allowing for inclusion of media and visualisations, along with format options for the source text. (<https://manifoldapp.org/>)

Open Monograph Press – OMP is an open-source solution for open access publishing, designed to be an end-to-end solution for publishing monographs.

(<https://pkp.sfu.ca/software/omp/>)

Digital Preservation

Digital Preservation Archives

CLOCKSS – “CLOCKSS, or Controlled [LOCKSS](https://clockss.org/) (Lots of Copies Keep Stuff Safe), is a shared dark archive that runs on LOCKSS technology. CLOCKSS’s content is hosted on 12 servers around the world, at leading academic libraries, with robust infrastructure and security.” (<https://clockss.org/>)

LOCKSS – “The LOCKSS project, under the auspices of Stanford University, is a peer-to-peer network that develops and supports an open source system allowing libraries to collect, preserve and provide their readers with access to material published on the Web. Its main goal is digital preservation.” (<https://www.lockss.org/>)

Global LOCKSS Network – “The GLN, the world's longest-serving LOCKSS network, ensures local custody, failover access, and post-cancellation access for subscription and open-access electronic journals and books at over 100 global research and academic libraries.” ([Global Lockss Network](https://www.lockss.org/))

Portico – “Portico is a community-supported preservation archive that safeguards access to e-journals, e-books, and digital collections. Our unique, trusted process ensures that the content we preserve will remain accessible and usable for researchers, scholars, and students in the future.” (<https://www.portico.org/>)

Internet Archive – “Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more.” (<https://archive.org/about/>)
Technically the Internet Archive is a web-archiving service rather than a digital preservation archive, but they are included here with this note, as they are a significant preserver of cultural and scholarly artifacts. The IA also appears to be in the process of creating a digital preservation service, more info here: <https://websiteservices.archive.org/pages/preservation>

***PKP Preservation Network** – The Public Knowledge Project (PKP) Preservation Network was created to digitally preserve OJS ([Open Journal Systems](#)) journals. The Network uses LOCKSS technology. The PKP PN ensures that journals that are not part of any other digital preservation service (such as CLOCKSS or Portico) can be preserved for long-term access. While presently the network only works to preserve journal articles, it is hoped that in the future there may be a similar network established for the OMP ([Open Monograph Press](#)) software. (<https://pkp.sfu.ca/pkp-pn/>)

Digital Preservation Software

[Archivematica](#) - A free and open-source digital preservation system that is designed to maintain long-term access to digital memory. “Archivematica provides an integrated suite of free and open-source tools that allows users to process digital objects from ingest to archival storage and access in compliance with the ISO-OAIS functional model and other digital preservation standards and best practices. Archivematica uses METS, PREMIS, Dublin Core, the Library of Congress BagIt specification and other recognized standards to generate trustworthy, authentic, reliable and system-independent Archival Information Packages (AIPs) for storage in your preferred repository.” Storage itself is not free and must be arranged elsewhere. (<https://www.archivematica.org/en/>)

[Archivum](#) – Commercial - An end-to-end digital preservation and archival safeguarding solution designed specifically for the Heritage, Libraries and Higher Education markets. The scholarly output strand of the service for Libraries and Higher Education markets ensures long-term accessibility of data while complying with Research Council regulations. (<https://arkivum.com/>)

[Preservica](#) – Commercial – “Preservica combines all the core functions for successful long-term active digital preservation and secure access into a single, intuitive and fully supported application aligned to the OAIS ISO 14721 standard.” (<https://preservica.com/>)

Digital Preservation Guides

[NDSA Levels of Digital Preservation](#) – First published in 2013 and updated in 2019 along with supporting documentation and additional resources, these guidelines are arranged in levels 1 to 5, in each of the following categories: storage, integrity, control, metadata, and content.

[Orbis Cascade Alliance Digital Preservation Step By Step Guide](#) – Follows the same five functional areas as the NDSA Levels of Digital Preservation, with guidance on how to improve preservation activities.

Digital Preservation Coalition (DPC) Resources

[Novice to Know-How: Online Digital Preservation Training](#)

[Digital Preservation Handbook](#)

[DPC Technology Watch Publications](#)

- **Reports:** in-depth reference guides around 40 pages; specific content or data types.
- **Guidance Notes:** brief overviews of 2 to 5 pages; address specific digital preservation challenges and solutions.
 - Wheatley, Paul. 'A Risk Driven Approach to Bitstream Preservation'. DPC, December 2022. <https://doi.org/10.7207/twgn22-02>.

