Preserving eBooks

Amy Kirchhoff and Sheila Morrissey

DPC Technology Watch Report 14-01 June 2014

Series editors on behalf of the DPC
Charles Beagrie Ltd.

Charles Beagrie

Principal Investigator for the Series
Neil Beagrie

DigitalPreservationCoalition

DPC Technology Watch Series

ISSN: 2048-7916

DOI: http://dx.doi.org/10.7207/twr14-01

First published in Great Britain in 2014 by the Digital Preservation Coalition.

Foreword

The Digital Preservation Coalition (DPC) is an advocate and catalyst for digital preservation, ensuring our members can deliver resilient long-term access to digital content and services. It is a not-for-profit membership organization whose primary objective is to raise awareness of the importance of the preservation of digital material and the attendant strategic, cultural and technological issues. It supports its members through knowledge exchange, capacity building, assurance, advocacy and partnership. The DPC's vision is to make our digital memory accessible tomorrow.

The *DPC Technology Watch Reports* identify, delineate, monitor and address topics that have a major bearing on ensuring our collected digital memory will be available tomorrow. They provide an advanced introduction in order to support those charged with ensuring a robust digital memory, and they are of general interest to a wide and international audience with interests in computing, information management, collections management and technology. The reports are commissioned after consultation among DPC members about shared priorities and challenges; they are commissioned from experts; and they are thoroughly scrutinized by peers before being released. The authors are asked to provide reports that are informed, current, concise and balanced; that lower the barriers to participation in digital preservation; and that they are of wide utility. The reports are a distinctive and lasting contribution to the dissemination of good practice in digital preservation.

This report was written by Amy Kirchhoff, Portico's Archive Services Product Manager, and Sheila Morrissey, Senior Researcher at Ithaka, both engaged in eBook preservation and access at Portico: the work of Portico is described later in this report. The report is published by the DPC in association with Charles Beagrie Ltd. Neil Beagrie, Director of Consultancy at Charles Beagrie Ltd, was commissioned to act as principal investigator for, and managing editor of, this Series in 2011. He has been further supported by an Editorial Board drawn from DPC members and peer reviewers who comment on text prior to release: William Kilbride (Chair), Janet Delve (University of Portsmouth), Sarah Higgins (University of Aberystwyth), Tim Keefe (Trinity College Dublin), Andrew McHugh (University of Glasgow) and Dave Thompson (Wellcome Library).

Acknowledgements

The authors thank Marcel Ras (formerly Program Manager for the International e-Depot, now Program Manager for the Netherlands Coalition on Digital Preservation (NCDD)), Trudie Stoutjesdijk (Senior Advisor for Collections), and Victor-Jan Vos (Team Leader, Collection Development) of the National Library of the Netherlands (KB) for discussing the KB's experience in the acquisition, management, and preservation of eBooks in the Netherlands.

We thank Leslie Johnston, Chief of Repository Development, Library of Congress, for sharing the Library of Congress's experience in collecting and managing eBooks; and Jeremy York, Assistant Director, HathiTrust, for his contribution on the experiences of HathiTrust.

We thank Caroline Dawson, Project Manager and Sharon Johnson, Head of Content Development Implementation, both of the British Library, for sharing their experience with us.

We thank Richard M. Davis, Manager, Academic & Research Technologies University of London Computer Centre (ULCC) for helpful insights into some of the preservation challenges presented by enriched eBooks, including novel experimentations with the capabilities of the eBook form.

We thank Stephen Abrams, Associate Director, University of California Curation Center, and other International Internet Preservation Consortium (IIPC) members who responded to his request for information, for data on the validity of HTML content harvested by IIPC members.

We also thank the anonymous peer reviewers of this report for their many helpful observations and suggestions.

And, finally, we would like to thank Neil Beagrie, Director of Consultancy, Charles Beagrie Ltd, for his advice and his many, very constructive, suggestions as this report was developed.

Contents

1. Abstract

This report discusses current developments and issues with which public, national, and higher education libraries, publishers, aggregators, and preservation institutions must contend to ensure long-term access to eBook content. These issues include legal questions about the use, reuse, sharing and preservation of eBook objects; format issues, including the sometimes tight coupling of eBook content with particular hardware platforms; the embedding of digital rights management artefacts in eBook files to restrict access to them; and the diverse business ecosystem of eBook publication, with its associated complexities of communities of use and, ultimately, expectations for preservation.

2. Executive Summary

The rapid convergence of technological advances, including the use of the Internet as a channel of content delivery, an increasingly 'digital native' population with new expectations (such as efficient automated search, retrieval, and re-use of information), and cost pressures on the production and storage of new publications, have made the eBook as a mode of publication a fact on the ground for the foreseeable future. Along with those preservation challenges common to all digital objects, eBook preservation entails some distinctive issues, many of which arise from its hybrid and sometimes hazy definition, embracing both digitized and born digital content. These issues include legal questions about the use, reuse, sharing and preservation of eBook objects; format issues, including the sometimes tight coupling of eBook content with particular hardware platforms; the embedding of digital rights management artefacts in eBook files to restrict access to them; and the diverse business ecosystem of eBook publication, with its associated complexities of communities of use and, ultimately, expectations for preservation.

The provision of long-term, permanent access to eBooks that have been licensed is ill-defined, and ownership of the responsibility for the preservation of different large categories of digital artefacts that fall under the rubric of eBooks is not clearly established. Nor are the costs for carrying out the preservation and establishing sufficient permanent funding to meet those costs.

The right to permanent possession, including perpetual access and preservation rights, is the exception rather than the norm in eBook licensing. Permanent access and preservation is further threatened by the widespread use of digital rights management (DRM) technologies in licensed artefacts. The stability of this leased content, even in libraries, is not assured: content can be modified, and even withdrawn under the terms of use of most eBook licences.

There are sometimes quite serious preservation risks associated with the formats in which eBooks are created. This is particularly the case for proprietary formats, particularly those closely tied to an individual commercial vendor's hardware platform and distribution system. The variants, deliberate and inadvertent, in even standard eBook formats such as PDF and EPUB call to mind the 'browser wars' of the early days of the Internet or the even earlier videotape format wars of Betamax versus VHS. There is also as yet no well-established scheme for uniquely and consistently identifying eBooks. Nor are the tools for validating and characterizing all these formats fully mature.

Large-scale digitization of print books has created valuable and widely used digital surrogates for those books, that are being put to uses (rapid search and discovery, text mining across corpora of collections) impossible with print books. However the scale of digitization has introduced quality assurance issues both for the digitized images and for the creation of descriptive metadata for these books. They have also, in some cases, embroiled institutions in legal entanglements arising from both the eBook's similarity to, and difference from, its print source.

