Though publishers have been taking advantage of digital book production workflows since the 1970s, the widespread availability of books as digital objects that readers access directly is a relatively recent phenomenon. Despite the experimentation and pioneering efforts of early digital text producers such as Project Gutenberg, e-books did not become widely read commercial products until the mid-2000s, when portable devices and interfaces for reading improved to the point that a viable marketplace could be maintained. With the availability of digital texts as purchasable objects came a series of linked circumstances that challenge producer-independent preservation into the future: the fluid boundaries of the e-book as a digital object, the use of digital rights management technologies and restrictive licensing agreements, and the evolution of file formats used for e-book file delivery and access.

All three of these conditions require coordinated preservation approaches that are moving beyond the resources of individual collecting libraries. With printed books, individual libraries have long taken responsibility for their
preservation in the stacks and special collections. The large number of libraries that house print books, paired with coordinated collections strategies such as shared print preservation agreements and last copy policies, could reasonably guarantee the preservation of an individual title. This perception of broad coverage and curation helps create a sense of “bibliodiversity,” to borrow a term from the world of independent publishing, that sees preservation activities happening across a range of organizations and methods. Today, the preservation of digital books is taking place on a similarly coordinated path at the level of library consortia and other national and international collaborative projects that have access to the resources and infrastructure needed to accomplish digital preservation work at large scales. Though the advantage of coordinated approaches means that resources for this undertaking can be shared, collaborative digital preservation also requires careful risk management and assessment. Fewer organizations preserving e-books on behalf of larger numbers of individual libraries means that there may be fewer preserved copies in the global digital stacks and significant potential gaps in the preserved record. This is especially true for the “long tail” of books produced outside of large multinational publishers, such as independent trade publications, open access materials, and self-published works that make up the diverse nature of digital publishing today.

The roles of regional library consortia, government depository libraries, and individual institutions are especially important in diversifying the preserved e-book record. This chapter reviews the approach we are taking at Scholars Portal, the technology service provider for the Ontario Council of University Libraries (OCUL), to preserve e-books on behalf of consortia members. I bridge the theory of e-books as digital objects and relevant theories and standards in digital preservation with an overview of our approach to translating these ideas into actionable practices for preserving a large corpus of e-book titles. The intention is to add to existing approaches and methods for e-books preservation with the goal of encouraging greater diversity and development in the field. It comes with the caveat that our approaches are very much now (and will likely always be) in development and refinement. I begin by providing an overview of e-books as digital objects, the current state of e-books preservation, and the context of Scholars Portal and OCUL’s preservation efforts to date. I then proceed to review the preservation approach we are taking with reference to the consortium’s licensing strategies, our preservation principles and conceptual workflow, and ongoing storage and management functions. I conclude with a short discussion of considerations for maintaining Scholars Portal’s Trustworthy Digital Repository certification.

E-Books as Digital Objects

An e-book is both a specific digital object and, as Eileen Gardiner and Ronald G. Musto define it, a “cultural practice” that extends to the complex relationships between authors, publishers, and technologies that result in production and dissemination.⁴ Although Alan Galey notes that the term e-book “implies a unity that does not exist in practice,”³ it is necessary for the purposes of preservation to prescribe some boundaries around the data one is seeking to preserve in order to guide the particular preservation processes that will take place. Such boundaries are suited to genres of digital objects: categories of objects that possess sufficiently common generic characteristics. A genre consists of individual member digital objects that themselves are composed of one or many constituent files or bit streams. Therefore, we have e-books as a genre of digital objects, an individual e-book title as a specific digital object requiring preservation according to the requirements of the genre, and the components that together enable the performance of that book (such as text files and images rendered together using software and hardware) that are the subject of defined preservation processes. Digital preservation processes keyed toward a particular genre of digital object ought to be specific enough to ensure that the preservation of members of the genre retain their intelligibility, form, and function as determined by the preserving organization (i.e., the object retains its “significant properties” for re-performance), but general enough to be hospitable to the smaller variations that might occur within these terms.⁴ Otherwise, new policies, processes, and workflows ought to be defined in response to new genres or subgenres. E-book preservation approaches must rely on the general predictability of e-books as digital objects, while also monitoring potential changes in how new materials received from providers might challenge previously made assumptions or boundaries. A second consideration is the preservation of e-book objects as independent from e-book-specific reading devices or interfaces. The general approach discussed in this chapter favors the preservation of digital objects independent of specific access platforms, since this is the agreed-upon form in which book files are transmitted to our organization. Furthermore, Scholars Portal’s definition of our designated community, which is a core aspect of the Trustworthy Digital Repository and Open Archival Information System (OAIS) standards, emphasizes “ease of searching, browsing, retrieval, and reuse,” which necessarily privileges accessible files over specific devices or interfaces.⁵ However, other approaches may be necessary for e-books that are more deeply entwined with their presentation method. Similarly, our focus remains on preserving the output from a provider: the e-book as
(generally) intended to be read. E-books live prior lives as production files that have additional preservation considerations from the point of view of publisher archives.\(^6\) The significant properties of e-books have not been formally defined, though a workshop on the subject was convened in 2013 at the British Library. Among the terms identified were \textit{structure, layout, embedded and linked content, versions, and metadata}, all of which resonate with our own approach at Scholars Portal.\(^7\) The Library of Congress’s “Recommended Formats Statement” section for textual works also alludes to aspects that might be inferred as significant properties, including “associated external files and fonts” and excluding “digital rights management or encryption.”\(^8\) The definition of significant properties is up to the preserving organization’s evaluation of the needs of its designated community.

