WP7 Scoping Report on Archiving and Preserving OA Monographs

Miranda Barnes, Emily Bell, Gareth Cole, Jenny Fry, Rupert Gatti & Graham Stone

June 2022

DOI: 10.5281/zenodo.6725309

copim

a project funded by

UKRI Research England

ARCADIA
A CHARITABLE FUND OF LISBET RAUSING & PETER BALDWIN

Table of Contents

1. Introduction

Technical methods for effectively archiving complex digital research publications and for creating an integrated collections of content in different formats have not yet been developed. As part of COPIM, an international partnership of researchers, universities, librarians, open access book publishers and infrastructure providers, WP7 (Work Package 7) have begun by compiling a digital preservation risk register. This report builds on that work in offering an overview of existing preservation solutions for Open Access (OA) research monographs. It brings together interviews conducted with representatives from several university presses and OA presses, and draws on the discussions that took place in a workshop held in September 2020 with a range of professionals in the archiving and preservation domain.

What the interviews and the workshop have indicated is the need for a consensus on file formats, further awareness and a culture shift to acknowledge and respond to the importance of digital preservation, further support and guidance for small and scholar-led publishers to assure equity in the publishing and preservation landscape, and a clear way forward regarding techniques to effectively preserve the components of complex digital monographs, including links and embedded content. A number of opportunities for future work have been highlighted, among them tools, guidance, developing new workflows, and nurturing a network of advocates in specific communities. These avenues for future work are further elaborated on at the end of this report.

1.1 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 publishers (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 OPERAS-P project and 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 of the COPIM Project will identify the key challenges associated with archiving research monographs in all their variation and complexity, and work towards developing new

solutions. The concept of a monograph as "just" text with the occasional image or table is increasingly outdated. "Books" now come in multiple digital formats (e.g. PDF, XML, EPUB) as well as hardcopy, and can also include embedded material such as videos and interactive 3D models. In some publications, users can interact directly with content hosted externally, such as databases and URLs. As individual objects, each of these formats—such as a PDF file or a video—appear in established guidance and standards for preservation and can be reliably archived with time, effort and resource. Yet how does one archive a "book" which consists of all of these?

1.2 Aims and Methodology

This report aims to provide a brief overview of existing technical methods for digital preservation of open access (OA) monographs, and offer some possible avenues for development. As the report is based on work package activities from 2020, further documentation of work package activities can be expected as these progress, including case studies and good practice guidance.

This report begins with a section detailing current practice in OA publishing, drawn from a series of semi-structured interviews of approximately 30 minutes conducted on Microsoft Teams in September and October 2020. The same interview questions were also supplied to several participants via email and via Online Surveys (formerly BOS). Though only a small set of responses were received, these have been incorporated into the findings. Opportunity exists for further discussions with publishers to enrich and advance the current findings. A wider analysis of the documentation pertaining to digital preservation at OA university presses and publishers was conducted during the same period, though documentation is limited.

Additionally, our first workshop took place on 16 September 2020, run jointly with the Digital Preservation Coalition. This workshop brought together COPIM teammates and experts in digital preservation, which included the following:

- Senior member, Archiving body (Participant A)
- Department head, National library (Participant B)
- Digital conservationist, National library (Participant C)
- Senior member, Data archive (Participant D)
- Senior member, Preservation archive (Participant E)
- Scientist, National laboratory (Participant F)
- Department head, National library (Participant G)
- Senior member, Community of practice organisation (Participant H)
- Senior technical officer, University library (Participant I)

These discussions feed into the points in section 3, as well as the concluding thoughts and summary of future work at the end of this report.

2. Current Practice in Preservation of OA Books and Journals

Based on the interviews with small publishers, university presses and preservation specialists, this section summarises some of the existing practice in archiving and preservation among university presses, scholar-led presses and libraries. These include the following key stakeholders:

- Senior member, University press A (Interviewee 1)
- Editorial board members, small OA press A (Interviewees 2 & 3)
- Senior member, OA journal publication platform (Interviewee 4)
- Senior member, preservation archive (Interviewees 5 & 6)
- Senior member, small OA press B (Interviewee 7)
- Senior member, mid-sized OA press (Interviewee 8)
- Senior member, University press B (Interviewee 9)
- Senior member, research academy (Interviewee 10)
- Senior member, small OA press C (Interviewee 11)

The sections below summarise responses to each question. Section 3 will draw out the themes arising from these interviews.

It should be noted that information about digital preservation at OA university presses and small publishers can often be difficult to find. The OPERAS survey was circulated to 56 partners in 17 countries and elicited only two responses. However, the Jisc landscape study (2017) included a question regarding current preservation solutions for NUPs in the UK, and received 43 responses overall, including 13 from HEIs with existing press initiatives. These responses are useful, but inconclusive, and indicate further clarity and investigation is needed.

2.1. In which format do you preserve publications and why?

Interviewees explained that the approach to preservation was mainly driven by format, with some emphasising the importance of content and version of record as the determining factor. The predominant format employed is PDF, but EPUB and XML are also significant formats, while XML and .zip package formats in particular appear to be growing in importance. However, the effort and resource required of many small OA publishers to produce XML files for their catalogue is more often than not prohibitive. (A note on XML: we appreciate that XML is not strictly a file format in the same sense as PDF or EPUB, but instead an encoding standard used as a basis for structuring information, which includes multiple file formats and metadata. However, for the purposes of simplicity and clarity in this report, it will be referred to as a file format used among publishers, alongside PDF and others.)

Answers to the question of preserved formats were broadly similar. Small OA press A indicated that they publish most titles as Print PDF, online PDF, EPUB, Kindle and HTML, and keep digital copies on personal devices and backup disks. Respondents from the mid-sized OA press responded that they similarly preserve all published formats (PDF, EPUB and MOBI), and will be adding XML to make transformation easier. Small OA press B preserves PDF, as does university press A and university press B (specifically their web-ready version). In terms of journals, the consulted OA journal publishing platform and journals published with university press B include JATS XML. Interviewee 4, a senior member of the referenced OA journal publishing platform, confirmed that the PDF was the highest priority for their platform as the version of record, but that the thinking behind JATS XML was to theoretically allow for forward migration into new formats. Interviewee 4 noted there had not yet been much uptake. The research academy respondent noted that they published PDFs, with indd/quark to be archived for further use and transformation into other formats. Formats such as .doc/.docx are also preserved just in case. Interviewee 11, a senior member of small OA press C, responded that they currently preserve PDF and HTML, but presently they do not have any direct

avenue for the preservation of XML, though they do produce these along with EPUB and MOBI. This is due to their present third-party relationship with a digital preservation body, by way of a central aggregator.

Two respondents from the digital preservation archive confirmed that the platform does not limit by format, instead focusing on content type; as with the publishers, the focus is on the version of record, i.e. what exactly was published. The respondents noted that, in general, content comes to the digital preservation archive as XML (with JATS and ONIX being common, for journals and books respectively; essentially, similar formats to those for indexing and dissemination) with PDFs and images. Supplementary files might be more varied, but the focus was on XML and PDFs with images. The organisation provides some guidance to depositors, primarily requesting consistency in deposits. The respondents noted that although the digital preservation archive can bit preserve all file types and trigger those for access (if required) there is no guarantee that those files will be usable (due to software obsolescence etc.).

2.2. What third-party material do you preserve (e.g. embedded material like videos and music or linked material like websites and URL references) and what steps do you currently take to preserve it?