Libraries and other preservation institutions will need to work with publishers to ensure that:

- explicit comprehensive rights to all objects embedded within an eBook are transferred along with rights to the eBook;
- there are explicit protocols for propagating updates to eBook content. This would include explicit negotiation for removal or darkening of – restricting public access to - collection items;
- there is publisher and preservation institution conformance to open standards for eBook formats, metadata, and identifiers. This would include, at a minimum, articulation and documentation of usage for standard formats like ONIX;
- there is minimization or elimination of digital rights technologies for restricting access to eBook content;
- they understand what long-term, permanent access options are available for each eBook;
- policies are articulated to ensure coverage of categories of eBook content by preservation institutions;
- there is co-ordination amongst institutions, both to establish there are no unintentional preservation gaps, and that there is no needless duplication of preservation efforts;

- eBook leases guarantee preservation rights, and prohibit DRM technologies in the preservation copy acquired from the vendor; and
- there is investment in maturing existing file format characterization tools (such as JHOVE[1] and DROID[2]), and in extending that toolset if it is determined to be necessary for the preservation community to develop the capacity for emulation (software that duplicates the function of other hardware and software) for rendering eBook formats closely tied to particular hardware platforms.

[1] See https://bitbucket.org/jhove2/main/wiki/Home

[2] See http://digital-preservation.github.io/droid/

3. Introduction

What is an eBook? Most simply, we might describe it as a ‘digital analogue’ of what we have come to perceive as a book, as that perception has developed out of our collective and increasing familiarity with the codex form over the last two millennia. Understood this way, an eBook is an instance of a ‘paper and ink’ book in an electronic medium.

The extreme plasticity of digital objects, however, means that this analogy has not been a strict one; nor is it likely to become one. Gardiner and Musto (2010), commenting on the still-evolving characteristics of eBooks, have observed both the elaboration of delivery mechanisms of the textual content of a book via audio equivalents, as well as the enrichment of that textual content with multimedia. Romano (2002) defines an eBook as simply ‘the presentation of electronic files on digital displays’, whether that content would have been presented in non-digital form as a book, a magazine, a newspaper, or a catalogue. He sees part of the functionality of an eBook as the ability to create a representation of the notion of a ‘page’ appropriate to each of those formats, adjusted to the size of the device’s screen; although the notion of a page has, since the publication of Romano’s article, become increasingly problematical with the creation of rendering tools that operate across diverse systems with different form factors (the physical size and shape of a device).

The eBook’s development was an emergent phenomenon as much as a deliberate goal, beginning with the conceptualization by Vannevar Bush of the Memex (a personal device with a compressed collection of all of a user’s various textual artefacts, including books, notes, and records) (Romano, 2002). Michael Hart’s digitization of the American Declaration of Independence as the first entry in Project Gutenberg[3] in 1971 is often conventionally marked as the first eBook (Hart, 1992).

As Manley and Holley (2012) detail in their history of the eBook, the eBook’s capabilities and features have evolved along with advances in both specialized and general reading and rendering devices, and with the means of delivering content to those devices (for example, locally attached optical discs, web-accessible static content, and streamed delivery over the network). eBook developments included various (often multi-media) products, appearing roughly between 1985 and 1990, typically released on optical discs. Examples include supplementary textual and visual content for audio compact discs, commercial encyclopaedias, and, memorably for the classic preservation challenges it ultimately presented, the BBC Domesday Book project[4].

Document content management – the need to create a copious number of frequently updated manuals associated with its hardware and software – led to the development at IBM of Generalized Markup Language (GML), the precursor to Standard Generalized Markup Language (SGML), and ultimately to the development elsewhere of perhaps the best-known SGML vocabulary, HyperText Markup Language (HTML).[5] The development of GML was followed in turn by the creation of commercial products to read and navigate such electronic documentation (Goldfarb, 1996). Beginning in the 1990s, an increasing number of purpose-built ‘book player’ devices, specialized for reading a digital analogue of a physical book, appeared on the market. A great deal of technical effort was expended on improving display technologies to approximate the experience of reading print on paper, and miniaturization to make the device at worst no heavier than a paperback book (Manley and Holley, 2012).

Simultaneous with these developments was John Warnock’s ‘Camelot Project’ (Warnock, 1991), proposed in 1991 at Adobe Systems[6]. Intended to solve the problem of making documents ‘viewable on any display and ... printable on any modern printer’, it became the Portable Document Format (PDF). Consequent upon Adobe’s decision to distribute its Acrobat Reader project free of charge, the format became ubiquitous, especially for distribution of document content on the burgeoning World Wide Web. Its ubiquity in turn fuelled the print-to-digital transition in scholarly journal publishing, beginning in the mid-1990s, and is currently used by most of those same scholarly publishers for scholarly monographs and other book-length publications.

[3] See http://www.gutenberg.org/wiki/Main_Page
[4] See http://www.atsf.co.uk/dottext/domesday.html
[5] See http://www.w3.org/community/webed/wiki/HTML/Specifications for links to specifications of versions of HTML
[6] http://www.adobe.com/

Google announced the Google Books Library Project[7] in 2004. This project undertook mass digitization of books in print, with the contents indexed for search, beginning with the holdings of the University of Michigan Library,[8] Harvard University Library,[9] Stanford University Libraries,[10] Oxford's Bodleian Library,[11] and the New York Public Library. The project has raised numerous issues, for example, the legality under various copyright regimes of certain uses of the artefacts of the project, as well as the status of the (sometimes faulty) digitization artefacts as objects or versions of objects, of information and study. The immediate accessibility of these collections to search and to (variously restricted) read and process, along with the availability of other large-scale book digitization projects such as Project Gutenberg, the Million Book Project's[12] Universal Digital Library,[13] and the Open Content Alliance,[14] has had an incalculable but quite discernible effect upon the level of expectation for, and increasing frequency of use of, digital surrogates for print books (St. Clair, 2008).

The convergence of technological advances, including the use of the Internet as a channel of content delivery, an increasingly 'digital native' population with new expectations (such as efficient automated search, retrieval, and re-use of information), and cost pressures on the production and storage of new non-digital publications, have made the eBook, as a mode of publication, a fact on the ground for today and the foreseeable future (Romano, 2002). Along with those preservation challenges common to all digital objects, eBook preservation entails some distinctive issues, many of which arise from its hybrid and sometimes hazy definition, embracing both digitized and born digital content. These issues include legal questions about the use, reuse, sharing and preservation of eBook objects; format issues, including the sometimes tight coupling of eBook content with particular hardware platforms; the embedding of digital rights management (DRM) artefacts in eBook files to restrict access to them; and the diverse business ecosystem of eBook publication, with its associated complexities of communities of use and, ultimately, expectations for preservation.