For my purposes, I adapt Gardiner and Musto’s general definition of an e-book as a freestanding text or media-based publication in digital form that is produced, disseminated, and read using a combination of computer networks, systems, software, and hardware.\(^9\) Such a definition could include a number of types of e-books: digitized copies of print books, born-digital e-books with or without equivalent print-based versions, and potentially more nontraditional forms, such as multimedia-based productions. At the same time, e-books are distinguished from other text-based content by having a length and depth of content coverage that is either the work of a single author, or produced in parts or as a whole by multiple authors. Their contents typically contain paratextual aspects that signal their status as books, including title pages, tables of contents, chapter divisions, pagination, notes, references, and indexes. The general presentation of e-books currently mirrors traditional printed codices through the structured presentation of text and images, although the use of reflowable text in formats such as XML and EPUB enables some limited flexibility within this structure.\(^10\) Furthermore, texts may be enhanced by internal hyperlinks to references and footnotes, and external links or supplementary material. Some authors have signaled that e-books could potentially become more media-rich materials embedding video and sound content. New forms of e-books may involve immersive, interactive forms that blur the boundaries between text, games, and videos.\(^11\) What is key to remember is that e-books contain “book-length” and “book-like content.”\(^12\) Similarly, the Portico project’s e-book format action plan makes such a definition even more explicit: “The fundamental archival unit for e-book content is a book.”\(^13\) In short, e-books are books, which, though a circular definition, is a necessary one to establish, lest we get overly distracted by all of the content that could be a book in one way or another, and that may have more specialized preservation needs and approaches.

E-book history has its roots in the earliest development of computers and networks. A commonly cited intellectual source for e-books is Vannevar Bush’s imagined “memex,” as presented in the 1945 essay “As We May Think,” which is a device that stores a large library of linked information for easy storage and retrieval. Developments in hypertext and network technology from the 1960s to the 1980s, followed by the dissemination of books via storage devices such as floppy disks and CD-ROMs through the 1990s and early 2000s, made this vision a reality. Initiatives such as Project Gutenberg began making digital texts available in the ASCII text format in 1971. At the same time, the publishing industry began to digitize production workflows. Canada’s Coach House Press pioneered the use of the Datapoint 2200 for book production beginning in 1974, and later they were key innovators of SGML typesetting in the 1980s. Despite the widespread digitization of publishing processes by the 1990s, printed books continued to be the primary method of access until the gradual commercial availability of e-books in the mid-2000s. The e-book market is now persistent, though it continues to comprise a smaller market share than print books. According to a consumer survey by BookNet Canada, e-book sales comprised roughly 16.8 percent of the Canadian market in 2016. Publishers reported that 46 percent of e-books had no sales, a figure which may be due to a larger volume overall of e-books made available due to backlist conversion. However, they also reported a rise in digital revenue overall. Furthermore, purchase methods among libraries continue to be in flux, as experiments in open access licensing, such as the Knowledge Unlatched project, and patron-driven acquisition and a transition away from “big deal” license packages to title-by-title selection, continue to evolve. What is common to all of these developments is the continued need to access e-book files into the future: as valuable assets central to a library’s ongoing operations, and as irreplaceable cultural and historical records that deserve continued accessibility in the long tradition of books as carriers of human knowledge.

