IJDC | *Brief Report*

Long-Term Preservation and Reusability of Open Access Scholar-Led Press Monographs

Dr Miranda Barnes
Loughborough University

Ross Higman
Open Book Publishers

Dr Gareth Cole
Loughborough University

Dr Rupert Gatti
Open Book Publishers

Dr Jenny Fry
Loughborough University

Abstract

This brief report outlines some initial findings and challenges identified by the Community-Led Open Publication Infrastructures for Monographs (COPIM) project when looking to archive and preserve open access books produced by small, scholar-led presses. This paper is based on the research conducted by Work Package 7 in COPIM, which has a focus on the preservation and archiving of open access monographs in all their complexity, along with any accompanying materials.

Submitted 13 June 2022 ~ *Accepted* 13 June 2022

Correspondence should be addressed to Miranda Barnes. Email: m.l.barnes@lboro.ac.uk

This report is based on a presentation at International Digital Curation Conference IDCC22, online, 13-16 June, 2022

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

COPIM[1] (Community-led Open Publication Infrastructures for Monographs) is an international partnership of researchers, universities (Coventry University[2]; Birkbeck, University of London[3]; Lancaster University[4]; and Trinity College, Cambridge[5]), established open access publishers (the ScholarLed consortium[6], which includes Mattering Press[7], meson press[8], Open Humanities Press[9], Open Book Publishers[10] and punctum books[11]), libraries (UC Santa Barbara Library [12]and Loughborough University Library[13]) and infrastructure providers (the Directory of Open Access Books[14] and Jisc[15]).

Work Package 7 (and COPIM as a whole) is focussed on archiving and preserving the outputs of small, scholar-led, open access presses. Examples include the ScholarLed group of presses, including Open Book Publishers (OBP).[16] The work package includes preservation experts, librarians, scholars and publishers from Loughborough University, Open Book Publishers, the Digital Preservation Coalition, the British Library, and Jisc. The inclusion of these different specialisms has been essential to our work because it has highlighted the various challenges faced by different stakeholders involved in the preservation of open access monographs. Without input from those who participate in this process, from many angles, the solutions we aim to develop or contribute to would be far less effective.

Preservation of these open access monographs is essential for their current use and future reuse. In addition, because monographs increasingly have elements associated (or directly included) which are not textual, the complexity of preserving born digital works is increasing. For effective use and future reuse, all essential components of open access monographs must be preserved, wherever possible.

In Work Package 7, we have been using Open Book Publishers as a test bed for potential workflows and processes. In particular, we have been using two publications which have embedded files that are essential to the understanding and re-use of the work. These two works were deliberately chosen due to the variety of content and file types:

- 'Rameau's Nephew' – 'Le Neveu de Rameau': A Multi-Media Bilingual Edition (Diderot, 2016)
- Image, Knife, and Gluepot: Early Assemblage in Manuscript and Print (Rudy, 2019)

Our work has focused on exploring different options to preserve these books in a way that will help to ensure their reusability over time. By reusability we don't just mean the textual

[1] https://www.copim.ac.uk/
[2] https://www.coventry.ac.uk/research/research-directories/current-projects/2020/copim/
[3] http://www.bbk.ac.uk/
[4] https://www.lancaster.ac.uk/
[5] https://www.trin.cam.ac.uk/
[6] https://scholarled.org/
[7] https://www.matteringpress.org/
[8] https://meson.press/
[9] https://openhumanitiespress.org/
[10] https://www.openbookpublishers.com/
[11] https://punctumbooks.com/
[12] https://www.library.ucsb.edu/
[13] https://www.lboro.ac.uk/library/
[14] https://www.doabooks.org/
[15] https://www.jisc.ac.uk/
[16] https://www.openbookpublishers.com/

element but the other associated and integrated content such as audio files and image files. We are also investigating how much of the preservation processes can be automated to help these smaller presses. Although technically similar to the challenges facing larger presses, the challenges small presses encounter also potentially include staffing, technological, and expertise deficits As part of our work, we have identified the main challenges facing these presses: .

1. Metadata loss when outputs are moved between different infrastructures (e.g., publisher to dissemination platform to preservation platform)
2. The formats created by smaller presses are less suited to preservation workflows than those created by larger presses and journal publishers (e.g., PDF vs XML)
3. Open access (e.g., should the archived version be "open" or "closed") and, linked to this, consideration of file formats in archive vs. "access" copies
4. The need to package their books in different ways for different organisations and platforms (e.g., metadata differences, format differences etc.)
5. How to preserve embedded or linked content which is essential to the understanding of the book
6. Limited staff resource may prevent the implementation of additional workarounds

As part of our research we have spoken, primarily through interviews, focus groups, and follow up meetings, to many scholar-led presses, preservation providers, and similar projects looking at preservation of publications (both journals and monographs).

By providing practical, and hopefully easily implemented, solutions we can make a big difference to the long-term reusability of the published books. The project will assist publishers to see preservation as "just another" dissemination route. COPIM has also created an open metadata management and dissemination system called Thoth[17]. Our aim is to embed an option for publishers to be able to send metadata (and associated files) to repository systems (e.g., DSpace, Eprints, Figshare etc.) in the same way publishers send metadata and files to aggregators, national libraries etc. By utilising the repository systems' APIs and enabling this quick "send to repository" option, it should help publishers to archive their books more efficiently. This will also enable the reuse of material by having multiple access and discoverability routes to the publications.