[Digital Preservation Policy Toolkit](#)

[Case Studies](#)

File formats

[DPC Digital Preservation Handbook: File formats and standards](#)

[Library of Congress Recommended Formats Statement 2022-2023](#)

Tools

[Digital Preservation Tools by Function](#) (DigiPres.org)

[COPTR](#) (Community Owned Digital Preservation tool registry)

- [COPTR Tools Grid](#)

[DROID file format identification tool](#) (UK National Archives, Free)

[EPUBcheck](#) (validates EPUB files, World Wide Web consortium, w3.org, Free)

[JHOVE](#) (OPF, open source, extensible software framework for identification, validation, and 118haracterization)

[Metadata2Go](#) (Online tool that allows you to access the hidden exif & metadata of your files, Free)

[PRONOM file registry database](#) (UK National Archives, Free)

[veraPDF](#) (OPF, validates PDF/A files, Free)

Appendix II: Glossary

The definitions below are provided to give clear context to how these terms are used for the purposes of this guidebook. They are often drawn from existing sector definitions, which will be cited within the definitions where applicable, as well as cited at the end of this glossary.

Terms:

access – continued, ongoing usability of a digital resource, retaining all qualities of authenticity, accuracy and functionality deemed to be essential for the purposes the digital material was created and/or acquired for. (Source: [Digital Preservation Coalition](#))

AIP - Archival Information Package. An Information Package, consisting of the Content Information and the associated Preservation Description Information (**PDI**), which is preserved within an **OAIS** (OAIS term). (Source: [Digital Preservation Coalition](#))

API – An application programming interface (API) is a way for two or more computer programs to communicate with each other. It is a type of software interface, offering a service to other pieces of software. (Source: [Wikipedia](#))

With an API, “it is possible for users to send a list of instructions within certain parameters to a data store, usually a server and a database maintained by the content provider. This list of instructions is then processed, and data is returned to the user.”⁶⁷ (See also: [How-To Geek: What Is an API, and How Do Developers Use Them?](#))

archiving - the storage and preservation of records of enduring value.* (Source: [SAA Dictionary](#))
**Note: it is possible to digitally archive something in an archive location that is not supported by a digital preservation layer, and therefore archiving does not always assure digital preservation.*

bag - A package of content that conforms to the BagIt Specification (specification available at <http://www.digitalpreservation.gov/documents/bagitspec.pdf>). Under the specification, a bag consists of a base directory containing a small amount of machine-readable text to help automate the content's receipt, storage and retrieval and a subdirectory that holds the content files. See also "Bagit Specification" and "Bagger." (Source: [NDSA Glossary](#))

⁶⁷ ‘Application Programming Interface (API) - Digital Preservation Coalition’. Accessed 26 April 2023. <https://www.dpconline.org/digipres/implement-digipres/computational-access-guide/computational-access-guide-approaches/computational-access-guide-approaches-api>.

Bagger - A graphical software application tool to produce a package of data files that conforms to the BagIt Specification. See also "BagIt Specification" and "bag." (Source: [NDSA Glossary](#))

BagIt Specification - An Internet Engineering Task Force (IETF) Internet-Draft specification for a hierarchical file packaging format for the storage and transfer of arbitrary digital content. Specification available at <http://www.digitalpreservation.gov/documents/bagitspec.pdf>. See also "Bag" and "Bagger." (Source: [NDSA Glossary](#))

bit - A bit is the basic unit of information in computing. It can have only one of two values commonly represented as either a 0 or 1. The two values can be interpreted as any two-valued attribute (yes/no, on/off, etc). (Source: [Digital Preservation Coalition](#))

bit preservation - A term used to denote a very basic level of preservation of digital resource as it was submitted (literally preservation of the bits forming a digital resource). It may include maintaining onsite and offsite backup copies, virus checking, fixity-checking, and periodic refreshment to new storage media. Bit preservation is not [digital preservation](#) but it does provide one 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 Coalition](#))

bit rot/bit loss - The corruption of the lowest level of information digital data in transmission or during storage. (Source: [SAA Dictionary](#))