[7] http://books.google.com/googlebooks/library/
[8] http://www.lib.umich.edu/michigan-digitization-project
[9] http://hul.harvard.edu/hgproject/index.html
[10] http://lib.stanford.edu/google-books/statement-support-participation-2005
[11] http://www.bodleian.ox.ac.uk/dbooks
[12] https://libwebspace.library.cmu.edu/libraries-and-collections/MBP_FAQ.html
[13] http://www.ulib.org/index.html
[14] http://www.opencontentalliance.org/about/

4. Issues

4.1. eBook Ownership

There is some question as to whether one can even speak of ‘selling’ and, correspondingly, ‘owning’ eBooks. With traditional books, an individual or library which purchases the print book owns the book, owns the ink on paper that makes up the book, and has the right to resell the book. With eBooks, the purchaser does *not* own the book or the bits that comprise the book. Those bits live in the cloud and are licensed to the purchaser for use. Amazon makes this extremely clear in its Terms of Use, declaring that ‘Kindle Content is licensed, not sold, to you by the Content Provider’ (Amazon, 2012). The nature of eBook purchases or licences where the content resides in the cloud and is never truly in the possession of either the individual or library purchaser brings serious questions to promises of long-term access that eBook publishers, sellers, or aggregators may make. What does it mean, for example, for a library to ‘purchase’ a book where perpetual access is only available if the library continues to pay its platform service fees? The implications of this business model for libraries are discussed further below in Section 4.2.2. Significantly for the preservation of eBooks, commercial licence agreements to date typically do not make explicit licence provisions for preservation rights, nor for the permanent and irrevocable transfer of the digital objects comprising the eBook to a preservation repository (Lynch, 2013).

4.2. The Business Models of Selling Books

eBooks are marketed and sold in different ways to different communities. These marketing and sales avenues tend to parallel the ways in which print books have been, and continue to be, marketed and sold. As preservationists, it is important to understand these differences, as the business models under which books are sold have implications for both the formats used for publication, and for expectations of the different ‘markets’ or communities of use for their preservation.

4.2.1. Mass-Market and Trade eBooks for Individuals

In the print world, there is a significant difference in how mass-market books and trade books are marketed, the physical quality of the artefact produced, and how they are distributed. The differences between mass-market and trade were outlined by the *New York Times Sunday Book Review* when they created a separate bestseller list for trade books in 2008. As explained by Elsa Dixler (2008), mass-market books are marked by lower production values and smaller ‘form-factor’, and typically are produced in large quantities. Usually affordably priced, they are sold in many venues, including supermarkets and newsstands. Trade books, on the other hand, generally have higher production values (including paper and binding); are often produced in smaller quantities, and typically are sold in bookstores (‘to the trade’ – hence ‘trade’ books). The distinctions noted by Dixler – in print production values, quantities printed, sales venue – are in general not applicable to eBooks. However, individual consumers of the eBook analogues of trade versus mass-market print books *might* continue to be distinguished by these same categories.

4.2.2. Mass-Market and Trade eBooks for Libraries

Public and academic libraries also license mass market and trade eBooks for use by their patrons (either to download to a patron-owned device or to download to a device the patron borrows from the library). When a public library purchases a print book, it has the right under copyright law to lend the book. With eBooks, however, as noted by Fryer (2013), ‘libraries do not own these [eBooks] outright. Instead they must negotiate licensing deals for each book they want to lend. They put the e-collections on servers run by computer firms such as OverDrive and 3M’. In the absence of universal government mandates for the legal deposit of all eBooks into national or other libraries, and given the lease rather than ownership business model for these particular eBook purchases, there is a risk that there will be no clear ownership of the responsibility for the preservation of mass-market and non-scholarly trade books and no guarantee of the long-term accessibility of these eBooks to the libraries communities.

Libraries and publishers are still experimenting with how to purchase or license eBooks and then how to lend them to patrons, including such disparate options as:

- books that expire after a certain number of loans;
- limited simultaneous use;
- books that expire a certain amount of time after they are licensed by the library;
- state and country wide licensing deals; and
- libraries paying publishers each time a book is lent out (Fryer, 2013).

4.2.3. Professional and Scholarly eBooks in the Scholarly Market

eBooks have existed in the professional and scholarly market for significantly longer than they have existed in the mass or trade book market. In fact in 2009, ‘professional and scholarly publishing titles represent[ed] 75.9% of the US eBook market, or $1.33 billion’ (Clarke 2009). The business models supporting eBooks produced for academia are quite diverse. Not only do the publishers of the books sell or license them to libraries, but eBooks in the professional and scholarly market tend to be widely available via a substantial number of aggregators – with each aggregator having, perhaps, different rights to the books.

Within the scholarly market (which includes trade books), publishers are still experimenting with different models for selling individual books and bundles of eBooks to academic libraries. The larger publishers have integrated eBooks into their journal delivery system and, from a user perspective, eBook chapters look remarkably like journal articles. Both smaller publishers and larger publishers are using aggregators as scholarly and professional eBook distribution methods. Whether purchased or licensed through the publisher or an aggregator, this type of eBook tends to be made available through a web-based platform. Typically, this content is not tailored for download to a personal reading device. Many aggregators are promising perpetual access to licensed or purchased eBooks, but as the libraries are never in possession of the actual books, there is no true guarantee of this promise.

The business models for selling professional and scholarly eBooks to academic libraries are diverse and include such options as:

- selling ownership of the eBooks to libraries;
- licensing eBooks to libraries with unlimited, simultaneous use;
- licensing eBooks to libraries with a limited number of simultaneous users; and
- patron-driven acquisition (PDA) (where libraries set up an account with the publisher or aggregator and the library patrons purchase books, until the limit on the account is reached).

This variety of ownership and licensing models presents preservation institutions with challenges. For example, if a library licenses an eBook for limited use, does that library have any preservation rights to and expectations of long-term access to the eBook? Can aggregators meet their promises of perpetual access when they do not have long-term preservation rights to the books? Preservation institutions are taking on the task of preserving eBooks for the long term for specific publishers. However, the access rights to these preserved eBooks for libraries which may have purchased or licensed the books through aggregators are not well established, and libraries should ensure their agreements with aggregators and publishers include plausible perpetual access and preservation clauses.

4.2.4. The Use of eBooks in the Scholarly Market

The most recent ITHAKA Faculty Survey (2012) investigated the area of eBook usage among university and college staff. The survey found that ‘Scholars are engaging with scholarly monographs in digital format, as 70% of faculty respondents indicated that they have “often” or “occasionally” used scholarly monographs in electronic format in the past six months . . . and only about 10% indicated that they have not done so at all, with little variation between disciplines' (Schonfeld *et al*., 2013). The survey also showed that academics prefer to use print books for cover-to-cover reading, whereas searching for a specific topic or exploring references was considered easier to do in a digital format.

4.2.5. Self-Publishing

With the advent of Kindle Direct Publishing,[15] NOOKPress,[16] and iBooks Author,[17] it is quite simple for an individual to publish an eBook. Self-publishing is increasing dramatically. As Klems (2013) notes, 'Bowker Market Research data shows that the number of self-published books released each year has grown exponentially, totalling some 235,000 print and e-books in 2012 (a conservative statistic, as the absence of ISBNs on many e-books makes them hard to track)'. eBooks self-published on one of the major platforms do not present any additional technical challenge for preservation, as they are produced and distributed in one of the standard eBook formats. However, they are a considerable institutional challenge for preservation, as there is no simple, scalable way to contact and negotiate with all these independent authors in order to discuss such issues as a permanent licence for preservation for this content.