As e-books have become more available and accessible as part of the general publishing marketplace, features of their status as digital objects have resulted in a shifting ground around access and preservation. The central issue is the easy replication of digital files that threatens e-books’ value as commercial products. To ensure scarcity and therefore support enduring commercial value, producers began wrapping e-books (initially published on CD-ROMs) with digital rights management technologies (DRM) in the mid-1990s. DRM can prevent certain uses of files, such as copying and distribution, and can limit the number of users able to read a certain book at one time. The need to protect DRM itself led to amendments to copyright legislation such as...
the Digital Millennium Copyright Act of 1998 in the United States, which made it illegal to circumvent or deactivate technological protections. Michael Widdersheim dates the shift from e-book ownership to licensing to this time, as new license terms that accompanied e-books essentially eliminated the right to first sale that typically enables the lending and resale of print books.\(^{22}\) Today, libraries primarily access e-books through subscription packages.\(^{23}\) The instability of subscription-based models has led to instances where e-books suddenly become unavailable to users, such as when vendors lose the rights to distribute the works, titles become inaccessible when subscriptions cease, or when publishers withdraw their titles from collections.\(^{24}\) The initial proliferation of e-book file formats was related to the need to lock down digital files for commercial reasons, though de facto industry standards such as the EPUB 3 format appear to be emerging.\(^{25}\) Proprietary file formats enable producers to lock users into the format of their choice, thereby restricting them to the devices and access methods of the producer’s choosing. The British Library’s e-book file format summary assessment report lists 30 file formats, 12 of which were either declared obsolete, have had their support ended, or have been replaced.\(^{26}\) Preserving institutions will have to review the kinds of formats in their collections, including the versions of these formats, in order to determine appropriate preservation approaches.

**E-Books Preservation Review**

E-books have been an area of interest for digital preservation in the library domain since studies began in the early 1990s. The initial focus was in response to rapidly degrading print books that were published with inferior-quality paper in the nineteenth and early twentieth centuries.\(^{27}\) Book digitization posed a potentially promising solution to a serious preservation issue. As Lisa Fox wrote in the introduction to the 1996 edition of *Preservation Microfilming*, “The technology now exists to scan textual and graphic materials and to store them as images or as text files . . . The accessibility of scanned materials offers an exciting application of our reformatting systems.”\(^{28}\) Fox offered some familiar caveats in favor of continuing to use microfilm as the preservation copy: optical and tape-based media have short life spans and require frequent data migration, and standards for the long-term maintenance of digital files had not been established,\(^{29}\) though the work that would become the OAIS standard was getting under way at the time.\(^{30}\) Early projects, such as Yale University’s Project Open Book (1991–99) and Cornell University’s
digitization-to-microfilm project (1990–93), focused on the initial definition of digital preservation as digitization and used microfilm in their workflows. Both projects acknowledged that long-term maintenance of the new digital surrogates was required, but neither explored this topic further. At the same time, concerted research efforts in the field of preserving born-digital materials were beginning with efforts such as the Commission on Preservation and Access and the Research Libraries Group’s Task Force on Archiving of Digital Information (1994–96). Among its findings, this task force’s final report noted that “there is little good experience yet in storing in digital form massive quantities of materials traditionally regarded as culturally valuable, such as books and serials.”

Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving, a 2002 Council on Libraries and Information Resources report, included a chapter on e-books that identified some of the key problem areas that were then coming to the fore, including the rise of DRM and the lack of ownership of books’ files, the variety of access devices and formats, and the problem of determining who should be responsible for preserving them. As Frank Romano notes in his chapter in the report, “the likelihood of long-term preservation by commercial enterprises may not be as assured as is preservation by certain libraries.”

Action in the area of e-books preservation has generally lagged behind that of e-journals preservation. The reasons behind this have yet to be explored in detail, but it is likely due to a few factors. E-journals were very rapidly embraced by producers and readers, whereas e-books have not enjoyed the same quick uptake. Secondly, e-books may be somewhat more complicated as a genre to preserve, relating in particular to what Amy Kirchhoff and Sheila Morrissey call their “hybrid and sometimes hazy definition” and the potential for more complex aggregates of files that comprise a single book. As a consequence, work in e-book preservation has generally been applied after serials preservation, though, of course, the two have much in common.

In 2005, the Association of Research Libraries released its statement “Urgent Action Needed to Preserve Scholarly Electronic Journals,” which goes into detail about many of the threats to serials preservation in higher education. Though many of the issues articulated in this statement apply to e-books (and many other different kinds of digital materials), books are not mentioned. Much e-journals preservation research in the United States has been funded by the Andrew J. Mellon Foundation beginning with the “Archiving Electronic Journals” project, which ran from 2000 to 2003. Several outcomes of the Mellon funding in 2000 have had a large impact on e-books preservation: the LOCKSS project and the Portico project. These two programs take fairly

---

different approaches to digital preservation, though both depend on agreements with providers to enable preservation activities to take place. LOCKSS is a distributed system that depends on libraries installing the software locally and selecting materials to enter into the system themselves, while Portico is centralized and ingests materials directly from providers. Both are similar in that the preserved content itself is not accessible to users until certain “trigger events” occur, though the specific trigger mechanisms differ between the two projects. LOCKSS (which stands for “Lots of Copies Keep Stuff Safe”) is an open source peer-to-peer software platform that was developed in 1999 for e-journals, but it has been applied to all kinds of digital content, including e-books. It uses a web-harvesting approach to capture a local copy of content to which the participating library has access, which is then verified across the same content stored on other LOCKSS boxes that participate in the global network. As such, there could be many copies distributed over the network.