born digital - Digital materials which are not intended to have an analogue equivalent, either as the originating source or as a result of conversion to analogue form. This term has been used in the Handbook to differentiate them from 1) digital materials which have been created as a result of converting analogue originals; and 2) digital materials, which may have originated from a digital source but have been printed to paper, e.g. some electronic records. (Source: [Digital Preservation Coalition](#))

checksum - A unique numerical signature derived from a file, used to compare copies. (Source: [Digital Preservation Coalition](#))

dark archive - An archive that is inaccessible to the public. It is typically used for the preservation of content that is accessible elsewhere. (Source: [SAA Dictionary](#))

digital archiving - This term is used very differently within sectors. The library and archiving communities often use it interchangeably with digital preservation. Computing professionals

tend to use digital archiving to mean the process of backup and ongoing maintenance as opposed to strategies for long-term digital preservation. (This guide uses the latter definition.)

(Source: [Digital Preservation Coalition](#))

digital preservation - Digital preservation refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. Digital preservation is defined very broadly for the purposes of this study and refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological and organisational change. Those materials may be records created during the day-to-day business of an organisation; "born-digital" materials created for a specific purpose (e.g. teaching resources); or the products of digitisation projects. This Handbook specifically excludes the potential use of digital technology to preserve the original artefacts through digitisation. (Source: [Digital Preservation Coalition](#))

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.

medium-term preservation - Continued access to digital materials beyond changes in technology for a defined period of time but not indefinitely.

long-term preservation - Continued access to digital materials, or at least to the information contained in them, indefinitely.

DOI (Digital Object Identifier) - A technical and organisational infrastructure for the registration and use of persistent identifiers widely used in digital publications and for research data. The DOI system was created by the International DOI Foundation and was adopted as International Standard ISO 26324 in 2012. <http://www.doi.org> (Source: [Digital Preservation Coalition](#))

emulation – A means of overcoming technological obsolescence of hardware and software by developing techniques for imitating obsolete systems on future generations of computers. (Source: [Digital Preservation Handbook](#))

EPUB - The epub format is an open standard for e-books created by the International Digital Publishing Forum ([IDPF](#)). The EPUB family of standards 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. The container file is based on the [ZIP](#) format and defined in the Open Container Format (OCF). It is

referred to in this description as an EPUB Container, but the term "OCF ZIP Container" is also used in the EPUB specifications. (Sources: W3.org and Library of Congress)

fixity check - a method for ensuring the integrity of a file and verifying it has not been altered or corrupted. During transfer, an archive may run a fixity check to ensure a transmitted file has not been altered en route. Within the archive, fixity checking is used to ensure that digital files have not been altered or corrupted. It is most often accomplished by computing checksums such as MD5, SHA1 or SHA256 for a file and comparing them to a stored value. http://en.wikipedia.org/wiki/File_Fixity (Source: Digital Preservation Coalition)

FTP (File Transfer Protocol) - is a reliable method of transferring files electronically over the Internet, involving uploading files to and downloading files from websites and other computers connected to the Internet.

institutional repository - An institutional repository is an archive for collecting, preserving, and disseminating digital copies of the intellectual output of an institution, particularly a research institution.⁶⁸

memory institution - A memory institution is an organization maintaining a repository of public knowledge, a generic term used about institutions such as libraries, archives, heritage (monuments & sites) institutions.

metadata - Information which describes significant aspects of a resource. Most discussion to date has tended to emphasise metadata for the purposes of resource discovery. (Source: Digital Preservation Handbook)

administrative metadata - Data that is necessary to manage and use information resources and that is typically external to the informational content of resources. (Source: SAA Dictionary)

descriptive metadata - Information that refers to the intellectual content of the material and aids the discovery of such materials. (Source: SAA Dictionary)

preservation metadata - Information about an object used to protect the object from harm, injury, deterioration, or destruction. (Source: SAA Dictionary)

⁶⁸ 'Institutional Repository'. In *Wikipedia*, 21 February 2023. https://en.wikipedia.org/w/index.php?title=Institutional_repository&oldid=1140705921#cite_note-1.

structural metadata - Information about the relationship between the parts that make up a compound object. (Source: [SAA Dictionary](#))

migration - A means of overcoming technological obsolescence by transferring digital resources from one hardware/software generation to the next. 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. 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](#))