The approaches to preserving third-party material described by the interviewees were mixed. From a lack of preservation of this material to relying on the somewhat black-box approach of CLOCKSS and LOCKSS, there was an evident absence of consistent practice in this area.

Some interviewees indicated that they did not currently take any steps to preserve embedded or linked material (small OA presses A and B), or that these efforts are very limited. The research academy archives reference works in progress on "rare occasions" using Zenodo. University press A stores additional material in their publishing repository, or requires data and audio-visual material to be stored on other institutional servers. University press B sends their web-ready PDFs with content embedded to a digital preservation archive, but also keep original manuscript and artwork files on the university server. The mid-sized OA press, which uses a different digital preservation archive, does not send images separately when there is a PDF, keeping to the embedded file, but they do preserve copies themselves. Supporting images and datasets are preserved in a repository. The publisher encourages creators to use persistent identifiers, and is moving to offering a video streaming platform through a new partnership as an alternative to YouTube. The OA journals platform does not currently preserve third-party references and links. The platform's senior member noted that it can be difficult to tell what digital preservation archive services do on this front, and that there is a delay between the time these services say they will ingest and when the content is processed. The copyright issues connected with this question were raised by several interviewees, particularly that if a publisher who has indemnified a service goes bankrupt, there is no recourse for the preservation service.

Small OA press C embeds images in all of their digital editions, with PDFs and HTML versions archived as previously mentioned, while audio and video materials are uploaded to the press's YouTube channel and the authors are also encouraged to archive the content in university or subject repositories. The press also uses either DOIs or handles to link to this content within the monographs. Links to external URLs used to archive websites referenced or linked to within the monographs are archived at the Internet Archive at point of publication.

The consulted digital preservation archive confirmed that they expect publishers will package up content and send it to them. The respondents noted that sometimes depositors might send something like a list of links, but that this was not preferred as it pushes the digital preservation archive into making editorial decisions. The respondents further highlighted that they rely on publishers telling them when things need to be preserved that are not expected; their processing will not automatically highlight these issues.

2.3. What is your current digital preservation process for open access monographs?

There were two clear trends in answers to this question. Most publishers indicated a specific preservation service with which they have minimal interaction (small OA press B, for example, uses an online OA deposit library and their processes) and an institutional repository for local back-ups.

University press B supplies PDFs to a digital preservation archive through a well-known digital library, keeping final files of all formats on the university server. The research academy indicated they use a local disk with a good back-up system, and stored published PDFs in the national library. At University press A, they follow the library's policy. Interviewee 4 of the OA journals platform mentioned a useful distinction between preservation and back-up: their platform uses a digital preservation archive, with an engineer having written a plug-in to automatically scan for new material in their publishing environment. Interviewee 4 suggests because this is automated, it is necessary to rely on the Keepers Registry to check on the safety status of content. The interviewee sees digital preservation archives as the "first line of defence" against total organisational failure. The second system is delegated third-party ingests such as the Internet Archive scholar service, and encouraging authors to submit to green repositories. These levels, while not seen as comprehensive, mitigate against different kinds of risk. A digital preservation archive is similarly the most crucial layer for the responding mid-sized OA press, who are also about to start a partnership with the Internet Archive to mirror content. Small OA press C preserves PDF versions in Portico via OAPEN, deposits digital editions in the British Library, and archives HTML versions with The Internet Archive.

In the interview with Interviewees 5 and 6 from the digital preservation archive, the question of when in the process the content gets preserved arose. The respondents noted that books are received after publication, but that it is not unusual to receive pre-print journal articles. They encourage submission as soon as possible in case of any problems such as publisher bankruptcy. Different publishers update content in quite different ways: some send content once a year, some once a month, some constantly, giving several versions with every small change. One of the respondents noted that with dynamic content that could be continually updated, preservation at the point of publication did not necessarily make sense; how do you know when something is 'completed' if it is open for comments and feedback, or is otherwise regularly updated?

When asked about the involvement of content creators in preservation, all interviewees indicated that this was very much only a conversation between the publishers and the preservers.

2.4. What are the benefits and limitations of your current approach?

Some interviewees felt they benefitted from being part of a memory institution which has a digital preservation infrastructure (e.g. personnel, technology and policy) in place. Others were aware that

their approach was far from systematic, which results in copyright issues and having to deal with obsolete formats.

Interviews indicated concern about wider issues, which were particularly problematic for small publishers that lack access to the type of infrastructure available to university presses. These included costs, environmental concerns associated with cloud-based services, copyright, and the long-term commitment. Sustainability was a key issue and the CLOCKSS service was seen as solution to this.

One challenge highlighted by the mid-sized OA press was the challenge of international publishing. Although the publishing process is now very smooth for standard books, the different places and repositories to register abroad, and different services that might need to be integrated, raise issues for the publications of monographs in a digital format.

University press A raised the important point that being connected to a library as a memory institution makes digital preservation easier for their press. At the research academy, in contrast, things were "not done systematically", and therefore there are obsolete formats and copyright issues. Small OA press B, meanwhile, raised issues of cost, the carbon footprint of cloud services, issues of copyright, and the question of how long we should be committing to preservation for. Small OA press C said that the difficulty with titles involving multi-media content or embedded material there is no way of systematically keeping all the content together to ensure it is accessible in its entirety in the future, but that presently the Internet Archive is the closest.

In terms of benefits, University press B noted Portico's sustainable community funding model and their confidence that the service would be provided long-term. The OA journals platform noted similarly that CLOCKSS was preferred due not only to its automation but due to its sustainable business model as a paid service. Interviewee 6 noted their digital preservation archive's approach works very well for content that looks similar to the print world. They are investigating web archives and more dynamic content, which is difficult to preserve at scale (such as GIS mapping and live visualisations).

2.5. Have any specific issues arisen in your digital preservation processes (e.g. around particular titles, or copyright issues, or access)

Few specific issues had arisen for interviewees. Some noted that the interaction with their preferred digital preservation bodies were relatively friction-less and that perhaps some items (e.g. those with unique characteristics) warranted further dialogue between the content provider and the service. Legal issues (e.g copyright, plagiarism) surfaced again as a fundamental concern.

The OA journal platform noted that interaction with their preferred digital preservation body is minimal, and individual articles are not debated. In agreement with the above comments by the responding digital preservation archive, Interviewee 4 noted that generic artefacts are easier to preserve. The mid-sized OA press noted that some services are not very reliable, redefining their processes and making it difficult to pin down what is a 'core' distribution system. Interviewee 8 (senior member of the mid-sized OA press) also noted the benefits of added value modules like DOAB, which validates peer review.

Interviewee 11, with small OA press C, again raised this issue that the existing process is not sufficient for any multimedia or embedded content. The press also relies on DOI and it is not clear who is responsible for maintaining DOI links once the publisher stops operations.

Interviewee 6 highlighted two issues with more traditional content: rights issues or plagiarism. Interviewee 6 also noted that, as a preservation agency, the digital preservation archive is not comfortable with deleting these files, but do ask for a new file without the offending text so that they can keep the original dark. In the interviewee's experience, publishers are open to this process though they start from a place of anxiety. Interviewee 5 noted that providing useable access to things outside of the book format, such as databases, is a big challenge. Interviewee 6 added that they could not afford to do the work multiple times for multiple things, and so needed supplementary material to be normalised so that it could be triggered in a similar way.