The LOCKSS trigger mechanisms relate to the general accessibility of the resource, such as when a website is down, and are commonly used as a backup access method for this reason. LOCKSS also resulted in a separate organization, CLOCKSS (Controlled LOCKSS), in 2006, which was founded in a collaboration between LOCKSS, academic publishers, and research libraries. CLOCKSS is essentially a specific private LOCKSS network that ingests and preserves material on behalf of libraries. Materials are collected both by web crawling and transfer processes directly from providers. As such, it represents something of a midpoint between global LOCKSS and Portico. Portico originated in JSTOR’s electronic archiving initiative and was set up in 2005 with Mellon funding. Portico’s approach is to load e-book, e-journal, and other digital collections content (such as digitized newspapers) from providers directly and process it for long-term preservation. Providers and libraries subscribe to Portico’s service individually. Content is not accessible to anyone other than Portico staff, though title lists are made available for analysis, until a “trigger event” occurs, or if publishers enable Portico member libraries to use Portico as a perpetual access method. Portico’s trigger events have a higher threshold for releasing access compared with LOCKSS. They relate to when a publisher ceases operations, when a title ceases to publish and has not been transferred to another publisher, or when materials become inaccessible on vendor sites and are not available otherwise, in addition to “catastrophic failure . . . for a sustained period of time.”

Portico, LOCKSS, and CLOCKSS are all fairly well-developed to the point that in the 2011 edition of No Shelf Required: E-Books in Libraries, the chapter treating preservation simply points to the three programs as sufficient to cover
a library’s needs. However, this scope is limited to whatever proportion of a library’s collections is covered by these programs through their agreements with content providers and publishers, and concern has been expressed about both the lack of coverage and potential overlap between programs. A few studies have been published on this matter regarding e-journals. I could not locate any published coverage studies treating e-books specifically. Serials preservation assessment research conducted by libraries has found a wide range of coverage values for LOCKSS and Portico. A 2007 collection development report at Rutgers University Libraries noted that approximately 50 percent of their purchased e-journals holdings were preserved by the two services combined. A 2011 assessment at Cornell University Library found that LOCKSS and Portico together preserved about 13 percent of e-journals holdings. A 2014 report by the Consortium of Swiss Academic Libraries found that an average of 56 percent of e-journal holdings were preserved by Portico. Of course, these figures are apt to change over time as LOCKSS, CLOCKSS, or Portico gain or lose agreements with producers. A recent effort to improve data in the area of e-journals preservation coverage is the Keepers Registry (www.thekeepers.org), which was founded in 2011. No comparable effort for e-book preservation is under way, and it is unclear if there is any demand for such a service.

In terms of e-book preservation research on a broader scale, Portico has been a primary driver in the field. Portico’s Amy Kirchhoff and Sheila Morrissey have published several articles and reports on e-books preservation topics. Research and practice outside of the United States has been driven primarily by national libraries in Europe. These include the British Library, the National Library of France, and the National Library of the Netherlands’ e-Depot service, which are preserving books received through digitization and legal deposit mechanisms. A related area is the preservation of theses and dissertations in institutional repository systems such as DSpace and Islandora. DSpace workflows using Archivematica for digital preservation functions are actively used. Finally, organizations such as the Internet Archive and HathiTrust focus on the preservation of digitized books, though their workflows are somewhat different from born-digital e-books in that they are preserving master scans at the level of individual pages. It is clear that the e-books preservation landscape is far from diverse, particularly in terms of content produced outside of the large multinational publishers that service academic libraries. Kirchhoff and Morrissey have signaled that there is “no clear ownership of the responsibility for the . . . preservation of mass-market and non-scholarly trade books” in the absence of enforceable government depository mandates in certain jurisdictions, and due to the fact that public libraries typically subscribe to, rather than

purchase, content. Where institutions and organizations feel they have a mandate and resources for preservation, public libraries, regional consortia, and government depositories should stake a claim in e-books preservation in order to ensure a diversity of approaches and content sources. OCUL and Scholars Portal have emerged within Canada as one key center for the digital preservation of published content. The remainder of this chapter will review the approach to digital preservation at OCUL and Scholars Portal, from the history of our program’s development to our licensing strategies and preservation approach as a potential model for coordinated approaches in e-books preservation.

