Key issues

Archiving and preservation are not synonymous. ‘Archiving’ refers to the practice of placing files in safe-keeping somewhere aside from the publisher’s local storage, which is a means to ensure the content continues to be available even if the publisher’s server crashes. Open access archiving means the location where the book files can be found is online and is free to access to anyone with an internet connection. ‘Digital preservation’, on the other hand, is a technical process usually performed by a digital preservation archive, which assures the files are preserved for the long term.

Archiving and preservation can be challenging for small and scholar-led OA publishers who cannot afford to subscribe to expensive membership in archiving/preservation portals that offer OA-policy-compliant technical solutions. Many small and independent OA presses do not have financial resources, personnel and/or technical expertise to use these portals. This means their content is at risk of disappearing should they cease to operate. In Europe, Canada, Australia and the US, Open Access mandates (e.g. UKRI or cOAlition S) require publishers to preserve articles on professional preservation platforms, such as CLOCKSS or PORTICO. However, it is expected that shortly, monographs will also have to be preserved in the same way. Therefore, all publishers of Open Access monographs are advised to think about their book preservation plans now.

This section of the Toolkit will address the following themes:

- Digital preservation archives
- Archival formats
- Metadata for preservation
- Submitting content to digital preservation platforms

Digital preservation archives

Digital preservation is the process of safekeeping digital information. As academic publishing has transitioned from print to web distribution, new ways of preserving content (e-books, e-journal articles and digital data) had to be invented to future-proof it and to keep it for the next generations.

Open Access monograph publishers have two options for archiving their publications in digital preservation archives – directly through membership with a digital preservation archive, or via a third party, for example, a distribution platform. The archives we present here: LOCKSS, CLOCKSS and PORTICO are all non-profits involved in the so-called ‘programmatic preservation’, i.e. cross-institutional effort to preserve scholarly content in trusted repositories. These platforms use one or

two preservation methods. Files can be “bit preserved” or undergo the process of ‘normalisation’. The former method means that files will be kept in the same form they were received in, with no changes; in the latter method, original files are converted to one of a few standard and non-proprietary formats, which allows the archives to avoid managing many different formats. However, in the process of normalisation, some file properties can be changed; therefore, some archives use both methods at the same time.

LOCKSS, CLOCKSS, and PORTICO are the three major digital preservation platforms where digital content is stored for long-term access and use.

- LOCKSS (Lots of copied keeps stuff safe), run by Stanford University, supports an open-source system/software allowing libraries to collect, preserve and provide their readers with access to material published on the Web.
- CLOCKSS – Controlled LOCKSS – is a shared archive using LOCKSS technology to host digital content on the servers of academic institutions in diverse geographical locations. CLOKSS uses bit preservation. There are no definitive file-type requirements for publishers, and all file types are permitted via either file transfer or web harvesting. Publisher participation costs can also be found on the CLOCKSS website. Fees include an annual fee based on publisher revenue and a one-off setup fee.
- PORTICO is another archive run by ITHAKA, a not-for-profit organization that helps academic institutions to preserve their content. PORTICO performs the normalisation process where possible on ingested content to their internal XML standard. However, original ingested files and any related information are preserved as well. PORTICO is also preparing to archive complex and experimental monographs.

Both CLOCKSS and PORTICO are ‘dark archives’, meaning that users do not have access to them unless a ‘trigger’ situation occurs, i.e. a publisher stops their operations (or part of their operations, e.g. an e-journal) or experiences a catastrophic failure, and their content is no longer available. But libraries may use this service also for more prosaic reasons, for example, after ending a subscription, they can get access to content for which they had a subscription (e.g. older issues of an e-journal). LOCKSS, on the other hand, is a real-time backup that provides access for users even during a brief downtime. This is possible due to its complex mechanism of file replication, format migration and repair.

While long-term digital preservation is preferred, if a small open access monograph publisher does not have anything in place to assure future access, the first step is making multiple copies safe in multiple locations, for example, university repositories. Another cost-effective option is to store their content in the Internet Archive, an American digital library founded by Brewster Kahle, an American librarian and OA advocate, and based in California, US. The Internet Archive allows anyone to register for a free account and download/upload digital content. It is an open repository providing the public with free access to collections of materials including websites, software

applications, music, audio-visual and digitized print materials. These materials are not exclusively academic, and a lot of the content data is collected automatically by its web crawlers. Using the Archive-It subscription service, publishers can build and preserve collections of digital content. Through a web application, subscribers can harvest, catalogue, manage, and browse their archived collections. Collections are hosted at the Internet Archive and are accessible to the public with full-text search. Subscription cost is based on the amount of data archived annually and can range from a few hundred to many thousands of dollars per year.