MOBI - MOBI files were the proprietary format for eBooks created by Amazon for the Kindle reader. Amazon has now retired the MOBI format, now recommending that for reflowable eBooks, use of EPUB, DOCX, or KPF file instead. (Source: [Kindle Direct Publishing](#))

monograph - a usually detailed, specialist written work regarding a single subject or an aspect of a subject, often by a single author, as a contribution to scholarly understanding.

normalisation - Converting files to a preservation format during the ingest process to aid long-term preservation. (Source: [Open Science Foundation](#))

obsolescence - A situation where digital content is no longer usable because the software or hardware it relies upon is unavailable or cannot be easily accessed using current technologies. (Source: [Open Science Foundation](#))

open access - free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. (Source: [Berlin Declaration](#))

PDF (Portable Document Format) - a set of formats and open standards maintained by the International Organization for Standardization for producing and sharing electronic documents originally developed by Adobe Systems. The original page description format has been elaborated over successive versions to enable the embedding of such complex objects as image, audio, and moving image files, hyperlinks, embedded XML metadata, and updatable forms. Specification for various versions and profiles of the format are now maintained by the International Standards Organization. (Source: [Digital Preservation Coalition](#))

<http://www.adobe.com/uk/products/acrobat/adobepdf.html>

PDF/A - Versions of the PDF standard intended for archival use. (Source: [Digital Preservation Coalition](#)) <http://www.aiim.org/Research-and-Publications/Standards/Committees/PDFA>

PID (Persistent Identifier) - A persistent identifier is a long-lasting reference to a digital resource. Examples are DOIs, handles, or ORCIDs. (Source: [ORCID](#))

render - To process a digital object (generally with a software application) in order to view, listen to, or interact with the content. This is usually done in a fashion consistent with the format encoding of the file. (Source: [Archives New Zealand Glossary](#))

scholarly infrastructure - A chain of interrelated actors, such as universities, academic publishers, data archives, and libraries, each of which serves a dedicated function.⁶⁹

trigger event – This terminology is used when specific conditions relating to an electronic publication and its continued delivery to users are met. If the publication is no longer available to users from the publisher or any other source for a variety of reasons then a trigger event is said to have occurred. They can set in motion access for users via an archive where the electronic publication may be digitally preserved. (Source: [Digital Preservation Coalition](#))

“Common trigger events can include the demise of the publisher, usually due to bankruptcy where there is no pick-up of their assets; discontinuation of a journal where a publisher removes all internet access; or, a disaster disrupting the publisher’s availability for an extended period of time.” (Kiefer, 2015)

validation - The process of making sure that data is correct and useful when checked against a set of data validation rules. These might include rules for package or file structure or specific file format profiles. (Source: [NDSA Glossary](#))

verification - The process of checking a copy of a data file to make sure that it is exactly equal to the original data file, or that a file remains unchanged over time. (Source: [NDSA Glossary](#))

web archiving - the process of collecting, preserving, and providing enduring access to web content. (Source: [SAA Dictionary](#))

⁶⁹ Plantin, Jean-Christophe, Carl Lagoze, and Paul N Edwards. ‘Re-Integrating Scholarly Infrastructure: The Ambiguous Role of Data Sharing Platforms’. *Big Data & Society* 5, no. 1 (1 January 2018): 2053951718756683. <https://doi.org/10.1177/2053951718756683>.

XML - Extensible Markup Language, a widely used standard (derived from SGML), for representing structured information, including documents, data, configuration, books, and transactions. It is maintained by the World Wide Web Consortium (W3C). <http://www.w3.org/XML/> (Source: [Digital Preservation Coalition](#))

Additional glossaries:

For more definitions of terms relating to digital preservation, archiving, and open access, see the following glossaries, all of which have provided a number of definitions within the above:

DPC Digital Preservation Handbook Glossary: <https://www.dpconline.org/handbook/glossary>

NDSA Glossary: <https://ndsa.org/glossary/>

Society of American Archivists Dictionary of Archives Terminology:
<https://www2.archivists.org/dictionary>

Working Definitions for the Levels of Digital Preservation (Version 2.0) – Open Science Foundation: <https://osf.io/rynmf>

References

Adema, Janneke, and Graham Stone. 'Changing Publishing Ecologies: A Landscape Study of New University Presses and Academic-Led Publishing'. Publication, 30 June 2017.