Digital Preservation at OCUL and Scholars Portal

The Ontario Council of University Libraries (OCUL) is a 21-member consortium of academic libraries in the province of Ontario, Canada. OCUL members range from large research universities, such as the University of Toronto (approximately 85,000 full-time equivalent students), to smaller undergraduate institutions, such as Algoma University (approximately 1,200 students). In total, OCUL covers over 484,000 university students, which is at least 30 percent of all the university students in Canada. The consortium engages in collective licensing, collaborative planning, advocacy, and professional development for its members. Among the consortium’s functions is the shared funding of Scholars Portal, the information technology service provider for OCUL members.

Scholars Portal was founded in 2002 and was initially developed through seed money provided by the provincial government of Ontario. Scholars Portal is a department of the University of Toronto Libraries, which acts as a service provider to OCUL. The emergence of Scholars Portal and OCUL as a site for digital preservation is due to a number of circumstances and priorities that developed gradually within the consortium and at the University of Toronto. Prior to the advent of Scholars Portal, the University of Toronto had been negotiating for the right to load copies of digital journal articles onto local servers beginning in the 1990s. This effort was intended to improve access to these materials, since publisher platforms were notoriously unreliable at the time. This approach was carried forward as OCUL moved into collective licensing in the late 1990s. OCUL began journal loading on a locally administered system in 2001, a practice that set the groundwork for Scholars Portal’s “light archive” access model discussed below.

coincided with the founding of Scholars Portal’s journals service in 2002, which gave a single point of access to these materials to licensed users. A similar path was followed for books, and the Scholars Portal books platform was released in 2009. In 2007, it was recognized that the large collection of e-journals materials required preservation oversight. Steve Marks, the first digital preservation librarian at Scholars Portal, was hired to coordinate these activities in 2010. In 2013, the journals repository was certified as a Trustworthy Digital Repository according to the TRAC checklist by the Center for Research Libraries. Reflecting the general trend of e-serials preservation occurring before e-books preservation, various circumstances, including gaps in staff resources and changing organizational priorities, delayed the development of e-book preservation requirements until 2016. A recent project to redevelop the e-books platform included key functional requirements pointing to e-books preservation, though e-book preservation processes are independent of the platform itself. The Scholars Portal e-books platform was initially developed using eBrary, which is based on the ISIS document management system. In 2011, ProQuest purchased eBrary and support for the system ceased. It was decided that a new platform would be developed in-house to address various needs emerging from users (i.e., both librarians at member institutions and readers), including improving rights management and the reading experience. Following a two-year development project, a beta version of the new books platform was released on May 14, 2018, and a full release for all users followed on July 2, 2018.

**OCUL’s Licensing Approach**

The high priority placed on access within OCUL has meant that e-book and e-journal preservation takes place in a “light archive” model. In this context, a light archive is one where materials are made available to those users who are entitled to access them. Rights are managed via entitlements modules in the relevant access platforms. By contrast, a “dark archive” does not provide access to materials until a predefined time, if ever (see the discussion of “trigger events” above). The advantage of the light approach means that what is preserved is transparent to users, and users have the opportunity to validate the contents of the archive immediately, as opposed to some future date when access might be granted and quality issues are difficult to fix. It is reasonable to expect that in many cases, sales rights will be transferred from provider to provider for as long as materials remain in copyright, meaning that most materials in dark
archives will remain inaccessible for decades to come. In addition to dedicated automatic and manual quality assurance processes, the Scholars Portal e-book and e-journal platforms receive frequent queries from users that help ensure the continued quality of the collection. Furthermore, the light archive approach meaningfully fulfills one of the key terms in both the OAIS and Trustworthy Digital Repository standards: that the contents of the archive are understandable to the defined designated community. Since users are constantly accessing the materials, issues in understandability can be flagged immediately.

A key to the ability to preserve journals content has been the use and adoption of model licenses that contain the required terms for preservation. OCUL provides two such licenses: a comprehensive model content license that contains the relevant terms for access and use of materials, and a local archiving and hosting agreement that enables the transfer of files from producer to archive and contains preservation-specific terms. Licensing issues are at the core of the present challenges in e-books preservation. The ability to negotiate for perpetual access models as opposed to subscriptions has emerged as one feature of the power of consortia in managing e-book licensing. The ability of OCUL to do so resides in the license agreement that is signed between the provider (the licensor), OCUL (the licensee), and the University of Toronto representing Scholars Portal (the service provider). These terms are related to three primary functions: the right to collect and store local copies of titles (local load); the right to provide access to users who have purchased access to the content, even after subscriptions cease (perpetual access); and the right to migrate content to new formats in order to prevent obsolescence (portability). I will discuss these three main preservation rights in turn, plus additional rights that support them.