Preservation through third-party distribution platforms

Some dissemination platforms offer preservation services for publishers whose content they host on their platform in collaboration with LOCKSS, CLOCKSS, or PORTICO. For example, 3 well-known platforms hosting books, OAPEN, JSTOR, and Project MUSE collaborate with PORTICO for digital preservation (Project MUSE subscribes to LOCKSS as well). OAPEN's Library database can be used by member publishers to upload books and book chapters in PDF and EPUB formats. The OAPEN Library is built using the DSpace repository software package. JSTOR is part of ITHAKA and Project MUSE is a digital distribution platform based at Johns Hopkins. All these portals charge for book distribution to libraries and knowledge bases and provide other services such as file conversion and preservation. Pricing can be found on their website.

The Thoth Archiving Network is working to offer a basic archiving option for OA monographs from small and scholar-led presses that use Thoth to manage their metadata. Thoth allows metadata to be exported to various catalogues, dissemination channels and platforms. The Thoth Archiving Network has moved from proof-of-concept work (completed during the COPIM project) to the building of a functioning network of participating university repositories and web archiving platforms, including the Internet Archive. Previously, the Thoth Archiving Network uploaded 600 works into the Internet Archive via a registered account: https://archive.org/details/thoth-archiving-network. A summary of the work performed can be found here.

Can university repositories be used as book archives?

Publishers should encourage their authors to deposit their monographs in university or subject repositories to save multiple archived copies in multiple locations. However, for those publishers who are affiliated with universities, a university repository can also be used as an archive.

Barnes and Higman reported on their attempt to use the Figshare repository to test manual and automated ingest of book metadata. The authors concluded that while manual ingest 'create[s] very specific and thorough metadata for the files, as well as to assure clearly articulated

connections between the files, both monograph text and supplementary content', the manual process of depositing books is so time-consuming that no small or medium size press will be able to do it on a regular basis due to shortages in staffing. On the other hand, automated ingest is easy to set up, and once set up, an automated workflow can be reliably repeated for large numbers of works. While more complex works (books with a lot of images, audio or video files, rich metadata) may need 'manual intervention to ensure they are correctly represented within an archive'. Therefore, even though this method of depositing is not perfect, it can be used with reasonably good results until the press has the resources to professionalise its preservation strategy.

Archival formats

Different content types, e.g. images, video, audio or text, have their own file formats with different functionalities. New formats (e.g. 3D model) and newer generations of software may lead to phasing out of older formats and software. This phenomenon is called 'format obsolescence'. Due to file-software incompatibility, data in older file formats may become unusable. Therefore, for digital preservation, book files must be future-proof. Since there is no certainty that any particular device or programme used today will exist or be common in the future, files need to be stored in formats that can be 'rendered' in the future. Rendering is how books are displayed in different apps and on different devices.

The digital formats we are familiar with are not suited for preservation. Converting files such as PDF or e-pub into one of the archival formats described below ensures that they will remain readable and accessible. However, it is important to pair the respective content type with a suitable preservation format so that its functionalities are preserved.

PDF/A

For conventional printed publications and e-books, preservation service providers recommend the PDF/A archival format. The first PDF/A-1 standard [ISO 19005-1:2005] was published in 2005, and this format forbade most embedded content (audio, video, 3D models) or dependencies external to the document, such as links to destinations outside the document (script, or 3D images, audio and multimedia). Using professional parlance, an ideal file for preservation had to be 'self-contained', 'self-dependent' or 'device independent' because any external dependence, on a link or a device, could potentially end in the loss or corruption of the file, which means that its visual appearance would not be preserved over time. A self-contained and independent file is more likely to survive regardless of the tools and systems used for creating, storing or rendering files.

The PDF/A family has progressed since this earlier version. PDF/A-2, also defined in ISO 19005-2:2011, extended the capabilities of PDF/A-1 and a new capability was to allow the embedding of PDF/A-compliant attachments. PDF/A-3 again advanced this capability, now permitting the embedding of a file or files in any format. However, for these, audio, video, and 3D artwork content

are still forbidden. Likewise, Javascript and the inclusion of executable files remained prohibited.

The most recent version of the PDF-A format, PDF/A-4, was released in November 2020. Within PDF/A-4 are subsidiary profiles: PDF/A-4f, a profile that allows files in any other format to be embedded, and PDF/A-4e (intended for engineering documents) supports Rich Media and 3D annotations as well as embedded files. Rich media in that context is defined as describing video, audio, 3-D models, or animated GIF files embedded within another file or piece of code. Embedding of rich media file formats is supported with this version, but it is worth noting that external dependencies still are not, and any embedded content must be within the file and not linked elsewhere.

In summary, overall the PDF/A standard (ISO 19005) was created primarily for archiving traditional books – this format would not be suitable for content that requires multiple external dependencies. In this sense, a fairly high percentage of experimental books and monographs would not be suited to the PDF/A standards family in terms of long-term preservation or archiving unless a static version could be agreed upon for preservation in the future. This is a conversation in progress amongst the experimental publishing community, alongside discussions of best practices around iterative work.