4.3. Expanding Definitions of eBooks

4.3.1. eBooks with Embedded and Networked Content

Many eBook formats support embedded audio or movie files, embedded fonts, and other types of embedded content. The eBook format becomes what is termed, on the Library of Congress 'Sustainability of Digital Formats' website, a 'wrapper format'. [18] Comprehensive preservation of such enriched eBook wrappers would entail traversing the content graph – the data structure - of the eBook, locating and identifying embedded objects, and taking appropriate preservation actions on the components within that wrapper, in addition to preserving the textual 'intellectual content' of the 'book' component of the eBook. As an additional complication, there may be independent intellectual property rights associated with those embedded objects distinct from those associated with the text.

Whether or not a preservation institution requires such comprehensive preservation of all components of an eBook artefact will be a policy decision. Preservation institutions must carefully consider what is technically feasible, what comprises an acceptable submission, and what levels of preservation they will support. It may be necessary for them to insist that their content providers adhere to guidelines that limit what format types may be embedded into an eBook. Additionally, preservation institutions will need to determine if their right to preserve an eBook includes the right to preserve objects embedded within that eBook.

Additionally, eBooks may contain hyperlinks to content or components outside the eBook itself. The preservation institution similarly must articulate its provisions for linked content, including determining what actions it intends to take with respect to preserving the permanence of those links, and to what content referenced by hyperlinks in an eBook comes under its preservation mandate.

4.3.2. Textbooks

Textbooks are rapidly becoming digital. They comprise an emerging category of an extended eBook definition, potentially containing both embedded and networked content, some of which (for example, annotations) are custom to each reader of the book. They are an instance of the potentially disruptive effects of eBooks on academia and on pedagogy in general.

[15] https://kdp.amazon.com/

[16] https://www.nookpress.com/

[17] http://www.apple.com/ibooks-author/

[18] http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

Many electronic versions of textbooks are much more than the print book in electronic form. Often they comprise modules to be integrated into a course management system, a massive open online course (MOOC) or other eLearning courses. They are subject to the preservation challenges listed above for eBooks with embedded and networked content, while presenting additional challenges for preservation to institutions that must determine to what degree any individual reader-specific customizations should be preserved.

4.3.3. Reference Books

Electronic reference books come in many forms. They might be published as just another trade eBook, in conventional eBook formats. However, often they are online databases, with information that is continually updated. Examples include the electronic version of the *Oxford English Dictionary*, various encyclopaedias, pharmacology references, and so on. This content is also deserving of preservation, but unless it is published in conventional eBook formats, it will require a completely different preservation regimen, more closely resembling the preservation of relational and other databases, or dynamic web content.

4.3.4. Digitized Books

There is an extraordinary amount of activity in the digitization of books originally published in print. National libraries, including the British Library, the French National Library, the National Library of Norway, the National Library of the Netherlands, and the National Library of Latvia, have long been involved in the digitization of primary sources, including books. In the current environment, with relatively inexpensive digitization broadly available to individuals and institutions, a tremendous amount of digitization activity is taking place both on a grand scale (see below) and on smaller scales in libraries across the world.

The larger digitization projects include:

- Project Gutenberg, mentioned in the Introduction, which is one of the earliest efforts focused on the mass digitization (including manual keying of the source book) of historical books. As of December 2013, Project Gutenberg had more than 42,000 free, out-of-copyright eBooks available for download in a variety of formats. Many of these books were created by manual transcription to electronic text. These books are available in various image and text formats (including ISO-8859/ISO-Latin, UTF-8, UTF-16 and other encodings for plain text), with generated files available for download in PDF, EPUB (for use on non-Kindle devices and applications), MOBI (for use on Kindle devices and applications), and numerous other options.[19]
- In 2004, as also mentioned in the Introduction, Google and several university and research libraries initiated the massive Google digitization program to digitize entire libraries.
- In October 2008, many of those university and research libraries that had participated in the Google digitization program came together to fund HathiTrust to store and manage their copies of this digitized content. As of December 2013, HathiTrust was managing more than 5.5 million digitized books. As noted below in Section 6.3, the primary preservation image formats for books in HathiTrust are ITU G4 TIFF and JP2.
- The Internet Archive digitizes and stores books on behalf of the Open Content Alliance, in addition to storing copies of books in Project Gutenberg. It provides a scanning service that creates JP2000 images and associated MARC and Dublin Core metadata, PDF for viewing, along with OCR (including DJVU XML).[20]

On a smaller scale, Jisc has long funded digitization programs in the UK;[21] similarly, in the US, much digitization has been funded by the NEH.[22] In addition to externally funded projects, many libraries are digitizing their own special collections.

Several of the national libraries have taken on the digitization of historical books in partnership with commercial entities. These arrangements are often structured so that the commercial entities (such as Gale

[19] See http://www.gutenberg.org/wiki/Gutenberg:File_Formats_FAQ
[20] See http://archive.org/scanning
[21] See http://www.jisc.ac.uk/whatwedo/programmes/digitisation.aspx
[22] See http://www.neh.gov/divisions/preservation

or ProQuest) provide delivery and access to this content. Access is provided at low or no cost back to the institution that digitized the content, but the content licence agreements typically allow the commercial entity to charge others for access to the content via subscription or purchase. Examples of this include Early English Books Online (digitized by the British Library, and access provided through ProQuest), and Eighteenth-Century Collections Online (also digitized by the British Library and access provided through Gale).[23] As discussed in Sections 4.1 and 4.2, it is important to articulate preservation rights and responsibilities as part of these engagements with commercial partners. In some cases, explicit provision for preservation has been made; for example, Gale and Adam Mathews Digital are preserving many of their collections in Portico.

Some challenges are unique to this content as distinct from born-digital eBooks. Books digitized as part of the Google Books project have suffered from quality assurance issues, including poor metadata and some defective page images (Nunberg, 2009). Paul Conway (2013), in addition to investigating methods of detecting and mitigating some of these defects, raises the interesting question as to the intellectual status of these digital objects, particularly if they are defective, and especially as they are considered in their roles of new paradigms of use, such as large-scale text mining.

These digitized books entailed a considerable capital investment by the community, and comprise a considerable cultural and intellectual asset, making consequent demands on the resources of the preservation community. Preservation institutions who have taken this content under their institutional mandates (national libraries, the HathiTrust) have undertaken the challenge of keeping this content safe for the long term, and in a cost-effective manner.

4.4. Digital Rights Management (DRM)

Digital Rights Management (DRM) is a set of technologies employed to protect commercial intellectual property rights in digital content, whether that content is an eBook, a piece of recorded music, a video, or any other digital artefact (Liu *et al* .,2003). It enforces the use of digital licences, which restrict a consumer's access to a digital object in certain ways, including frequency and duration of access to the object, as well as restriction of rights of transfer, or the ability to copy the object. DRM can also be used to warrant the tamper-resistance of the object. It typically entails encryption of the bit stream comprising the object, though it may be confined to implanting a digital watermark in the object to trace users of the object, thus making digital piracy detectable. There are a number of DRM schemes in use with eBooks, including Adobe Adept DRM, Apple FairPlay, and proprietary DRM employed by Amazon and by the now-defunct Microsoft Reader's proprietary format.[24] In many cases, DRM is managed by a third-party, rather than directly by the eBook publisher.