Local load is both the name of the practice of storing local copies of purchased materials, as well as the short name of the agreement that is signed. The “local archiving and hosting agreement” is attached to a content agreement as an appendix, which means that it can be negotiated separately from OCUL content agreements and it can be attached to other agreements that a library might have with national licensing bodies or through a locally negotiated purchase. A “local load light” agreement is available for noncommercial providers that wish to load their content for access and preservation as well. The basic agreement for local load is the promise that the licensor will provide a copy of the relevant licensed materials to the service provider within an agreed-upon time frame.

Perpetual access rights are contained in both the content agreement and local load agreements. There is a distinct content agreement for e-books, as well as another for e-journals and database materials. The perpetual access
licensing term in the e-books content agreement is as follows: “Except for termination for cause . . . Licensor hereby grants to Licensee a nonexclusive, royalty-free, perpetual license to use the Licensed Materials that were accessible during the term of this License.” In the local load agreement, under the “Term” section, the language is as follows: “The Licensed Materials will remain accessible through the Service Provider in perpetuity.” It should be noted that these perpetual access clauses are somewhat different from those guaranteeing perpetual access on a vendor platform, which is referred to in an “Archiving” section of the OCUL content license as a second method of ensuring ongoing access. Perpetual access in this context is the right to provide continued access via the locally hosted platform. Perpetual access is a contentious term in the context of licensing scholarly materials. Jim Stemper and Susan Barribeau’s 2006 review of the subject defined perpetual access as explicitly separate from preservation rights, with the note that few institutions were exercising local load rights due to the cost and difficulty of doing so. However, we seek perpetual access rights as a key support for the ability to physically preserve digital materials. The Charlotte Initiative for Permanent Acquisition of eBooks by Academic Libraries project includes “Provision of irrevocable perpetual access and archival rights” among its three key principles for e-books.

Portability (also called “transformation”) rights are contained in the local load agreement and reserve the right to migrate content to new formats. The term language is as follows: “Service Provider will maintain the integrity of the Licensed Materials but shall have the right to migrate the Licensed Materials to new formats, in response to technological changes, in order to ensure ongoing access to Authorized Users . . . Licensor acknowledges and agrees that Digital Rights Management technology shall not interfere with this right.”

A series of related rights supports the three main preservation rights. Providers must supply DRM-free source files so that the negotiated access and preservation activities can take place. Scholars Portal reserves the right to preserve an archival copy of material that has been withdrawn. If the licensor loses ownership of the materials, it will make best efforts to ensure that Scholars Portal retains the right to archive and provide access to the materials, which references the TRANSFER code of practice (another service initially designed for e-journals whose applicability to e-books has yet to be defined).

on negotiated access rights mean that Scholars Portal has had to integrate DRM within its access platform. In most cases, DRM simply takes the form of restricting access to licenced users or restricting the amounts that can be downloaded to certain portions of a book. For other content, such as books loaded from the Association of Canadian University Presses, a licensing arrangement between the parties imposes additional access restrictions such as limited concurrent use, which is managed by the platform.\textsuperscript{66}

OCUL seeks consistency for preservation terms across licenses. However, terms are subject to negotiation, other agreements have predated the model license approach and have not been renegotiated as of yet, and some may differ in the actual language that is agreed upon by the parties. Generally, preservation-specific terms tend to be quite stable. “Publishers have learned that we have core requirements if they wish to do business with us,” Tony Horava writes. “As the scale of licensed content on Scholars Portal has grown, and our model has become widely known, publishers have become more familiar with our interests and requirements.”\textsuperscript{67} The presence of the three key licensing terms is linked to the level of preservation accorded to a specific e-book. Objects with the first two rights present (loading and perpetual access) are accorded bit-level preservation, which means that Scholars Portal commits to preserving the files at a bit level, that is, with the continued assurance that the files remain unchanged. Objects with all three rights (loading, perpetual access, and portability) are accorded full preservation, which means that file formats will be monitored and migrated as necessary to ensure continued access. About half of all e-book licenses are accorded full preservation status versus bit preservation status, and the intention is to continue to review licenses as they come up for renewal in order to provide for portability rights, which were missing from earlier licensing negotiations. The particular considerations for preservation as the three licensing principles are put into practice will be discussed below.

