COPIM

COPIM Archiving and Preservation Workshop, September 2020

Emily Bell¹

¹**COPIM**

Published on: Oct 27, 2020

DOI: 10.21428/785a6451.0e666456

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

As part of [COPIM Work Package 7](#) (Archiving and Digital Preservation), we are running a series of workshops in 2020 and 2021 aimed at gaining a better understanding of existing best practice in the preservation of OA monographs, and developing possible solutions for the technical issues of archiving books, including embedded content and links. Our [first workshop](#) took place on 16 September 2020, jointly run with the [Digital Preservation Coalition](#). This workshop brought together representatives from the Archaeology Data Service, Cambridge University Library, Educopia, Internet Archive, Library of Congress, Los Alamos National Lab and Portico, as well as team members from Loughborough University, Open Book Publishers, the British Library, the DPC and Jisc. During the event, which took place on Zoom, we engaged in group discussions about the challenges of archiving third-party content, and breakout conversations around specific issues arising from a set of titles.

When and where does the digital book end?

We kicked off the workshop with a discussion about the challenges 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.

A key issue underlining these questions is the relative benefits and limitations of preserving different formats which may contain the embedded material. One solution suggested by our workshop participants was packaging content together in digital preservation rather than expecting one format, like EPUB or PDF, to retain everything, i.e. using a tool like Bagger, which is already used in digital preservation to package digital content with metadata and documentation. A key takeaway was that there is no 'silver bullet' or single solution, and that preservation needs to take a more scattergun approach. Participants discussed this in terms of breadcrumbs that might be left, using platforms like [the Internet Archive's 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.

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? And how do you create shared understanding of the boundaries of the book?

Some of the copyright and access issues were touched on as well, and will be discussed further in a future workshop, but the initiatives of the COVID-19 pandemic have encapsulated some of the challenges represented by the various different levels of OA access and licence: many organisations have made their collections, or subsets of them, open in 2020 due to the global crisis. How can we factor in these kinds of change in status when archiving third-party content?

Book or database?

Following our opening discussion, we broke into groups to discuss specific issues around complex ebooks, raising questions about how we might keep the link between an archived book and connected resources, the best formats to preserve, and where in the life cycle interventions might be needed to facilitate archiving and preservation. One particular title we discussed, [*A Lexicon of Medieval Nordic Law*](#) by Jeffrey Love, Inger Larsson, Ulrika Djärv, Christine Peel and Erik Simensen (Open Book Publishers, 2020) was typeset programmatically, directly from a database which is also used to generate a [website](#). As the website is expected to update over time, how can we deal with versioning? If this takes us away from the realm of books and towards other forms, would it be best to encourage creators to consider books as datasets from the outset?

While this would be an extreme answer to the set of challenges raised by a book like this, our discussions highlighted wider cultural issues facing digital preservation. Some authors and publishers would strongly oppose considering a book in this way, and some still favour printing books as a method of ensuring they are preserved rather than focusing on digital preservation. Our workshop participants drew attention to issues around the willingness of content creators to engage with preservation issues, the reluctance of some organisations, and a general ideological scepticism summed up well by the questions of who we are archiving for, and for how long. How do we justify the financial and economic costs of digital preservation, and are we talking about five, ten, fifty, a hundred years?

Culture shift

While these are complex issues that will take time to solve, our participants mentioned some effective ways of developing skills and knowledge base: creating simple

guidelines and templates to start building good relationships with publishers and content creators, putting questions of digital preservation in people's minds earlier, and the possibility of partnering with funders and organisations to implement training schemes. As we're focused on academic publishing for the COPIM project, we spoke about ways of integrating training into the professional development of researchers. Advocates in specific domains and academic communities can disseminate knowledge and shape best practice, and workshop participants mentioned the importance of personal relationships, meetings in person and so on in getting content creators on board with these questions as early as possible in the publication process. In general, publishers we have spoken to have indicated that there is greater awareness and willingness to engage now than there has been previously, suggesting there is already a shift in the broader OA publishing landscape.

Looking past 2020

A key question raised by Karen Hanson, Senior Research Developer at Portico, is what is possible to preserve at scale. Hanson, who is also pursuing this question as part of a new grant on *Enhancing Services to Preserve New Forms of Scholarship* (involving NYU Libraries, Portico and CLOCKSS), has suggested essentially producing a risk assessment, that identifies problems and gives alternative routes with a better chance of long-term survival. In terms of solutions, one thing emphasised by workshop participants was that we should be flexible in the decisions we make now, bearing in mind what might change in the future. As such, packaging several documents together might be a good answer here as well, and might make it possible to make necessary changes in future rather than preserving a book as a single file. A looser definition of the 'book' might serve long-term digital preservation better.

Ultimately this is a blog post of questions rather than answers: the workshop was a first step in our work to consider what technical interventions can be made, what the legal issues might be and how we can overcome them, and ultimately how we can facilitate preservation in a cost-efficient way for small presses. The workshop has shown that we should concentrate on facilitating multiple technical solutions, alongside encouraging culture shift and encouraging communities to share knowledge. We look forward to continuing the conversation as we move forward with the project.

Thank you to all the participants for their time and input. We will be holding further workshops in the next year, including a session focused on the legal issues surrounding digital preservation of OA monographs.

Header image: *The Future of Books*, Johan Larsson. *Flickr*:
<https://www.flickr.com/photos/johanl/6966883093/>