2.6. Other issues

Several other issues were raised across the interviews. Interviewee 4 noted that there needed to be more trigger events to show what processes would happen. Interviewee 4 also noted concerns about the redundancy procedures of preservers, not just in terms of hardware failure but social failure. The responding mid-sized OA press is considering establishing its own repository. The research academy noted that it is difficult to explain to users that digital preservation is important, while University press A noted that they were in a beneficial situation due to the institutional connection and that the long tail of small and independent publishers, or scholar-led platforms, may not have the same institutional back-up and therefore find it difficult to meet the fees of services.

Interviewee 6 similarly noted the pressure libraries could put on publishers to take digital preservation seriously, adding that some of their publishers had paid the digital preservation archive from marketing budgets because it kept their customers, the libraries, happy. Interviewee 6 said that publishers do understand and want the content to be available, but struggle with smaller income streams. Even the time it might take a small press run by a small team is prohibitive. Interviewee 6 concluded that communities needed to come together and work out how to fund preservation, how to get permission, etc. Interviewee 5 further noted that researchers as content creators could be better informed and advocate for digital preservation, and also noted the importance of standardisation in making the process of preservation cheaper.

3. Discussion Points

In this section, the threads of the summarised projects, points from the interviews, and points raised by the participants in our first WP7 workshop are synthesised.

3.1. Technical Considerations

Formats

It is clear from the projects and interviews that there are specific formats that technical solutions should focus on. These are the PDF (as the current version of record), XML (preferably a JATS equivalent, such as BITS), and WARC for web archives. There is a wider culture shift in the sector taking place towards XML, though the interviews showed there are still some difficulties prohibiting

uptake of XML for smaller OA publishers, particularly in terms of resource, and that for many of those publishers involved in the workshop and interviews PDFs will continue to be the primary focus. One solution suggested by workshop participants was packaging content together for digital preservation purposes, rather than expecting one format, such as EPUB or PDF, to retain everything, i.e. using a tool like Bagger or BagIt, which is already used in digital preservation to package digital content with metadata and documentation.

The role of repositories

There could be a role for institutional repositories in the archiving of multiple copies of OA monographs produced by OA presses, in particular for small scholar-led presses that may be run by a single individual or a very small team, which do not benefit from a memory institution. The resource limitations of these presses mean that while some use paid-for preservation services, many cannot afford these services or simply do not have a digital preservation policy in place due to the aforementioned resource limitations, which include lack of technical knowledge. Investigations on this front would be worthwhile pursuing, including the possibility of automation for the presses. However, it is important to note that unless the institutional repositories approached have their own digital preservation system in place, this is an exercise in archiving rather than preservation, and would only represent a beginning step in the process.

There also appears to be a pattern of preservation services not directly taking responsibility for any linked content or embedded content, as there is an expectation that publishers only require the deposited content to be preserved. Some publishers may deposit associated datasets in institutional repositories, but it is not a common practice. This will inevitably cause some issues in terms of maintaining the integrity of the digital book as an object, and even when there is a single, traditional PDF book being preserved, interviewees explained that it was difficult to keep track.

Automated services are appreciated but do not always provide reassurance that the content is ingested. One of the workshop participants suggested that a Keepers Registry, not only for publications with an ISSN, but for links and datasets, would be a step in the right direction. There is also a range of open source, free software such as LOCKSS, which could allow for the establishment of local repositories. However, the Open Preservation Foundation community survey has shown that only half of its respondents (mostly from academic research libraries, national libraries, archives and museums worldwide) have a digital preservation policy, and only 22% have a policy that is openly published. While a third were developing one, nearly 20% were not (Open Preservation Foundation, p. 10). This would mean the corresponding repositories may not have active digital preservation incorporated as part of their operations. Lack of time and resources were identified as key problems (p. 12), but what is highlighted is the need for clear and openly available documentation from both publishers and archives.

Emulation: a possible preservation tool?

There is a wide range of tools for web archiving, including solutions that capture more dynamic website content. Emulation was rarely mentioned, though it was raised as possibly the most faithful rendition of the content. Emulation, as a process, aims to reproduce or 'emulate' the original look and feel of the content as it originally appeared, by way of recreating the file within the environment of its original software. This is done by emulating applications, operating systems, or hardware platforms in

order to prevent the loss of original functionality by delivering the same user experience as the original platform.

The key benefit of emulation is to ameliorate the risk of losing the original functionality or the danger of the content becoming obsolete to a current platform. Emulation raises potential issues of proprietary software and appropriate licenses for use; otherwise, emulators need to be able to support these formats without relying on access to the original software by creating some other type of virtual environment. This content is also difficult to export, as noted in the workshop, raising the question of how this content is to be used and who will be using it. This technique may be worth further investigation, particularly where the functionality or presentation of preserved content are essential to the content, such as with experimental and dynamic content. However, for most applications, the necessary considerations appear to lie more within file formats and preservation workflows.

Preserving links

As "link rot" is a key factor in the disappearance of linked content from digitally preserved monographs, as well as journals and other materials, a focus in future investigations should involve clear methods in effective preservation of externally linked content. Digital monographs continue to evolve and become more complex, and in many cases depend on externally linked content particularly for hosting supplementary materials such as video and audio files, therefore broken or outdated links will likely cause a growing problem. Not only this, but referenced websites essential to scholarly material may cease to exist between the point of research or publication and that of access. Concerns along these lines highlight the need for active preservation practice in this area.

The Wayback Machine and the Internet Archive dominate the link archiving space, though there are other options which promise better rendering of live website elements. Combining this approach to links with a tool like Robust Links adds another level to this preservation of sources. Opportunities may exist for further incorporation of archived links into various monograph formats.

A common theme in our discussions was the social aspects of archiving and preservation: how do you convince content creators that long-term preservation is important when it comes to embedded and linked material? Linked material is a particular challenge here: a web page, on average, only lasts 90-100 days before changing, moving or disappearing completely. How far should the digital preservation arm reach? Further discussions with content creators (authors), publishers, and archiving bodies could prove useful in coming to a consensus.

Workflows

For more dynamic books, the very serious issues of cost and time associated with the work of preservation are even more vital to resolve. In depositing multiple copies of all associated content in multiple locations, the onus will often fall to the publisher, and smaller publishers are already limited in their time and staffing resource. A system of sustainable workflows would benefit small and scholar-led OA publishers in particular, but could also assist university presses and platforms with larger content profiles. There are possibilities for incorporating preservation into existing workflows, or in automating the process, that may provide benefits in the future.

A key factor in this will be formulating a practice to the archiving and preservation of all the components of a complex or dynamic monograph. Although the interviews demonstrated that content creators/authors are rarely involved in the digital preservation process, for increasingly multimedia monographs it will be necessary to consider these questions from the beginning. Preservation services do not want to make editorial decisions, and so these decisions need to involve authors much earlier. It is worth noting that the Embedding Preservability project at NYU is working to place digital preservation experts at the heart of the publishing workflow, where preservability as a concern may be introduced at the beginning of this process. However, the involvement, or at least the education of authors in digital preservation may be a key opportunity in the future, as decisions made at the content creation stage can have a significant impact on the ultimate preservability of the published content.

3.2. Concluding thoughts & opportunities for future work

There have been many opportunities for further work and investigation highlighted in this report, not least in terms of determining good, if not best, practice in archiving and preserving complex digital monographs. Certainly, a consensus on file formats would be useful, and what that means for levels, or tiers, of preservation that could be possible for different publishers and their needs. Additionally, there is work to be done in determining what workflows are possible to incorporate into existing publishing systems, and what new workflows could be developed and automated to support the limitations of smaller publishers.