<https://repository.jisc.ac.uk/6666/>.

Adema, Janneke, Gary Hall, and Gabriela Méndez Cota. 'Combinatorial Books - Gathering Flowers - Part I'. *Community-Led Open Publication Infrastructures for Monographs (COPIM)*, 28 April 2021. <https://doi.org/10.21428/785a6451.d3ecc6cc>.

Adema, Janneke, Tobias Steiner, Simon Bowie, Marcell Mars, and Tobias Steiner. 'Part 2: A Typology of Experimental Books'. *Community-Led Open Publication Infrastructures for Monographs (COPIM)*, 29 January 2021. <https://doi.org/10.21428/785a6451.cd58a48e>.

'Application Programming Interface (API) - Digital Preservation Coalition'. Accessed 26 April 2023. <https://www.dpconline.org/digipres/implement-digipres/computational-access-guide/computational-access-guide-approaches/computational-access-guide-approaches-api>.

Arias, Javier, and Lucy Barnes. 'Thoth, Open Metadata and Building Structural Equity: An Interview for Open Access Week'. *Community-Led Open Publication Infrastructures for Monographs (COPIM)*, 27 October 2021. <https://doi.org/10.21428/785a6451.c7ddbe7d>.

Bolton, Kevin, Jan Whalen, and Rachel Bolton. 'Guidance for Digital Preservation Workflows'. 11 March 2022. <https://cdn.nationalarchives.gov.uk/documents/digital-preservation-workflow-guidance-web-server-copy.pdf>.

Bowie, Simon. 'A New Model for Computational Book Publishing'. *Community-Led Open Publication Infrastructures for Monographs (COPIM)*, 14 March 2023. <https://copim.pubpub.org/pub/computational-book-model/release/1>.

Brown, Adrian. 'Developing Practical Approaches to Active Preservation'. *International Journal of Digital Curation* 2, no. 1 (27 July 2007): 3–11. <https://doi.org/10.2218/ijdc.v2i1.10>.

'COalition S Statement on Open Access for Academic Books | Plan S'. Accessed 25 April 2023. <https://www.coalition-s.org/coalition-s-statement-on-open-access-for-academic-books/>.

Center for Media and Social Impact. 'Code of Best Practices in Fair Use for Scholarly Research in Communication'. Accessed 30 March 2023. <https://cmsimpact.org/code/code-best-practices-fair-use-scholarly-research-communication/>.

Chassanoff, Alexandra, and Colin Post. 'OSSArcFlow Guide to Documenting Born-Digital Archival Workflows | Educopia Institute'. Educopia Institute, 23 June 2020. https://educopia.org/wp-content/uploads/2020/06/OSSArcFlow_Guide_FINAL-1.pdf.

Cole, Gareth. 'Workshop on Copyright in the Context of Archiving and Preservation of Open Access Books'. *Community-Led Open Publication Infrastructures for Monographs (COPIM)*, 14 April 2023. <https://doi.org/10.21428/785a6451.bbefb58>.

'Creating Digital Materials - Digital Preservation Handbook'. Accessed 13 April 2023. <https://www.dpconline.org/handbook/organisational-activities/creating-digital-materials>.
Fanning, Betsy. 'Preservation with PDF/A (2nd Edition)'. Second. Digital Preservation Coalition, 31 July 2017. <https://doi.org/10.7207/twr17-01>.

'Extensible Markup Language (XML) and Its Role in Supporting the Global Justice XML Data Model', 2004. https://bj.a.ojp.gov/sites/g/files/xyckuh186/files/media/document/What_is_XML_article.pdf.

Ferwerda, Eelco, Frances Pinter, and Niels Stern. 'A Landscape Study on Open Access and Monographs: Policies, Funding and Publishing in Eight European Countries'. Zenodo, 1 August 2017. <https://doi.org/10.5281/zenodo.815932>.

Ferwerda, Eelco, Tom Mosterd, Ronald Snijder, and Pierre Mounier. 'UKRI Gap Analysis of Open Access Monographs Infrastructure'. Zenodo, 5 April 2021. <https://doi.org/10.5281/zenodo.5771945>.