Main Challenges

The main challenges noted above can be expanded upon as follows:

Metadata Loss and File Formats

Our discussions with the smaller presses have highlighted the fact that many of them rely on third party aggregators and dissemination/discovery platforms to undertake the preservation of their material. This has obvious advantages for the press in that there is no additional staff time or expertise needed. However, our current research has shown that there is a loss of some metadata along the route from publisher to aggregator to preservation platform. In addition, not all versions (i.e., file types) are sent from the publisher to the aggregator. For instance, a publisher may have XML, EPUB and MOBI versions of the monograph text along with the PDF. In many cases only the PDF is sent to the aggregator which means that only the PDF version of the work is archived and preserved. The combined loss of both metadata and file types has implications for both discovery and reuse (e.g., in a number of cases we have identified where the open access licence has not been archived with the open access book).

[17] https://thoth.pub/

Preservation & Access

Whether or not a digital preservation archive holds archived open access content closed or open will depend on the archive. Certain archives keep all archived content closed and inaccessible to the public as a matter of practice (a "dark archive"). This is likely due to the fact that most content in these archives arriving from proprietary publishers, with the intent for preservation but not access, unless the publisher ceases to operate. However, others have a system in which they trigger all open access content open. This is often done during the onboarding process with a publisher, who confirms all open access material when they begin participating in the preservation process. Older content that may have been open access when published, but which does not have an easily identifiable reuse license, is more difficult to trigger open as a matter of course.

Several of the open access presses we have worked with (although not all) have highlighted that for native open access materials, the "archived" copy should also be publicly available. While this is policy for some large digital preservation archives, it is not necessarily standard practice for all. There is also the issue of discoverability: while open access content may be triggered open in some cases, it won't necessarily be easily found by those searching for it.

Automation in archiving & dissemination

Many aggregator and dissemination platforms use different standards of metadata. Understandably, the publishers we spoke to during our research have focused their efforts on dissemination. As such, their limited staffing numbers have not enabled preservation to be a priority and there has been a reliance on the third-party aggregators as outlined above. Having an easily implemented, automated, process to deposit straight from the publisher to an archive will hopefully mitigate some of these challenges. Using Thoth (see above), COPIM not only aims to ease the archiving of content but also the dissemination of this material to the aggregators by ensuring the various metadata standards can be created automatically from with the system.

Embedded and linked content

This is a challenge for all publishers (whether monograph or journal) but particularly for the smaller monograph presses we have been working with. The first challenge is that this content, particular linked content, may not be under the control of the press or the author and could be held anywhere on the web. For linked content, solutions do exist (e.g., creating permanent archive links through services such as the Internet Archive[18]) but their creation and inclusion relies, primarily, on authors rather than publishers. As part of COPIM we examined options to automate the creation of archive links and embedding these in the final manuscripts (alongside the live links as in some cases it is important for the current version of the page to be read).

However, fully automating this whole process (including updating links in the manuscript) has so far proved too complicated. Embedded content is equally challenging but there are more solutions available for smaller presses especially where the material is under the control of either the author or press. Using Thoth we hope to include this embedded/associated material in the dissemination to repositories. At its simplest, Thoth will have links to both the textual material and the associated material, and the system will use those links to transfer the available content to a partner repository. Our aim is to simplify this process to a single "Send to archive" button. We have conducted proof of concepts and are hopeful that this will be achievable by the end of the project in April 2023.

[18] https://archive.org/

Staff numbers & automation

As mentioned above, staffing levels and expertise are particular challenges for smaller presses. Currently, very few of the presses we have spoken to have established preservation procedures or processes for this reason. However, by automating the processes we hope to ensure that establishing a basic preservation option is possible, even for publishers with a single member of staff. The importance of automation for enabling archiving and preservation has been shown by our research. Unless the steps necessary are automated, simplified, and do not require additional resource, it is unlikely for many of these smaller presses there will be bandwidth for further archiving and preservation procedures to be introduced.

A COPIM staff member, with extensive knowledge of the relevant repository software (Figshare), took three days to manually transfer two books and all associated content and metadata from Open Book Publishers to Loughborough University's test repository system. Automating this process requires some initial set up resource (e.g., confirming the schema etc.) but after that, our research and proof of concept has shown that subsequent book transfers should be quick and painless for the publisher.

Automating processes as outlined above should aid both preservation and reuse. The publishers can be more confident that their material is archived and have more control over the content they archive (rather than relying on a third-party aggregator with the limitations noted above). In addition, using Thoth to send material to aggregators and other dissemination platforms will hopefully aid discoverability and reuse because the quality of the metadata will be higher than at present.

Concluding thoughts

Although the options we outline in this paper are not gold-plated, we do see them as a good starting point for smaller, scholar-led presses. With these foundations in place, we can then build on these solutions, including better advocacy and guidance materials for presses and authors, to ensure that current manuscripts are useable and reusable for scholars, policy makers, the general public, and other interested parties into the future.

References

Diderot, D. (2011, June) 'Rameau's Nephew' – 'Le Neveu de Rameau': A Multi-Media Bilingual Edition (Ed. M. Hobson. Translators: K.E. Tunstall and C. Warman. Music: P. Duc). Retrieved from https://doi.org/10.11647/OBP.0098.

Rudy, K. (2019, July). Image, Knife, and Gluepot: Early Assemblage in Manuscript and Print. Retrieved from https://doi.org/10.11647/OBP.0145.