PDF vs. XML vs. Other formats

More information is necessary to understand how and what content is preserved and what techniques will allow for the most reliable and effective preservation of digital monographs, particularly as their content and supporting materials grow more complex. While PDFs are the most adopted format for the publication of digital monographs at present, further explorations are needed to determine how well complex versions of monographs will fare in terms of archiving. We recognise that there is a necessity for work within the existing landscape to assure that, where possible, there is robust support provided for PDFs.

Additionally, while presently there is less uptake of XML among OA monograph publishers, there appears room for encouraging a broader uptake of this format, given its potential benefits in terms of preservation. However, this would require clear explanation, use cases, and documentation. There still exists the potential barrier of technical and staff resource for small publishers. Following on, scope exists for determining potential ways to package a monograph's content using looser linking rather than relying on the PDF as the version of record – moving past the current focus on digital preservation as revolving around a 'finalised' (i.e. printed) product.

Third-party materials

An additional aspect for further research is the challenge of archiving third-party material. Participants raised the problem of how archivists and preservation systems know what is there, i.e. the boundaries of the book as a digital object, and what issues there might be with third-party content, such as videos, links, digital appendices or scannable codes. Although preservation services currently focus on the book as close to the point of publication as possible, there need to be earlier conversations between

archivists and publishers, and publishers and authors. (This is currently being addressed by the Embedding Preservability project, which is a welcome step in a useful direction.)

Culture shift, good practice & communication

A substantial need for further awareness raising and communication exists within the academic community as well as the community of OA monograph publishers, in terms of enabling a culture shift around the importance of digital preservation and archiving. There are evident possibilities for facilitating training and tools for authors, funders, and publishers alike, and this could be expanded over time to involve librarians, universities, and academic-training organisations. Potential tools and guidance could be produced, for instance to encourage links to be 'robustified' by the author to reduce the pressure on the publisher.

Further efforts could be made surrounding the nurturing of a network of advocates within specific communities (for example, OABN) including partnering with funders and organisations to implement training schemes. On a practical level, the production of simple guidelines and templates that can be adapted by small publishers to give to authors and to begin conversations with repositories would extend the values of scaling small within the digital preservation landscape.

While there is not yet any firm 'best' practice, particularly in such a varied body of publisher sizes and resource levels, certainly encouraging the base principle of ensuring at least four copies are preserved in different locations in different forms of storage is a first step in establishing guidelines. "Deposit in at least one trusted preservation repository with full metadata" (Waters, 2016). From here, there is an opening to develop workflows for depositing different formats in different places, focusing on national libraries, domain-specific repositories, institutional repositories for datasets, and web archives for links. An openness in archiving and preservation methodology amongst publishers would encourage and expand good practice for the sector, so a solid objective to persuade publishers in making documentation available that outlines their preservation process, preferred formats, and any normalisation needed for supplementary material would be a positive way forward.

Next steps on a much longer path

A key takeaway from the workshop and interviews conducted is that there is no 'silver bullet' or single solution, and that digital preservation of OA monographs needs to take a more scattergun approach. Participants discussed this in terms of breadcrumbs that might be left, using platforms like the Wayback Machine alongside institutional repositories. Knowing that websites may disappear, certain files may become corrupted, and formats may become obsolete and therefore difficult to work with in future, it's clear that multiple technical solutions will be needed.

Possible next steps could include further workshops or smaller focus group-type sessions with stakeholders, workflow experimentation, consultations with related projects, and consultations with archiving and digital preservations specialists. As requirements for OA scholarship increase and the digital monograph itself grows more complex, further understanding of what will be needed for effective digital preservation will need to be uncovered via multiple processes. Certainly, there will be a requisite of flexibility in the decisions made now, bearing in mind what might very well change in the future. What appears evident now is that these concerns will continue to grow in their impact, and collaboration is a useful and efficient route towards finding new and effective solutions.

Appendix 1: Existing Preservation Solutions and Resources

1.1 Preservation Solutions

This section provides an alphabetical overview of major preservation solutions as identified by workshop participants, interviewees and project members, as well as a literature search. It is not exhaustive but offers broadly comparable information about a selection of solutions including formats supported, third-party content support, features, and costs, where the information is available. This information is drawn from the websites of these services, and in some cases supplemented via direct contact. This section also includes web archiving and emulation services. Important to acknowledge is the Western-centric and dominantly global North aspect of these examples, as the entities identified throughout this section, as well as the overall report, are primarily based in the United Kingdom, United States, and the European Union. There are, of course, preservation solutions based in other global regions not included here.

Name: Archive-It
Link: archive-it.org
Summary: Archive-It is a subscription service that allows institutions to build and preserve collections of born digital content. Archive-It partners can harvest, catalogue, manage, and browse their archived collections. Collections are hosted at the Internet Archive datacentre and are accessible to the public with full-text search. See also **Internet Archive**.
Format types: WARC
Third-party content support: Unspecified.
Features: The Archive-It web application allows users to add, import, and export descriptive metadata, and allows for public browsing and full-text search via archive-it.org. Archive-It also provides APIs and other tools to facilitate external integrations with local websites and repositories or third-party discovery or preservation storage services. The Archive-It service maintains a minimum of two copies of each collection online, a primary and a back-up copy.
Costs: Subscription starts from around $2,500.

Name: Archivematica
Link: www.archivematica.org
Summary: Archivematica is an integrated suite of open-source software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model. Users monitor and control ingest and preservation micro-services via a web-based dashboard.
Format types: Maintains the original format of all ingested files to support migration and emulation strategies. Normalises files to preservation and access formats upon ingest. Groups file formats into format policies (e.g. text, audio, video, raster image, vector image, etc.). Archivematica's preservation formats must all be open standards.
Third-party content support: Not specified.
Features: Archivematica uses METS, PREMIS, Dublin Core, the Library of Congress BagIt specification and other recognised standards to generate Archival Information Packages (AIPs) for storage in a preferred repository. Good format compatibility and integration with third party systems to allow for easy moving between systems as needed. Good documentation.
Costs: Open source with additional costs for support and hosting.

Name: Arkivum Perpetua
Link: arkivum.com/heritage-higher-education-and-corporate-archives
Summary: Digital preservation solution for the heritage, libraries and higher education that combines fully managed service with open-source software.
Format types: Not specified.
Third-party content support: Not specified.
Features: Local control, cloud-based management, format upconverting management, centralised access point for all formats.
Costs: Varies, starting at 1TB. Can use on-site storage or existing storage contracts.

Name: APTrust
Link: aptrust.org
Summary: The Academic Preservation Trust (APTrust) is a consortium of higher education institutions committed to providing both a preservation repository for digital content and collaboratively developed services related to that content. The APTrust repository accepts digital materials in all formats from member institutions, and provides redundant storage in the cloud. It is managed and operated by the University of Virginia.
Format types: All types of digital content from its member institutions, including but not limited to print, audio, video, and encrypted files
Third-party content support:
Features: The Amazon Web Services (AWS) platform provides APTrust with unlimited capacity that usually generates costs only as it is needed or used. A portion of the APTrust annual membership dues covers preservation of the first 10 terabytes of content per member, and members may purchase additional capacity in increments of 1 terabyte at a current cost of $420 per year. Those core service costs provide for three preservation copies of content in separate S3 availability zones in the AWS datacentre in Virginia and three preservation copies of content in separate Glacier availability zones in AWS's Oregon datacentre. APTrust conducts fixity checks on deposited content every three months. Long history of success and a very active community; good documentation and best practice.
Costs: Membership dues are $20,000 per year.