Gatti, Rupert, Vincent W. J. van Gerven Oei, and Livy Onalee Snyder. 'Developing Thoth's "Software as a Service" Model'. *Community-Led Open Publication Infrastructures for Monographs (COPIM)*, 30 August 2022. <https://doi.org/10.21428/785a6451.8ffabe1b>.

'Getting Started in Digital Preservation: Taking Your First Steps', 10 June 2020. <https://blog-ica.org/2020/06/10/getting-started-in-digital-preservation-taking-your-first-steps/>.

Gonzalez-Fernandez, Pedro. 'More Open EBooks: Routinizing Open Access EBook Workflows | The Signal'. Webpage. The Library of Congress, 25 March 2020. [//blogs.loc.gov/thesignal/2020/03/more-open-ebooks-routinizing-open-access-ebook-workflows](https://blogs.loc.gov/thesignal/2020/03/more-open-ebooks-routinizing-open-access-ebook-workflows).

Gregg, Will, Christopher Erdmann, Laura Paglione, Juliane Schneider, and Clare Dean. 'A Literature Review of Scholarly Communications Metadata'. *Research Ideas and Outcomes* 5 (5 August 2019): e38698. <https://doi.org/10.3897/rio.5.e38698>.

Greenberg, Jonathan, Karen Hanson, and Deb Verhoff. 'Report on Enhancing Services to Preserve New Forms of Scholarship', December 2021. <https://doi.org/10.33682/0dvh-dvr2>.

Greenberg, Jonathan, Karen Hanson, and Deb Verhoff. 'Guidelines for Preserving New Forms of Scholarship', September 2021. <https://doi.org/10.33682/221c-b2xj>.

'Guidance on the Implementation of Plan S | Plan S'. Accessed 25 April 2023. <https://www.coalition-s.org/guidance-on-the-implementation-of-plan-s/>.

'Introducing the New NDSA Levels of Preservation - Digital Preservation Coalition'. Accessed 25 April 2023. <https://www.dpconline.org/blog/introducing-the-new-ndsa-levels-of-preservation>.

Johnson, Duff. 'Glossary of PDF Terms', 7 July 2021. <https://www.pdfa.org/glossary-of-pdf-terms/>.

Johnson, Duff. 'The Only Archival Digital Document Format - Digital Preservation Coalition'. Accessed 31 March 2023. <https://www.dpconline.org/blog/wdpc/the-only-archival-digital-document-format>.

Kiefer, Randy. 'Digital Preservation of Scholarly Content, Focusing on the Example of the CLOCKSS Archive'. *Insights the UKSG Journal* 28 (5 March 2015): 91–96. <https://doi.org/10.1629/uksg.215>.

Kirchhoff, Amy. 'Portico Content Type Action Plan: Supplied Files for E-Book Content, v. 1.0'. Portico, 22 March 2016. <https://www.portico.org/wp-content/uploads/2017/12/Portico-Content-Type-Action-Plan-Supplied-Files-for-E-Book-Content.pdf>.

Kirchhoff, Amy. 'Portico Content Type Action Plan: E-Book Content, v. 1.2'. Portico, 27 March 2016. <https://www.portico.org/wp-content/uploads/2017/12/Portico-Content-Type-Action-Plan-E-Book-Content.pdf>.

Kirchhoff, Amy, and Sheila Morrissey. 'Preserving EBooks'. DPC Technology Watch Reports, June 2014. <https://www.dpconline.org/docs/technology-watch-reports/1230-dpctw14-01/file>.

Laakso, Mikael, Alicia Wise, and Ronald Snijder. 'Peering Into the Jungle: Challenges in determining preservation status of open access books.' *Conference Proceedings - IPres 2022*, 1 December 2021. <https://ipres2022.scot/conference-proceedings/>. p. 388 – 391.

Laakso, Mikael, Lisa Matthias, and Najko Jahn. 'Open Is Not Forever: A Study of Vanished Open Access Journals'. *Journal of the Association for Information Science and Technology* 72, no. 9 (2021): 1099–1112. <https://doi.org/10.1002/asi.24460>.

Todd, Malcolm. 'File Formats for Preservation'. DPC Technology Watch Reports, 2009. <https://www.dpconline.org/docs/technology-watch-reports/375-file-formats-for-preservation/file>.