DRM constitutes a challenge in the preservation of eBooks. It is challenging from the business model perspective – if a book is 'sold' with limited use, what preservation rights come with that sale? And it is also challenging from a technical perspective, where DRM can impede the preservability of an eBook. Further, a change in digital rights technologies can be seen as a specialized case of format obsolescence. This has already threatened to become an immediate issue even for current owners of eBooks and other content under DRM, when Adobe announced a non-backwardly compatible change to its DRM technology.[25]

As noted in Section 6.1, some institutions themselves apply DRM in the delivery version of certain eBook content. If however DRM – and, more particularly, its encryption – cannot be removed by preservation institutions from the preservation copy of an object, that object will be opaque to future viewers. Preservation in such cases would be byte-level preservation of an unintelligible digital object.

[23] See http://www.bl.uk/reshelp/findhelprestype/prbooks/britprcoll1501to1800/britishprintedcols.html
[24] See http://en.wikipedia.org/wiki/Digital_rights_management#DRM_and_e-books
[25] See http://www.the-digital-reader.com/2014/02/03/adobe-require-new-epub-drm-july-expects-abandon-existing-users/#.UvE9QbRO2vA

4.5. Content Stability

It is not only the reference books discussed in Section 4.3.3 that are subject to more or less continual content modification. While the ability to provide almost instant updates and corrections to an edition of an eBook can understandably be seen as an attractive design feature of the format, it complicates the task of preservation institutions. Will these updates happen automatically and, potentially, silently? How many versions of an eBook will a preservation institution be required to preserve?

Most seriously, what happens when the modification of an eBook is its retraction or withdrawal? This occurred, famously, in 2009, when Amazon deleted some editions of George Orwell's novel *1984* from the Kindle devices of customers who had purchased them.[26] Memory institutions will need to be able to ensure the stability of eBook content in their collections, and maintain control of any withdrawal or de-accessioning of that content.

4.6. Business Models to Support Preservation

As with the preservation of any digital resource, preservation of eBooks is not free. It is expensive to identify content for preservation, gather it, perform initial actions on it, and then preserve that content for the long term.

There are a number of different exemplars of approaches to sustainable preservation of eBook content, including:

- Collective model. HathiTrust is an example of this approach. Launched initially by the 13 member universities of the Committee on Institutional Cooperation (CIC) and the University of California System, it now has more than 90 institutional partners (Christenson, 2011). The member institutions, whether consortium or individual institution, have agreements with the University of Michigan (which has legal responsibility for the repository), but arrangements through HathiTrust with one another. These partners share basic infrastructure costs, with an attempt to pro rata those costs to participating institutions based in part on 'an evenly distributed share of the cost to support public domain volumes in HathiTrust', and a pro rata cost for the number of in-copyright volumes in the collection that overlap with volumes held in an institution's print collection.[27] HathiTrust's operations and programs are supported by membership fees from the partners. No grant funding supports these operations and programs, though institutions participating in HathiTrust may receive grants on their own, which HathiTrust may leverage (as is the case with Copyright Review work at Michigan).
- Subscription service. Portico[28] and CLOCKSS[29] are examples of this approach. The costs of preservation are supported by subscription fees from participating libraries and publishers.
- Government support. The national libraries of the United Kingdom, France, the Netherlands, and other European countries are examples of this approach.

Each of these approaches is of course sensitive to losses in its primary source of funding. Each typically looks to diversify or supplement those primary sources of funding with other sources, such as grant funding.

4.7. Legal Deposit

Many countries require publishers to deposit publications with the national library. The broad publication of eBooks is a new enough phenomenon (the first Kindle was sold in November 2007, although some scholarly publishers were making eBooks available on their delivery platforms before then) that, country by country,

[26] http://www.nytimes.com/2009/07/18/technology/companies/18amazon.html?_r=0
[27] See http://www.hathitrust.org/cost
[28] See http://www.portico.org/digital-preservation/
[29] See https://www.clockss.org/clockss/Home

laws regulating the legal deposit of eBooks are still being resolved. As early as 1996, a study by a Working Group of the Conference of Directors of National Libraries determined that a number of countries (including Canada and Norway) were already requiring the deposit of electronic books (CDNL, 1996). As was the case then, countries have varying degrees of legal deposit requirements (CDNL, 2010). In some countries, such as the United Kingdom, eBooks must be deposited.[30] In other countries, such as the United States, only eBooks that have no print counterpart must be deposited with the copyright office; otherwise the print version must be deposited (as it is considered the best edition). The Netherlands has had great success with its voluntary deposit system (van Trier, 2006). National libraries that accept or require the deposit of eBooks see it as their mission to preserve these books for their citizens. Very often, that long-term preservation limits access to the reading rooms of the national library.

5. Technology and Standards

When considering the preservation needs of eBooks, it is necessary to understand the more common formats and standards used in their publication and distribution. eBook formats belong to one of two types: fixed layout and reflowable text. Fixed layout formats are most closely linked to eBooks' print ancestry. The structure of a page, including the relative position of its components, remains fixed across the devices used to display it. Reflowable text, on the other hand, adapts its presentation to the form factor of the device on which it is displayed. The concept of the page might be all but lost in such a format. With some qualification, PDF would commonly be characterized as a fixed layout format. Plain (ASCII) text, HTML, EPUB[31] and other purpose-designed eBook formats described in the following sections, are reflowable formats. These purpose-designed formats, like EPUB, are largely based on eXtensible Markup Language (XML)[32] or the XML-based Extensible HyperText Markup Language (XHTML).[33]

In addition to the 'intellectual' or text content of a book and information necessary for specifying layout and display, eBook formats typically include other information, such as bibliographic markup (title, author) and structural markup (heading, section). This information, along with possible supplementary digital objects such as those described in Section 4.3.1, may be encapsulated within the eBook format (for example, within a PDF or an EPUB file). In the absence of such encapsulation, preservation institutions will need to determine the 'packaging' rules for associating the components of an eBook.

5.1. General Formats Used for eBook Publication

5.1.1. HTML

As noted in the introduction, historically HTML emerged out of early formulations of document formats. HTML, particularly if enhanced by Cascading Style Sheets (CSS),[34] can comprise a reflowable format for publishing eBooks. HTML tags however lack the expressiveness to provide the sort of enriched bibliographic markup available in other formats. Commercially published eBooks typically are not formatted in HTML, although, for example, Project Gutenberg does offer its content in HTML.