Name: CLOCKSS
Link: clockss.org
Summary: CLOCKSS provides a sustainable dark archive to ensure the long-term survival of web-based scholarly content. CLOCKSS (Controlled LOCKSS) employs a unique approach to archiving (Lots of Copies Keep Stuff Safe) that was initiated by Stanford University librarians in 1999. Digital content is stored in the CLOCKSS archive with no user access unless a "trigger" event occurs. The LOCKSS technology regularly checks the validity of the stored data and preserves it for the long term. CLOCKSS operates 12 archive nodes at institutions worldwide, preserving 200,000 book titles and a growing collection of supplementary materials and metadata information. As of March 2020, 64 titles have been triggered and made available open access. CLOCKSS participants include 300 libraries and 286 publishers.
Format types: Deliveries should include: both content and all related metadata materials as discrete file data: in a single directory, or in a directory tree, or packaged in a non-proprietary package format

(e.g., tar, zip). consistent metadata to content relationships: one metadata file per content file (1:1) or one metadata file per multiple content files (1:M), with metadata file names that include a timestamp or other unique identifier. Final content, no pre-publication content, non-proprietary formats only (e.g., PDF, HTML), a maximum of one non-text format for each item (e.g., PDF, EPUB, MOBI). Metadata text formats (e.g., XML, RIS), standard metadata schemas preferred (e.g., JATS, ONIX, PubMed, Crossref).
Third-party content support: In addition to the file types above, CLOCKSS also offers web harvesting.
Features: To allow CLOCKSS access to the publisher's source files, the publisher needs to place them on a designated FTP site. CLOCKSS boxes located at Rice, Indiana, and Stanford Universities ingest the content the publisher made available. The content is preserved through a system of audit and repair. The CLOCKSS boxes continually communicate over the internet to audit the content they are preserving. If the content in one CLOCKSS box is damaged or incomplete, that CLOCKSS box will receive repairs of the content based on other CLOCKSS boxes' holdings and/or by referring to the publisher's original presentation files. This cooperation between the CLOCKSS boxes avoids the need to back them up individually. It also provides unambiguous reassurance that the system is performing its function and that the correct content is always available.
Costs: Supporting library fees start at $485 per year for a library with a materials budget under $1 million.

Name: Conifer
Link: conifer.rhizome.org
Summary: Conifer is a web archiving service that creates an interactive copy of any web page that you browse, including content revealed by your interactions such as playing video and audio, scrolling, clicking buttons, and so forth.
Format types: WARC, ARC, HAR
Third-party content support: Recreates more of the user experience than other web archives (e.g. navigation, embedding).
Features: Conifer is a user-driven platform. Users can create, curate, and share their own collections of web materials. This can even include items that would be only revealed after logging in or performing complicated actions on a web site. On the technical side, Conifer focuses on "high fidelity" web archiving. Items relying on complex scripting, such as embedded videos, fancy navigation, or 3D graphics have a much higher success rate for capture with Conifer than with traditional web archives.
Costs: Free tier with 5GB of storage space with some networking quota restrictions. Access to collections that users made public is always free of charge and unlimited. $20 a month for 40GB of storage ($5 add on for an extra 20GB, or $200 a year).

Name: Digital Bedrock
Link: www.digitalbedrock.com
Summary: Secure, managed digital preservation services in an off-cloud architecture.
Format types: Not specified
Third-party content support: Work specifically with creative digital works.
Features: Store three copies, geographically dispersed. Combines object storage concepts to access unstructured data with off-cloud green storage. Migrates data from obsolete legacy LTO media to the cloud, current LTO media, or hard drives.

Costs: Starts at $360 per year and a one-off processing fee.

Name: Emulation as a Service Infrastructure (EaaSI)
Link: www.softwarepreservationnetwork.org/eaasi-gitlab
Summary: The EaaSI project encompasses the design, development, and implementation of scalable infrastructure and services for software emulation, including distributed management, description, sharing, and access.
Format types: It is built on open source platforms which can run proprietary formats (OpenOffice for Microsoft Office), GIMP (vs. Photoshop), SciLab, FreeMat, GNU Octave (vs. MATLAB), Scribus (vs. InDesign), FreeCAD, QCAD (vs. AutoCAD). A list of supported environments can be found here: https://eaasi-sandbox.softwarepreservationnetwork.org/eaasi/#/portal/environments.
Third-party content support: See above.
Features: Emulated CD-ROM environment sharing service, virtual reading rooms service, scientific software portal, API to automatically render objects in original software via emulation.
Costs: Currently free to test with more grants planned in future. Information about becoming a 'node host' is available on inquiry (limited to the US only due to copyright). Membership of the Software Preservation Network is $5,000 per year.

Name: Fulcrum
Link: www.fulcrum.org
Summary: The University of Michigan Library's ebook hosting, preservation, and media integration platform, developed on top of the Samvera repository platform. Aimed at preserving dynamic publications including collections of film and video clips for comparison, and visualising excavation records through three-dimensional interactive models.
Format types: Focus on Bit preservation rather than specific formats.
Third-party content support: Specialise in supporting a range of dynamic content.
Features: Contracted with CLOCKSS and have begun a pilot with HathiTrust to preserve content in their repository networks. U-M Library is a member of APTrust and routinely deposits all Fulcrum content into their cloud-based, consortially-governed service, providing further organisational and geographic redundancy for everything published on Fulcrum.
Costs: Begins at $2,500 per title.

Name: GitHub Archive Programme
Link: archiveprogram.github.com
Summary: GitHub has partnered with the Long Now Foundation, the Internet Archive, the Software Heritage Foundation, Arctic World Archive, Microsoft Research, the Bodleian Library, and Stanford Libraries to store multiple copies of software using their platform, on an ongoing basis, across various data formats and locations, including an archive designed to last at least 1,000 years.
Format types: Various (unspecified)
Third-party content support: Unspecified
Features: On every push to GitHub, they replicate Git data to multiple datacentres around the world. Additionally, they store backups of Git data, Issues, Pull Requests on GitHub in multiple locations. All of this data is available live via the GitHub API. GHTorrent monitors the GitHub public event

timeline, archives those events, and makes them queryable using BigQuery. You can also download snapshots by hour, day, or month. GHArchive monitors the GitHub public event timeline, archives those events, and recursively crawl and archive their contents and dependencies. Those archives will then be made available for download on a daily or monthly basis. The Wayback Machine will crawl GitHub's public repositories—including new repositories, issues, pull requests, wikis, and more—and store copies on hard drives in San Francisco and other locations. These archives will be publicly available via git and https. The Software Heritage Foundation will crawl GitHub on a regular basis and add its public repos to their archive, to which they provide public API access. Oxford University's Bodleian Library will provide redundancy for the Arctic Code Vault by keeping GitHub's 10,000 most-starred and most-depended-upon repositories in their depository as duplicate Piql film reels. On February 2, 2020, GitHub captured a snapshot of every active public repository, to be preserved in the GitHub Arctic Code Vault. This data will be stored on 3,500-foot film reels, provided and encoded by Piql, a Norwegian company that specialises in very-long-term data storage. The film technology relies on silver halides on polyester. This medium has a lifespan of 500 years as measured by the ISO; simulated aging tests indicate Piql's film will last twice as long. The GitHub Archive Program is partnering with Microsoft's Project Silica to ultimately archive all active public repositories for over 10,000 years, by writing them into quartz glass platters using a femtosecond laser.
Costs: Free.