McGlone, Jonathan. 'Preserving and Publishing Digital Content Using XML Workflows'. IDS Project Press, 2013. <http://deepblue.lib.umich.edu/handle/2027.42/99563>.

'OA Books Toolkit'. Accessed 4 April 2023. <https://oabooks-toolkit.org/lifecycle/14788396-research-is-reused/article/4859829-metadata>.

'Open Access and Monographs: Report on Roundtables and Events'. Universities UK, March 2019. <https://www.universitiesuk.ac.uk/sites/default/files/uploads/Reports/open-access-and-monographs.pdf>

Rieger, Oya Y., Roger C. Schonfeld, and Liam Sweeney. 'The Effectiveness and Durability of Digital Preservation and Curation Systems'. Research Report. Ithaka S+R, 19 July 2022. <https://sr.ithaka.org/publications/the-effectiveness-and-durability-of-digital-preservation-and-curation-systems/>.

'Plan S Principles | Plan S'. Accessed 25 April 2023. https://www.coalition-s.org/plan_s_principles/.

'Punctum Books Helps Build Streamlined System for Archiving Open Access Monographs | Internet Archive Blogs', 22 February 2023. <https://blog.archive.org/2023/02/22/punctum-books-helps-build-streamlined-system-for-archiving-open-access-monographs/>.

'Right to Data Portability'. ICO, 17 October 2022. <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-data-portability/>.

Snijder, Ronald. 'Jisc and OAPEN: Metadata for Open Access Monographs (2016)'. Jisc and OAPEN, February 2016. <https://oapen.fra1.digitaloceanspaces.com/55111e5ee86a4653a0f9708034d5f9ae.pdf>.

Stone, Graham, John Rupert James Gatti, Vincent WJ van Gerven Oei, Javier Arias, Tobias Steiner, and Eelco Ferwerda. 'Building an Open Dissemination System'. Report, 27 July 2020. <https://www.repository.cam.ac.uk/handle/1810/310885>.

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>.

'Technical Guidance and Requirements | Plan S'. Accessed 25 April 2023. <https://www.coalition-s.org/technical-guidance-and-requirements/>.

Todd, Malcolm. 'File Formats for Preservation'. DPC Technology Watch Reports, 2009. <https://www.dpconline.org/docs/technology-watch-reports/375-file-formats-for-preservation/file>.

'UUK Open Access and Monographs Evidence Review'. Universities UK, October 2019. <https://www.universitiesuk.ac.uk/sites/default/files/uploads/Reports/UUK-Open-Access-Evidence-Review.pdf>.

van der Knijff, Johan. 'EPUB for Archival Preservation'. Open Preservation Foundation. KB/ National Library of the Netherlands, 20 July 2012. <https://openpreservation.org/system/files/epubForArchivalPreservation20072012ExternalDistribution.pdf>.

Velte, Ashlyn, and Olivia M. Wikle. 'Scalable Born Digital Ingest Workflows for Limited Resources: A Case Study for First Steps in Digital Preservation'. *Preservation, Digital Technology & Culture* 49, no. 1 (1 April 2020): 2–13. <https://doi.org/10.1515/pdtc-2020-0004>.

Verhoff, Deb, Karen Hanson, and Jonathan Greenberg. 'Preservation strategies for new forms of scholarship.' *Conference Proceedings - IPres 2022*, 1 December 2021. <https://ipres2022.scot/conference-proceedings/>. p. 81 – 88.

Wheatley, Paul. 'A Valediction for Validation? - Digital Preservation Coalition'. Accessed 28 April 2023. <https://www.dpconline.org/blog/a-valediction-for-validation>.

Wheatley, Paul. 'A Risk Driven Approach to Bitstream Preservation'. DPC, December 2022. <https://doi.org/10.7207/twgn22-02>.

Wyatt, Peter. 'ISO 19005 (PDF/A)', 25 August 2005. <https://www.pdfa.org/resource/iso-19005-pdf/>.

Wyatt, Peter. 'ISO 32000 (PDF)', 15 July 2021. <https://www.pdfa.org/resource/iso-32000-pdf/>.



Research
England



Published April 2023
The COPIM Project

Barnes, Cole, Fry, Gatti, and Higman.

All content released under CC-BY 4.0 License
unless otherwise specified.

<https://doi.org/10.5281/zenodo.7876048>