It is also worth noting that regular PDF formats DO NOT have the same limitations as those adhering to the PDF/A standard for archiving.

EPUB3

Unlike static PDF/As, EPUB files are ‘flowable’; that is, files have different representations depending on the reading device: tablets, smartphones, and compatible eBook readers, computers, or online websites. EPUB files depend on internet-connected readers to link to external content, which is why EPUB files usually are small enough to be distributed and kept on reading devices. Such files would be unsuitable for preservation because external links can be easily broken. The archiving format for such files is EPUB3, which can be created alongside public-facing ebooks. EPUB3 contains both XHTML files with the content and all the supporting files: graphics, interactive elements, videos, audio, etc.

HTML

HTML is a hypertext markup language used to format webpages and ‘tell’ web browsers how a document should look. Some publishers, such as Open Book Publishers and Open Humanities Press, create HTML versions of their Open Access monographs so that their books can be read online in a

web browser without downloading. Like other webpages, these can be ‘webcrawled’ by the Internet Archive and preserved for future viewing. We introduce the Internet Archive and how it can be used elsewhere in the Archiving and Preservation section of the Toolkit.

XML

XML – Extensible Markup Language – is a markup language used to describe a document’s content and its data structure. The XML can be transformed into several readable formats such as HTML, PDF or EPUB. XML offers many benefits: it is searchable, preserves content in multiple formats, can be repurposed, and is, therefore, the golden standard in preservation. However, to be able to produce books in XML, publishers need to establish an XML workflow, which requires considerable investment and expertise. XML files must be 'well-formed', which means that the data must conform to very strict rules that govern XML; otherwise, the files won’t work. Small and scholar-led publishers may find XML versions of their monographs beyond their financial and technical capacity.

Some digital preservation specialists recommend also preserving an immediately readable PDF or EPUB version alongside the XML to ensure the layout and intended structure are preserved accurately. If all versions are preserved together, given that they are being made available by the publisher, this would help cover most eventualities, including unwanted changes in the layout during rendering.

Note on standards

Following defined standards for different archiving formats is necessary to successfully preserve files in the long term and to ensure their renderability. If XML files are created in nonstandard ways, for example, using bespoke open-source software rather than standard software, this could affect proper rendering and their future useability. This is also true about PDF/As, where Adobe is recommended.

The PDF file standard was previously a proprietary format controlled by Adobe but was released as an open standard on July 1, 2008, and Adobe relinquished control of the PDF file format to ISO (the International Standard Organisation). The PDF/A is also an open standard (ISO 19005). Publishers can look for free/open-source software that is approved by the PDF Association, and they can also use the PDF Association's free PDF and PDF/A verification service online: https://pdfa.org/pdfa-online-verification-service/

Metadata

As we pointed out in book 11 of this Toolkit (Metadata management), metadata is crucial for effective book dissemination and discovery. It is also critical for preservation. Thorough and complete metadata makes it possible to find a book among millions of records in a digital preservation archive; bad metadata means a book may never be found, and the effort to preserve it will be futile.

Unfortunately, in the case of books, there are no established standards or a minimum set of requirements for generating 'thorough and complete metadata'. According to the authors of the Good, Better, Best recommendation for archiving OA monographs, at the very least, book metadata should include information in the following categories:

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

Until a consensus is reached on minimal metadata standards for books, presses are advised to make sure that these fields in their metadata records are complete.

Submitting your content to digital preservation platforms

Packing files: submission information package (SIP)

To submit a book to a preservation platform, publishers will need to prepare a submission information package (SIP) which includes

- an archiving file with the content (for example, PDF/A or EPUB3),
- a separate metadata file,
- verifiable file manifest (list of all files in each file package and their location),
- checksums (different for each file, checksum is a unique numerical signature derived from a file).

If a book has multiple formats (PDF, XML, and EPUB, for instance), all versions should be packaged.

The Good, Better, Best report (p. 36-37) recommends the following way of organising files for submission:

- Create a folder and use consistent naming conventions across all your content files. You may wish to create subfolders – one for the content (e.g. called “content”) and one for any documentation about the content (e.g. called “metadata”). Whatever your system of file naming and file hierarchy, be consistent.
- Understand what you have: Create a list of the content that is being transferred (archived/preserved). You can use software, such as DROID (which is free), to identify what you have and create a list of the content. Always include as much information as possible: file names, file paths, sizes, file formats (especially important), last modified date etc.
- Prepare a “metadata” folder and save the list in an open format (e.g. CSV or XML, rather than a proprietary format, like .xls).

Revision #3
Created 4 November 2023 18:30:33 by Izabella Penier
Updated 5 November 2023 17:49:38 by Izabella Penier