Name: HathiTrust
Link: www.hathitrust.org
Summary: HathiTrust Digital Library is a digital preservation repository and access platform. HathiTrust provides long-term preservation and access services for digitised content from a variety of sources, including Google, the Internet Archive, Microsoft, and in-house member institution initiatives. Items in the public domain are in full view for everyone and items held in copyright are searchable. The members ensure the reliability and efficiency of the digital library by relying on community standards and best practices, developing policies and procedures to manage content and services at scale, and maintaining a modular, open infrastructure.
Format types: TIFF, ITU, G4, JP2 (JPEG 2000 part 1), and Unicode OCR with and without coordinates. Do not support schemas that describe publication structures (e.g. DocBook, TEI, EPUB), or derivative image formative (JPEG or PNG).
Third-party content support: No specific processes specified.
Features: Two synchronised instances of storage with wide geographic separation (located in datacentres in Ann Arbor, MI and Indianapolis, IN), and an encrypted tape backup with 6 months of previous-version retention (located in a third datacentre several miles from the Ann Arbor storage instance). The need for continuous integrity checking is fundamental to HathiTrust's data management strategy and underlies the choice of online (spinning magnetic disk) media for primary storage. Internally, each storage instance uses N+3 Reed-Solomon parity redundancy, which is analogous to but more fault-tolerant than conventional RAID 5 storage due to the additional parity redundancy. The storage system internally performs in-flight data integrity checks as well as periodic integrity checks of all at-rest data, and makes use of parity redundancy to permanently repair any errors encountered. External to the storage system, HathiTrust also conducts periodic validation of data with stored checksums to ensure that data has been ingested correctly and remains intact. Storage equipment is typically refreshed every 4-5 years. The storage system is modular and virtualised, with files split into blocks that are distributed across nodes of a cluster and automatically redistributed as needed to balance storage utilisation equally. Storage nodes that have reached retirement age may be

removed from the cluster with an administrative command, and new nodes may be added, with all movement of data managed internally while employing the in-flight integrity checks described earlier.
Costs: Tier-based fee system calculated on a cost-per-volume basis and total library expenditure, beginning at $7,146 per year.

Name: Internet Archive
Link: archive.org
Summary: The Internet Archive, a 501(c)(3) non-profit, is a digital library of internet sites and other cultural artifacts in digital form that provides free access. Contains 475 billion web pages, 28 million books and texts, 14 million audio recordings (including 220,000 live concerts), 6 million videos (including 2 million Television News programs), 3.5 million images, 580,000 software programs. Has newly launched an Internet Archive Scholar search engine. For books, see **Archive-It**.
Format types: Users input URLs, transformed to WAC
Third-party content support: When a dynamic page contains forms, JavaScript, or other elements that require interaction with the originating host, the archive will not contain the original site's functionality.
Features: Crawls are contributed from various sources, some imported from third parties and others generated internally by the Archive. The frequency of snapshot captures varies per website. Datacentres in three Californian cities: San Francisco, Redwood City, and Richmond. To prevent losing the data in case of e.g. a natural disaster, the Archive attempts to create copies of (parts of) the collection at more distant locations, currently including the Bibliotheca Alexandrina in Egypt and a facility in Amsterdam.
Costs: Anyone with a free account can upload media to the Internet Archive. The WayBack Machine archives webpages via various crawls and the archived sites are free to link to and access.

Name: LIBSAFE
Link: www.libnova.com
Summary: LIBSAFE Go is a fully automated cloud-based digital preservation platform.
Format types: Not specified.
Third-party content support: Not specified.
Features: Replicates content into four self-healing, geo-dispersed copies. Active Emulation Viewer allows you to open up to 75 file formats.
Costs: Price on application.

Name: LOCKSS
Link: www.lockss.org
Summary: The Lots of Copies Keep Stuff Safe (LOCKSS) programme one of the longest-running digital preservation initiatives still operating today. Provides the foundation for robust digital preservation of all types of digital content for libraries, publishers, and other content providers and stewards. The Global LOCKSS Network provides post-cancellation and perpetual access for eligible web-harvestable open access and subscription publications to participating libraries.
Format types: Supports a range of formats. WARC as default for back-end storage abstraction.
Third-party content support: Uses plug-ins to automatically ingest content, so does not specifically pull out or identify third-party content.

Features: LOCKSS networks can use a variety of mechanisms to ingest content. Currently, these methods are part of workflows that make content available for retrieval over the web. The LOCKSS software currently expects to be able to retrieve content for ingest over HTTP, in keeping with the original use case of harvesting web-based scholarly publications. However, any digital content can just as well be ingested and preserved, provided it can be made accessible over HTTP. There is a growing library of plugins for parsing the content and metadata made available by a variety of source platforms, according to a variety of packaging standards. As LOCKSS systems are managed locally, they do not trigger global OA release.
Costs: The LOCKSS software and technical documentation are available at no cost and under open licenses. Participation fees for the LOCKSS Alliance sustain ongoing improvement of core LOCKSS technologies and provide for technical support by the LOCKSS Program. Costs start at $2,642 depending on Carnegie classification.

Name: MetaArchive
Link: metaarchive.org
Summary: The mission of the MetaArchive Cooperative is to foster better understanding of distributed digital preservation methods and to create enduring and stable, geographically dispersed "dark archives" of digital materials that can be drawn upon to restore collections at member organisations.
Format types: Not specified.
Third-party content support: Not specified.
Features: Member institutions prepare their content for preservation, producing packages of content according to their local needs and workflows using tools such as Islandora, Archivematica, BitCurator, Fedora or Hydra. MetaArchive has produced a set of tools and scripts to assist members in preparing content for ingest into the preservation network, including tools to help with BagIt based ingests, and validating the quality of files before ingesting. All tools are publicly available.
Costs: Membership starts at $2,500 with additional fees for cache hosting, technology and storage.

Name: Perma.cc
Link: perma.cc
Summary: When a user creates a Perma.cc link, Perma.cc archives the referenced content and generates a link to an archived record of the page. Regardless of what may happen to the original source, the archived record will always be available through the Perma.cc link. Maintained by the Harvard Law School Library in conjunction with university law libraries across the country.
Format types: Perma provides two formats for each preserved web page: (1) capture and (2) screenshot. The capture uses the WARC file format. The screenshot is a .png file.
Third-party content support: As above.
Features: Links will be preserved as a part of the permanent collection of participating libraries.
Costs: Tiered subscription rate.

Name: Portico
Link: portico.org
Summary: A community-supported preservation archive that safeguards access to e-journals, e-books, and digital collections.

Format types: Most often, books are sent as batches of XML with one or more PDFs, figure graphics, and supplementary files. Some publishers supply EPUB. No formal limit to supported filetypes. See Hanson, 2019 for an overview.
Third-party content support: Mostly expect that the publisher will package everything and send it to them. Occasionally have received other material such as lists of links and dealt with this on an individual basis. Risk-averse in terms of copyright.
Features: Can download from interface, validate and migrate file formats, perform fixity checks. Current approach works very well for content that looks similar to the print world; Portico are also exploring solutions for more dynamic content and reference databases where access is networked rather than hierarchical, with a particular focus on what can scale.
Costs: Tiered model based on revenue, starting at $1,000. Additional one-off set-up fee.