HTML5, the latest version of HTML, introduced a number of new elements that are useful (for example, as employed in XHTML in EPUB3) for creating reflowable and media-enriched eBook content. These additions include elements for navigation, bi-directional text, embedded audio and video, and structural elements such as 'article,' 'aside,' 'header,' and 'footer'.[35]

[30] http://www.bl.uk/catalogues/search/non-print_legal_deposit.html
[31] See http://idpf.org/epub
[32] See http://www.w3.org/TR/REC-xml/
[33] See http://www.w3.org/TR/xhtml1/
[34] See http://www.w3.org/Style/CSS/
[35] See http://www.w3.org/TR/html5-diff/

At present, of course, there are multiple viewers for rendering HTML. The very many tools used to create HTML, and the varying proficiency of those who use those tools, means, however, that a large percentage (in some collections, over 90 per cent) of web pages harvested by the International Internet Preservation Consortium (IIPC), for example, do not conform to any version of the HTML specification (Abrams, Stephen 2014, pers. comm., 25 April). There is the further complication of the division of HTML standards development between the World Wide Web Consortium (W3C) and the Web Hypertext Application Technology Working Group (WHATWG).[36] [37] As HTML is at base a simple text format, it seems unlikely that this constitutes a grave preservation risk, at least for the textual content of an eBook in HTML. There is the risk for the loss of some semantics (for example, visual cues for emphasis provided by the tags for bold or italic text, or colour changes effected by CSS), as well as the possible loss of any 'behaviours' introduced by embedded scripts, if future browsers do not support the scripting language, or if the resources invoked by the script are no longer available.

5.1.2. PDF

As described in the Introduction, PDF was designed with the intent of preserving a page's image across different devices – originally, printers and computer displays. Often a part of production work flows for print products, it became an unsurprising choice for the publication of book content in electronic form.

The format is widely used, with an enormous number of creating and rendering applications, as well as an open specification that has moved under the International Standards Organization (ISO) umbrella. It has the capability (although that capability is not always, or even often, employed) to embed the sort of descriptive bibliographic metadata described above. It is also capable of enrichment via embedded and linked digital objects, with those objects and links encapsulated within a single file. The PDF specification is however complex, and occasionally ambiguous, resulting in varying interpretations, of varying degree of conformity to the specification, by the developers of PDF readers and writers. This complexity is thought by some to comprise a threat to its long-term viability as a format, even in its archival profiles (the various versions of the PDF/A standard[38]) (Morrissey, 2012).

5.2. Formats and Standards for eBooks Read on Dedicated Readers and Applications

There are two primary eBook formats in use (along with variants of each) by dedicated eBook readers and applications:

- MOBI,[39] originally developed for eBook readers on Palm devices and purchased by Amazon in 2005, is one format. MOBI allows DRM protection to be used, and supports reflowable content (fixed format text is supported only by inserting an image of a page, (mobipocket.com, n.d.)). Reflowable content in a MOBI package is included as XHTML. MOBI also includes CSS for rendition of the XHTML content. MOBI files are delivered as compressed binary files. The format may include advanced navigation controls and supports indexing. MOBI also allows JavaScript[40] and frames. A MOBI eBook may have either a .mobi or .prc file extension. Amazon's KF8[41] and AZW[42] are two variants of MOBI specific to Amazon's Kindle eBook reader device; they both use a different DRM scheme from the

[36] See http://www.w3.org/TR/html5-diff/#history
[37] See http://wiki.whatwg.org/wiki/FAQ#What_is_the_WHATWG.3F
[38]See http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920 (PDF/A-1), http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920 (PDF/A-2),and http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57229 PDF/A-3)
[39] See http://www.mobipocket.com/dev/article.asp?BaseFolder=prcgen
[40] See https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference
[41] See http://wiki.mobileread.com/wiki/KF8#The_Format
[42] See http://wiki.mobileread.com/wiki/AZW

standard MOBI format. AZW does not allow JavaScript (KF8, the more recent Kindle format, does allow it).

- EPUB[43] is the other eBook format in use by dedicated readers and applications. EPUB is a standard of the International Digital Publishing Forum (IDPF).[44] EPUB 2, released in 2007, was a successor to Open eBook Publication Structure (OEB), developed in 1999. Following a maintenance release in 2010 (EPUB 2.0.1), EPUB 3 was released in 2011, with significant improvements over earlier versions of the format. The disabled community was intensely involved in the development of EPUB 3. The DAISY Consortium endorsed EPUB 3 as a distribution format (The DAISY Planet, 2010). Although EPUB was created to be a general purpose packaging and encoding format, it is currently most widely used to support eBooks. EPUB supports both reflowable and fixed format text (using XML and metadata, as opposed to PDF).[45] Reflowable content is provided as an XHTML serialization of an HTML5 profile, along with CSS style sheets. EPUB may include advanced navigation controls and supports indexing. It is distributed as a compressed, binary file in ZIP format (although with an .epub rather than a .zip file extension).

Additionally, there is a variant of EPUB, the iBooks format. iBooks is the format produced by the Apple iBooks Author application. Though based on EPUB, it has extensions that make it proprietary, and prevent it from being read or edited on applications outside iBooks and iBooks Author.

It should be noted that Kindle devices and applications will not render books in EPUB format, whereas Nook, iBooks, and Kobo will not render books in MOBI (or KF8 or AZW) format. End users may use a desktop application such as Calibre[46] to convert from one format to another (a book downloaded as an EPUB from Project Gutenberg can be converted to MOBI via Calibre and then placed onto a Kindle or Kindle application for reading).

5.3. Formats and Standards for eBooks Read on Publisher Platforms

Many platforms deliver books to end users as HTML and PDF. Typically, the books are encoded in XML format in the publisher's systems and then transformed for display purposes into HTML with CSS used to style the HTML. For eBook markup, most publishers currently use either the NCBI Book Tag Set[47] or a version or variant of their own proprietary, internal electronic journal XML format. While their scholarly books are available on Amazon in a MOBI-based format, publishers tend to deliver the XML+PDF version of the eBooks to preservation institutions.

Many publishers will deliver their scholarly books to preservation institutions segmented into chapters. Each chapter is provided as a separate PDF file, and each chapter's PDF file is referenced from the XML markup. Sometimes publishers will provide the books both in a chapter-based and a book-based package (the entire book is provided as a single PDF file, and, in addition, each chapter is provided as a separate PDF file). When publishers provide books only in a book-based package to preservation institutions (without PDFs of each chapter), often the accompanying book metadata is provided as XML in the ONIX[48] format.

Very few scholarly books are delivered to preservation institutions as full-text XML at this time. Rather, the XML provided with the PDF image of the book content contains 'header-only' bibliographic metadata. This differs from the content provided in formats designed for dedicated eBook readers and applications – as for example both MOBI and EPUB are based on an expectation of full-text XHTML content for the book.

[43] See http://idpf.org/epub
[44] See http://idpf.org/
[45] See http://www.idpf.org/epub/fxl/
[46] See http://calibre-ebook.com
[47] See http://dtd.nlm.nih.gov/book/
[48] See http://www.editeur.org/83/Overview/

5.4. Other Formats

OEB or Open eBook Publication Structure (OEBPS)[49] was developed in 1999, and superseded by EPUB in September 2007. OEB encodes book text in XHTML, using CSS for styling the XHTML. The files that comprise the eBook are compressed into a ZIP archive. OEB informed the development of both EPUB and MOBI.