**E-Books Preservation Overview**

This section takes the reader through the general workflow intended for e-book preservation. As it stands, this workflow is yet to be developed: the intention is to develop, test, and implement it after the work to resolve any outstanding issues with the books access platform is complete. The total number of e-books currently eligible for preservation at the time of writing is 368,979 titles. Of this, 7 percent are in full-text XML format (from two providers), while the other 93 percent are in PDF format. A handful of EPUB files are also in the
mix. A project to provide initial insight into PDF versions using JHOVE is also under way. We are exploring the use of the preservation package generator Archivematica to process e-book files for preservation in contrast to our e-journals workflow, which uses a process built in-house, in order to create more robust preservation metadata (particularly PREMIS metadata) and to simplify the development process. Our intention is to improve the existing e-journals preservation workflow by using the experience gained through developing the e-books workflow.

Our preservation process takes advantage of the concept of the Archival Information Unit (AIU) and Archival Information Collection (AIC) as described in section 4.2.2.4 of the OAIS standard as a method to organize and manage multiple potential versions or portions of a single e-book that may arise, such as additions, updates, and corrections. Though we preserve the “final” output of providers, this output contains its own instabilities, since e-books are subject to post-release errata and additions. The ability to handle updates and corrections was signaled by books support staff as a key issue in the preservation process. As such, our intention is to accommodate e-books that are more fluid in nature, such as a book that is slowly built by new chapters over a period of time or one that is updated regularly, which is a publishing model that may become more common in the future. An AIU is a specialization of an AIP representing a “the content and metadata for a single ‘atomistic’ object.” This object can contain multiple files. In this case, an individual e-book submission as received from a provider is an AIU. More than one AIU representing the same book object or a portion of that object (that is, it has the same ISBN rather than being a new edition) enables the creation of an AIC to relate them to one another. An AIC “consists of multiple AIUs that have been grouped together into a distinct collection.” An AIC can be a conceptual link: it contains metadata that describes relationships and points to AIUs without physically containing them in the AIC itself.

Transfer

An e-book begins its life with a vendor or publisher that has the right to provide the data that makes up the book. After the licenses have been signed, processes are set up to facilitate the transfer of the licenced materials to Scholars Portal. This typically involves a transfer of book files and metadata from a provider’s FTP site or the sending of files to a local FTP site by the provider. Once the files are received, they are checked and the loading process begins. Loaders are scripts customized to providers that are run against received files, normalizing...
metadata from the source to the Books Interchange Tag Suite (BITS) standard and unpacking and organizing the files into individual book titles for upload into the platform. As part of the initial preservation workflow, the e-book preservation process will compare the file size of files on the originating site to the file size of the received files as a basic integrity check. Checksums will be generated for all received files upon receipt. While it would be ideal to have producers create checksums or apply the BagIt standard for transfer, it is currently beyond our ability to enforce such an approach. The future negotiated use of web-based deposit tools, however, might enable such a process to take place in order to provide greater proof of integrity during transfer. In general, Scholars Portal receives e-book files comprising the book either in a single PDF file or split into chapters, or a full-text XML file plus figures, in addition to publisher-supplied descriptive metadata and, potentially, supplementary files. All of these together comprise the e-book object.

From Transfer to SIP

In order to create a Submission Information Package (SIP), a number of processes must run on the files prior to their ingest into Archivematica. First, it must be validated that the received e-book will indeed be loaded. We sometimes receive book files in error, and it is important to verify first that we actually have the right to preserve and make accessible the files. The first step is verification of the loading of a book into our books database, which uses a MarkLogic NoSQL system as its back end. Once this is completed, the book has been loaded into the access platform and the preservation workflow can begin. First, descriptive metadata will be recorded in a CSV file that is formatted appropriately for ingest into Archivematica. We have made a provisional mapping of BITS to Dublin Core for descriptive metadata in the final METS file contained with the AIP, but we may also evaluate using a simplified BITS record in place of Dublin Core. Secondly, PREMIS rights information will also be imported via CSV. Finally, a processing file to automate the Archivematica transfer will be placed in a designated folder. This processing file directs certain preservation actions, such as normalization, which would be related to the preservation level defined by the licensing terms as discussed above. Once this process is completed, a package containing the original book files and associated metadata files will be transferred to a watched folder for automatic ingest into Archivematica. Archivematica continues the SIP processing as part of its transfer functions by validating checksums; identifying, characterizing, and validating files; and beginning to write metadata to the METS file, among other functions.