Name: Preservica
Link: preservica.com
Summary: Cloud-hosted and on-premise digital preservation software.
Format types: Migration pathways for over 1,600 file formats.
Third-party content support: Not specified.
Features: Preservica combines core functions into a single application aligned to the OAIS ISO 14721 standard. This includes upload, online preparation, active preservation, safe storage, flexible management and access and discovery. Designed as a scalable platform with in-product assistance and extensive APIs to enable connectivity with other systems.
Costs: Not specified.

Name: ReplayWeb.Page
Link: replayweb.page
Summary: ReplayWeb.page is an open-source browser-based system to replay archived web pages as accurately as possible, with interactive elements preserved. Archives can be loaded as static files from anywhere a browser can connect to, or from a user's local machine.
Format types: WARC (preferred), CDX (supported), WACZ (in progress), HAR (in progress), WBN (experimental)
Third-party content support: Recreates
Features: High-Fidelity replay of web archives in several formats and from various locations directly in the browser. Several ways to explore web archives: story view, page search and URL search. On-demand, incremental loading of large archives. Several options for fully functional offline usage. Available as a standalone desktop app with Flash support. Support for versioned embedding of web archives. Webrecorder Desktop can be downloaded as a desktop app to do the same thing.
Costs: Free to use but does not provide any storage of its own.

Name: RODA
Link: www.roda-community.org
Summary: A digital repository solution that delivers functionality for all the main units of the OAIS reference model. RODA is capable of ingesting, managing and providing access to the various types of digital objects produced by large corporations or public bodies. RODA is based on open-source technologies and is supported by existing standards such as the OAIS, METS, EAD and PREMIS.

Format types: Supports all but has a normalisation policy (see https://www.roda-community.org/#theme/Format_Normalization_Policy.md).
Third-party content support: Not specified.
Features: Digital objects storage and management, supports any XML-based format as descriptive metadata, off-the-shelf support for Dublin Core and Encoded Archival Description, configurable multi-step ingestion workflow, supports pluggable preservation actions, integrated risk management, integrated format registry. Good documentation.
Costs: Price on application.

Name: Samvera
Link: samvera.org
Summary: Open-source repository framework based on the premise that no single system can provide the full range of repository-based solutions for a given institution's needs and that no single institution can resource the development of a full range of solutions on its own.
Format types: Not specified.
Third-party content support: Not specified.
Features: Based on: the Fedora repository software providing a robust, durable repository layer for persisting and managing digital objects; Solr indexes, providing fast access to information about an institution's repository content; Blacklight, a Ruby on Rails plugin that sits above Solr and provides faceted searching, browsing and tailored views on objects; Samvera gems: Ruby on Rails components that integrate the building blocks to form a complete, flexible and extensible digital repository solution.
Costs: Free and open source.

Name: WebCite
Link: www.webcitation.org
Summary: An on-demand archiving system for web references (cited webpages and websites, or other kinds of Internet-accessible digital objects). No longer accepts new archiving requests but continues to serve existing archives as of July 2019.
Format types: All types of web content, including HTML web pages, PDF files, style sheets, JavaScript and digital images can be preserved.
Third-party content support: As above.
Features: It is not crawler-based; pages are only archived if the citing author or publisher requests it. No cached copy will appear in a WebCite search unless the author or another person has specifically cached it beforehand.
Costs: No charge to individual users or publishers to use the service, but publishers who want their publications analysed and cited references archived must pay a fee.

1.2. Resources and Communities

This section provides an alphabetical list of links to additional resources, particularly those connected to the development of publishing and preservation workflows, with summaries drawn from the resource website.

Name: Archives Unleashed
Link: archivesunleashed.org
Summary: The Archives Unleashed project aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past. Supported by The Andrew W. Mellon Foundation, they develop web archive search and data analysis tools to enable scholars, librarians and archivists to access, share, and investigate recent history since the early days of the web.

Name: BitCurator Consortium
Link: bitcuratorconsortium.org
Summary: A community of organisations that support practitioners responsible for the curation of born-digital materials, especially through the application of free and open-source tools. Extensive documentation is available on their website.

Name: Beyond the Repository Curatorial Toolkit
Link: doi.org/10.17605/OSF.IO/GEJQS
Summary: The goal of this toolkit is to assist cultural heritage organisations in choosing materials to send to distributed digital preservation systems, networks in geographically dispersed locations designed to perform preservation actions.

Name: eArchiving Standards & Specifications
Link: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving+Standards
Summary: The eArchiving specifications are based on common, international standards for transmitting, describing and preserving digital data.

Name: DigiPres Commons
Link: www.digipres.org
Summary: Community-owned digital preservation resources.

Name: Digital Preservation Coalition
Link: www.dpconline.org/digipres
Summary: The Digital Preservation Coalition exists to secure our digital legacy through community engagement, advocacy, workforce development, capacity building, good practice and standards and management and governance. Their website contains a range of toolkits and guides.

Name: Digital Preservation Outreach & Education Network
Link: www.dpoe.network
Summary: DPOE was created by the Library of Congress in 2010 to provide digital preservation training across the US. They offer emergency hardware support for eligible institutions within the US.

Name: National Digital Stewardship Alliance
Link: ndsa.org
Summary: An international membership organisation that supplies advocacy, expertise, and support for the preservation of digital heritage. Documentation includes a very helpful guide to levels of digital preservation.

Name: Open Preservation Foundation
Link: openpreservation.org
Summary: A global not-for-profit membership organisation working to advance shared standards and solutions for the long-term preservation of digital content. The website offers extensive resources.

Name: OSSArcFlow
Link: educopia.org/ossarcflow
Summary: The Educopia Institute, in collaboration with the University of North Carolina at Chapel Hill School of Information and Library Science (UNC SILS), LYRASIS, and Artefactual, Inc., are investigating, synchronising, and modelling a range of workflows to increase the capacity of libraries and archives to curate born digital content. These archival workflows will incorporate three leading open source software (OSS) platforms—BitCurator, Archivematica, and ArchivesSpace—and the project will generate findings that can be generalisable to settings that are using other platforms and applications.

Name: ParCore
Link: parcore.org
Summary: Preservation Action Registries (PAR) has been set up to address a set of fundamental issues that affect the digital preservation community to facilitate the sharing of preservation actions between disparate systems, and provide the ability for well-formed preservation workflows, including sets of compounded preservation workflows to be imported and used within local environments from one or more external trusted sources.

Name: PubSweet
Link: pubsweet.coko.foundation
Summary: PubSweet is a free, open source framework for building state-of-the-art publishing platforms. PubSweet enables you to easily build a publishing platform tailored to your own needs. It is designed to be modular and flexible. PubSweet consists of a server and client that work together, and both can be modified and extended with components to add functionality to the system. There's also a command-line tool that helps manage PubSweet apps. Could be used to design publishing workflows including preservation and archiving steps.

Name: Robust Links
Link: robustlinks.mementoweb.org
Summary: Robust Links adds two elements to existing links using HTML5's attribute extensibility mechanism. A URI is added, the URI of a snapshot (e.g. in Wayback Machine) is added, and a date of linking is provided.

Name: Signposting the Scholarly Web
Link: signposting.org
Summary: Uses Typed Links as a means to clarify patterns that occur repeatedly in scholarly portals. For resources of any media type, these typed links are provided in HTTP Link headers. For HTML resources, they may additionally be provided in HTML link elements.

1.3. Relevant Projects

This section provides a non-exhaustive alphabetical list of projects relevant to the work of COPIM WP7 with summaries drawn from the project website.