Microsoft LIT[50] was a format created by Microsoft to be used in its Microsoft Reader. Microsoft stopped producing this reader in August 2012. LIT is a proprietary format based on Microsoft's Compiled HTML Help format (CHM) and is DRM enabled.

The DAISY – ANSI/NISO Z39.86 standard[51] was developed by the DAISY Consortium to support the needs of individuals with vision impairment. The DAISY standard specifies the format and content of the electronic files that comprise a digital talking book (DTB), as well as the requirements for DTB playback devices. A book in DAISY format is an audio substitute for the print. The format is based on MP3[52] and XML and includes sophisticated features that allow for precise navigation. DAISY endorsed EPUB 3 as a suitable distribution method for accessible books; however it encourages the production of accessible content in DAISY4 (The DAISY Planet, 2010).

The Text Encoding Initiative guidelines (TEI)[53] are a set of guidelines and XML schemas developed within the digital humanities section of the scholarly community. P5 is the current release of the TEI guidelines, which include an XML schema that allows for the representation of texts. Few publishers use TEI, but it is in relatively wide use among libraries and museums doing digitization work, and is also used by Project Gutenberg.

5.5. Format Challenges

As with other content types, one must consider whether one should simply preserve eBook content in the format in which it was originally published or provided to the preservation institution, or normalize the content to another format and then save both instantiations, or save only the normalized content. In general, resources permitting, it is advisable to keep the original files even when choosing to convert to a more robust preservation format.

eBook formats with open specifications can at least potentially be validated independently of publisher or vendor platforms by preservation institutions. The various HTML, XML, PDF and EPUB formats can be identified by standard characterization tools such as DROID[54] and the UNIX file utility.[55] Given an openly available XML document type definition (DTD) or schema file, as for the NCBI Book tag set, XML eBook content can be validated by standard XML parsers, including those employed by JHOVE[56] and JHOVE2.[57] JHOVE and JHOVE2 can validate and perform feature extraction on HTML and XHTML files. The W3C and WHATWG also provide online HTML validation. There is at present no PDF validator that is completely conformant with the PDF standard, although much useful characterization information, including information about some violations of conformance, can be obtained from JHOVE and from Apache's PDFBox[58] tool, as well as from the 'Preflight' tools provided by Adobe's Acrobat Pro software. Encouragingly, Duff Johnson, Vice Chairman of the PDF Association, has been urging industry support for the development of a free and open-source, standards-conformant PDF validator.[59] The EPUB validator tool has shown encouraging signs of

[49] See http://www.digitalpreservation.gov/formats/fdd/fdd000054.shtml
[50] See http://en.wikipedia.org/wiki/Microsoft_Reader
[51] See http://www.niso.org/workrooms/daisy/Z39-86-2002.html
[52] See http://en.wikipedia.org/wiki/MP3
[53] See http://www.tei-c.org/Guidelines/P5/
[54] See http://www.nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm
[55] See http://unixhelp.ed.ac.uk/CGI/man-cgi?file
[56]See http://jhove.sourceforge.net/
[57] See http://www.jhove2.org
[58] See http://pdfbox.apache.org/
[59] See http://duff-johnson.com/2014/01/24/are-your-documents-readable-how-would-you-know/

improvement (van der Knijff, 2013) over the past year, suggesting an increasingly viable ecosystem of EPUB tools.

We have mentioned above some of the risk factors associated with PDF as a preservation format for eBooks. There is a growing consensus for the desirability of converging on standard open preservation formats for eBooks: in particular for EPUB 3 and the NCBI Book Tag Set. EPUB 3 is most appropriate for reflowable text. One of the benefits of EPUB 3 is that it not only provides for XHTML mark-up of the eBook, but it also requires a certain structure for packaging of the eBook (and working with the unique packaging of many different publishers is one of the most costly aspects of preservation of publications). The NCBI Book Tag Set provides a more robust set of metadata for books than is possible in EPUB 3. The authors recommend both be used in tandem, as appropriate for the content (EPUB 3 for the visual layout, packaging, and full text of the document and the NCBI Book Tag Set for robust metadata about the eBook).

Although Amazon’s Kindle formats (MOBI, AZW, and KF8) dominate in the market of books for use on dedicated readers or applications, the proprietary nature of those formats, especially in combination with their DRM regimes, comprise a long-term preservation risk. EPUB 3 meets many of the requirements for a preservation-robust format. It has the endorsement of the DAISY consortium, natively supports MathML,[60] and is compressed using the widely-used ZIP format (which allows it to be decompressed by any ZIP tool). However, analysis by Johan van der Knijff at the National Library of the Netherlands (van der Knijff, 2012), updated recently (van der Knijff, 2013) suggests that concerns remain (although they are lessening) about reader support, especially for EPUB3, and about the stability of EPUB’s underlying format specifications (in particular, for XHTML5 and CSS3), which are still somewhat in flux. As with other formats, DRM would have to be removed for preservation purposes.

As mentioned in Section 4.3.4, the digitization of historical collections, including books, has resulted in large collections of image and text files, including TIFF, JGP, JPEG200, plain text in various encoding, and text marked up in a number of XML vocabularies. There have been several projects to research and articulate guidelines for digitization intended for preservation, including the imaging guidelines developed as part of Metamorfoze, the joint venture of the National Library of the Netherlands and the National Archives of the Netherlands to digitize and preserve works on paper (van Dormolen , 2012). Similarly, in the United States, the Federal Agencies Digitization Guidelines Initiative[61] was undertaken to develop a common set of practices for federal agencies engaged in digitizing historical content. The Federal Agencies Still Image Digitization Working Group have developed technical guidelines for digitizing cultural heritage material,[62] as well as a file format comparison matrix.[63]

5.6. Standards for Metadata, Holdings and Identifiers

5.6.1. Metadata and Information Exchange Formats

Metadata for eBooks may be exchanged between publishers, libraries, and preservation institutions in a number of different formats. When academic libraries purchase or license eBooks, they typically need to receive robust MARC[64] records of the books. Preservation institutions must decide whether or not a MARC record is required for their purposes when they preserve eBooks. MARC records are typically based on the AACR2 (Anglo-American Cataloguing Rules) cataloguing standard[65] or its successor, the Resource Description and Access (RDA) cataloguing standard.[66] The quality of MARC records supplied ‘varies greatly among publishers’ (Polanka, 2011, p. 106), which is a factor that preservation institutions must take into account

[60] See http://www.w3.org/Math/
[61] See http://www.digitizationguidelines.gov/about/
[62] See http://www.digitizationguidelines.gov/guidelines/digitize-technical.html
[63] See http://www.digitizationguidelines.gov/guidelines/File_format_compare.html
[64] See http://www.loc.gov/marc/
[65] See http://www.aacr2.org/
[66] See http://www.rda-jsc.org/rda.html

when deciding whether or not to preserve these records and use them as a source of bibliographic metadata for the books.