From SIP to AIU

As part of Archivematica’s ingest process, the option to normalize files is given. For e-book files in PDF format, we are not currently exploring the normalization of PDFs to PDF/A, since the quality of programmatic normalization to this format has not been proven. It is clear that normalization to PDF/A without access to source files such as fonts may create more problems than it solves at the moment. For book packages with supplementary files, these will be normalized using Archivematica’s standard preservation policies unless the book has bit-level preservation rights only. More important for long-term preservation is investment in strong file identification, characterization, and validation tools, including the general-purpose JHOVE tool for PDF identification and validation (jhove.openpreservation.org), and the newly developed VeraPDF tool (www.verapdf.org) for PDF/A validation (for PDF/A files that we receive directly from providers). Having robust data relating to the file formats in the repository will mean better-informed preservation actions as they become necessary down the line. Once an AIU is stored, preservation metadata will be written to the MarkLogic books database. More than one AIU will be related by an AIC record that is also created in the database.

DIP Creation

Access derivatives come in several forms: PDFs in chapter form for download by users, and HTML files for reading within the platform. The book-loading process generates both access versions where applicable. HTML files are rendered using a tool called pdf2htmlEX, which are readable via the platform in an infinite scroll. Updates and corrections will be integrated into access versions. Finally, some preservation metadata such as checksums will be surfaced to users as a method of providing transparency about the archive. The Scholars Portal journals platform currently makes preservation metadata available to users.

Storage and Management

AIUs will be stored on OCUL’s OpenStack Swift-based cloud storage network, the Ontario Library Research Cloud (OLRC, cloud.scholarsportal.info). The OLRC consists of five nodes located at institutions throughout Ontario. Materials uploaded to the OLRC are instantly replicated to three of the five nodes and are constantly checked by the system for integrity. We may develop a method to harvest Swift’s system integrity checks into our MarkLogic database, or we may run separate periodic fixity checks. Metadata updates, including
descriptive metadata updates, may be periodically written to the stored METS file by taking advantage of Archivematica's metadata-only re-ingest feature, but we are unsure if fixity checks would be included in the scope of these updates.

Trustworthy Digital Repository Certification

Since Scholars Portal's e-journals platform is already certified as a Trustworthy Digital Repository (TDR), the intention is to retain this certification to include e-books as part of a regular disclosure to the Center for Research Libraries. There is little that will need to change at the highest level, since items such as the designated community definition and collection policy documents will remain stable. However, our TDR documentation was initially written for journals-specific content and workflows. As a consequence, journals-specific language is mixed in with broader language in response to the TDR certification statements and supporting documents. As we update our documentation to reflect e-book preservation practices, we will be making our documentation more modular in order to reflect potential future updates with different content types. Recent internal research found that 17 of the total 109 responses to individual TDR requirements embedded journals-specific language, with the majority (12 responses) focused in the digital object management section. To avoid this, each statement will be split into a general statement of practice; provide room for multiple format-specific statements; identify relevant roles, responsibilities, and anticipated risks where applicable; and provide a link to relevant documentation in an effort to make the documentation more hospitable to content updates in the future.

Conclusion

This chapter has demonstrated a path for e-books preservation that could be replicated among libraries and service providers, from broad principles to actionable processes. It has identified a licensing approach that has found success to date in our organization. And it has attempted to draw a lineage for e-books preservation from initial experimental digitization work in the 1990s to the current state of somewhat disconnected archiving agencies and processes that have yet to preserve the true diversity of e-books production. What is exactly being preserved, in what form, where, and by whom has become a strangely difficult question to answer in the world of e-books. The challenges of file formats and DRM will continue, though in general these issues are stabilizing.
as providers gain comfort in the marketplace and standardization proves to have commercial value from the perspective of users. Indeed, we will likely see more stabilization in the formats of e-books, but less stabilization in the forms of e-books: e-books that push the boundaries of the book as we know it. There are also strong opportunities to link print preservation strategies with digital ones in order to provide a holistic view of the preservation of a library’s collections on a macro scale. With a greater diversity of institutions taking a share in e-books preservation, e-books may one day shed that awkward “e” qualifier, and return to their former status as books in their own right, preserved as comfortably as their print counterparts are today.

Acknowledgments

Many thanks to the Scholars Portal staff who have worked hard on supporting e-book preservation to date and who provided information for this chapter: Ravit David, Kate Davis, Bartek Kawula, Sadia Khwaja, Annie Thomas, and Jacqueline Whyte Appleby.

NOTES


15. Ibid., 296.
19. Ibid., 20.
20. Ibid.
22. Ibid.
28. Ibid., 17.
29. Ibid., 18.


45. Rutgers University Libraries, “LOCKSS, CLOCKSS and Portico.”


48. Mary Elder, e-mail message to author, August 29, 2017.


68. Lavoie, Open Archival Information System, 14.
69. Ibid.