Name: ENABLE! Community
Link: www.enable-oa.org
Summary: This project aims to jointly and in partnership develop an open access publication culture in the social sciences and humanities.
Preservation: Since 2006, they have been operating the disciplinary full-text server Social Science Open Access Repository (SSOAR).

Name: Embedding Preservability
Link: https://guides.nyu.edu/blog/The-Andrew-W-Mellon-Foundation-Awards-NYU-502400-For-Libraries-Project-to-Expand-Capabilities-F
Summary: The Mellon-funded follow-up project to Enhancing Services to Preserve New Forms of Scholarship, Embedding Preservability is collaborating with preservation service providers (Portico, LOCKSS), digital preservation specialists, and publishers to "embed" and operationalise preservation guidelines into the publishing process, from the first steps.
Preservation: "The goal is to support publishers in making design choices that result in publications, including very complex ones, that can be preserved at scale without sacrificing functionality."

Name: Enhancing Services to Preserve New Forms of Scholarship
Link: www.dpconline.org/blog/enhancing-services-to-preserve-new-forms-of-scholarship
Summary: Mellon-funded project, awarded to NYU Libraries in 2019 (in collaboration with CLOCKSS and Portico) to test the limits of current capabilities and create a clearly defined range of currently preservable technologies and guidelines and best practices.

Preservation: The team has identified several formats that are used for enhanced monographs among the participating presses: EPUB3, HTML5, and web publications. Not creating new technologies but seeing what is possible with existing ones. Working with web archiving technologies including Rhizome's Webrecorder, emulation in cases where a specific software environment is necessary for accurate playback, consulting with the Emulation as a Service Infrastructure (EaaSI) team at Yale University. The team has also published their set of guidelines, which are an outcome of the project: https://preservingnewforms.dlib.nyu.edu/guidelines

---

Name: Hiberlink
Link: hiberlink.org
Summary: The focus of the Hiberlink project is to assess the extent of so-called 'reference rot'. This two-year study investigates how web links in online scientific and other academic articles fail to lead to the resources that were originally referenced.
Preservation: "Link rot", or the phenomenon of hyperlinks breaking over time, so that they no longer lead to or reference the file, server, or webpage intended, is a key issue in digital preservation. When resources are relocated or cease to be available at the intended address, the content is lost. While Hiberlink offers no preservation itself, the work highlights a keen requisite for effective online, or linked, content preservation.

Name: Project JASPER
Link: https://doaj.org/preservation/
Summary: Project JASPER aims to close the gap in preservation coverage that currently exists among open access journals, in response to research that concludes online journals, both open and closed access, can disappear from the internet.
Preservation: Phase One of Project JASPER is a pilot project between CLOCKSS, DOAJ, Internet Archive, Keepers Registry and PKP. It is a scoping exercise aiming to find a solution that will reduce the number of unarchived open access journals. JASPER aims to use a process that establishes what archiving options are the best fit for individual publishers, the level of effort manageable by the publisher, and provide guidance to the journal's representatives to facilitate archiving based on these factors.

Name: Next Generation Library Publishing
Link: educopia.org/next-generation-library-publishing
Summary: Educopia, California Digital Library (CDL), and Strategies for Open Science (Stratos), in close partnership with LYRASIS, Confederation of Open Access Repositories (COAR), and Longleaf Services are working to advance and integrate open source publishing infrastructure to provide robust support for library publishing.
Preservation: The focus of the NGLP is on journals and on creating an open source, adaptable and interoperable library publishing infrastructure. There is not any overt focus at present on preservation. Currently the data storage component of the software is delivered via WDP API, providing "a file persistence layer, abstracted so that the platform is not tied to any particular storage provider. That layer is provided by MinIO15, which implements the widely-used Amazon S3 API, but independently of Amazon Web Services (AWS)." While additional physical storage can be added by the organisation

employing the software in a number of ways, there is not a clear preservation solution incorporated at this time.

Name: Research Collections and Preservation Consortium
Link: recap.princeton.edu
Summary: The ReCAP facility consists of a preservation repository and resource sharing services, and is located on Princeton University's Forrestal Campus. ReCAP is jointly owned and operated by Columbia University, Harvard University, The New York Public Library and Princeton University.
Preservation: The majority of the ReCAP collection is physical content held in a temperature and humidity-controlled archive, though item requests can be fulfilled digitally via their Digital Delivery option. There does not appear to be a specifically digital repository for archiving digital content.

Name: Sustainable History Monograph Pilot
Link: longleafservices.org/blog/the-sustainable-history-monograph-pilot
Summary: An Andrew W. Mellon Foundation–funded initiative to publish open digital editions of high-quality books from university presses in the field of history.
Preservation: Does not appear to be central to their mission, despite the name. The website mentions indexing books in OAPEN and placing things in the Internet Archive "as well as many other open platforms".

Name: Strategies for Sustaining Digital Scholarship at University of Maryland
Link: https://ischool.umd.edu/research/projects/community-centered-strategies-sustaining-digital-humanities-scholarship
Summary: A Mellon-funded study of how research communities interpret, impact, and implement the sustainability of their own digital scholarship. The goal of this project is to identify strategies that digital humanities research communities may use to contribute to and increase the sustainability of their own digital resources and collections. Running May 2020 to October 2021.
Preservation: Though the initial grant announcement did not specifically mention preservation, the PI (Katrina Fenlon) has published a conference preprint that highlights the challenges and opportunities around "Sustaining Digital Humanities Collections" (link), particularly in terms of preservation. Fenlon's paper details the findings of the study, which "surfaced four main challenges confronting the sustainability and preservation of digital humanities collections: (1) Discontinuity between the essential interactivity of digital collections and the paradigm of artifactual preservation; (2) The importance and vulnerability of "connective tissue" within and between collections; (3) Ambiguity of institutional contexts and roles; and (4) Lack of infrastructure for collaborative humanities workfows."

Name: TimeTravel
Link: timetravel.mementoweb.org
Summary: Time Travel helps you find and view versions of web pages that existed at some time in the past. These prior versions of web pages are named Mementos. Mementos can be found in web archives or in systems that support versioning such as wikis and revision control systems.

Preservation: Draws from web archives: archive.today, Archive-It, Arquivo.pt: the Portuguese Web Archive, Bibliotheca Alexandrina Web Archive, DBpedia archive, DBpedia Triple Pattern Fragments archive, Canadian Government Web Archive, Croatian Web Archive, Estonian Web Archive, Icelandic web archive, Internet Archive, Library of Congress Web Archive, NARA Web Archive, National Library of Ireland Web Archive, National Records of Scotland, perma.cc, PRONI Web Archive, Slovenian Web Archive, Stanford Web Archive, UK Government Web Archive, UK Parliament's Web Archive, UK Web Archive, Web Archive Singapore, WebCite, Bayerische Staatsbibliothek. Also transactional web archives that run the SiteStory software that self-archives web server content, systems that have a bespoke version API but for which a TimeGate server was implemented, for example, arXiv.org, GitHub, and MediaWikis that have installed one of the Memento extensions for MediaWiki, for example, https://www.w3.org/wiki/Main_Page.

Name: Towards an Open Monograph Ecosystem
Link: www.openmonographs.org
Summary: A five-year pilot project of the Association of American Universities (AAU), Association of Research Libraries (ARL), and Association of University Presses (AUPresses) to develop a sustainable open monograph ecosystem.
Preservation: Almost no mention of archiving or preservation on their website.