ONIX for Books is heavily used by publishers when providing preservation institutions with metadata for books and when distributing books to aggregators and third party platforms. ONIX for Books is an XML-based standard. Unlike other metadata standards with a heavy emphasis on bibliographic metadata, ONIX for books includes a significant amount of information useful to all the supply chain partners (for example, the price of the book in different countries). It is known to be a flexible standard that allows a single piece of knowledge to be encoded in multiple ways, which can make it difficult for preservation institutions that normalize the content to a preservation standard.

Bibliographic metadata about individual eBooks can be encoded into the book formats themselves, although anecdotal evidence suggests that the metadata (such as title and author) within formats such as MOBI and EPUB is neither robust nor accurate (Goyal, 2013).The NCBI Book Tag set has a specific set of tags to support robust book metadata. Metadata may also be provided as Dublin Core,[67] typically within a broader XML structure.

5.6.2. Identifiers

Identifying books is particularly difficult for preservation institutions (and libraries). The proliferation of versions (for example, the MOBI version and the EPUB version and the version on the publisher's platform) and the difficulty in identifying specific editions tremendously complicate the identification of a particular work. The ISBN[68]is designed to allow all participants in the book supply chain to track the sales and distribution of specific versions of books, and, per the rules of the ISBN agency, the MOBI, EPUB, and platform versions of a published work with the exact same text, publication date, edition, author and publisher will have different ISBNs (International ISBN Agency, 2010). From a preservation standpoint, it may not be necessary to preserve all the versions, yet the proliferation of ISBNs may make it difficult to identify duplicates or confirm with those supplying content to preservation institutions exactly what has been preserved (as it currently makes it difficult for libraries to provide discovery services for these books).

Many of the new self-publishing tools do not force the use of an ISBN. For example, neither Amazon Kindle Direct Publishing nor Barnes & Noble NookPress require creation of an ISBN for publishing. They will instead assign their internal identifiers, such as an ASIN[69] or BN[70] identifier. Nor is it clear how widely publishers are implementing consistent use of the recommended multiple ISBN policy.

The International Standard Text Code (ISTC)[71] is a recent attempt to provide more robust identification of the intellectual work that is represented as a piece of text (as opposed to the ISBN, which identifies the specific format of an intellectual work). ISTC is an ISO standard, ISO 21047, and was published in 2009. Unlike an ISBN, an ISTC is not tied to a work and publisher pairing; the same ISTC should be used even if the work is published or distributed by two or more different publishers. Thus far, there has not been wide take-up of the ISTC.

The Digital Object Identifier (DOI)[72] is a widely used identifier of scholarly eBooks. Within scholarly publishing, DOIs are often assigned to each chapter of an eBook, in addition to the book as a whole. The DOI Foundation recommends that DOIs created for eBooks incorporate the book ISBN in a format called ISBN-A,[73] an 'actionable ISBN' (International DOI Foundation 2012). This means that the identification issues introduced through multiple ISBNs for the same intellectual book could also become an issue in using DOI for book identification.

[67] See http://dublincore.org/

[68] See http://www.isbn.org/

[69]See http://www.amazon.com/gp/seller/asin-upc-isbn-info.html

[70] See http://cp-barnesandnoble.kb.net/kb/?ArticleId=4354&source=Article&c=12&cid=28#tab:homeTab:crumb:7:artId:4354

[71] See http://www.istc-international.org/html/

[72] See http://www.doi.org/

[73] See http://www.doi.org/factsheets/ISBN-A.html

As it is not yet clear how preservation institutions will want to manage and compare eBook holdings in the future, it is advisable that institutions preserve all identifiers provided to them by the publisher.

6. Case Studies

These are taken from preservation institutions that are currently acquiring and preserving eBook content.

6.1. National Library of the Netherlands

The vision of the National Library of the Netherlands (KB), founded in 1798, is 'to offer everyone everywhere access to everything published in and about the Netherlands.'[74] The KB collects eBooks for two purposes: 1) to provide content for reading to on-site users and 2) to create a deposit collection of content for preservation.

The eBooks collected for use are acquired via patron-driven acquisition (PDA), and are web-accessible. The KB, in compliance with Dutch privacy laws, does not keep any user-specific information about use of its content.

As there is no legal deposit requirement for eBooks in the Netherlands, deposit into the preservation collection is voluntary (as is contribution to the print book collections) and the KB has made arrangements with eBook publisher associations to acquire and preserve publications. Access to these eBooks is restricted to on-site users. The KB 'keeps the bits' of the preservation content, and has a permanent preservation right to the content. The use of DRM in eBooks is explicitly disallowed in the KB's agreement with contributing publishers, and the Library uses format characterization tools to confirm that DRM is absent. They plan to use copy protection themselves for content they deliver to readers on site. For eBooks with embedded content, the KB requires that the publisher warrant that they hold all rights to all the content in the eBook, and can grant them preservation rights. The Library will, after consideration of a publisher request, make an eBook 'dark' and unavailable for access, but does not delete the eBook from the preservation archive.

The KB characterizes most of their eBook content as scholarly. They have a large number of government publications which may also be classified as eBook content, and are also slowly acquiring Dutch mass-market eBooks for preservation. The KB's eBook content is comprised of both born-digital and digitized print books. Current formats in the archive include PDF and Microsoft Word document files, and the KB expects to be receiving content in EPUB 2 and EPUB 3 in the near future; they have not received any full-text XML eBooks. The Library does accept supplemental files in any format that accompanies the eBooks. Because the content to date is provided in PDF and Word, no special eBook reader hardware or software tools beyond standard browser plug-ins are required for rendering eBook content.

The KB uses METS with extended Dublin Core and PREMIS metadata to manage their collection. Metadata is received in ONIX or MARC and normalized to METS. ISBN and DOI identifiers come from the publisher (one identifier per book). The Library also assigns their own unique persistent identifier to each file belonging to a publication.

Concerns expressed by the KB include validating consistency between an eBook and its accompanying descriptive metadata; scaling up workflow processes to handle the expected increasing volume of both print and digital books (up to 50,000 a year); the very plastic nature of eBooks, and the definition of the boundaries of a book or any other publication in electronic form.

6.2. Library of Congress

The Library of Congress (LOC) was founded in 1800 as a reference library for the Congress of the United States. Its mission is to 'support the Congress in fulfilling its constitutional duties and to further the progress

[74] See http://www.kb.nl/en/organization-and-policy/our-mission-and-vision

of knowledge and creativity for the benefit of the American people.'[75] Beginning with the copyright law of 1880 (sponsored by then Librarian of Congress Ainsworth Rand Spofford), all copyright applicants were obliged to submit two copies of their works to the Library.

The Library is only in the beginning stages of adding eBooks to its collections, to be followed by managing the licensing of access to eBooks. They have recently begun to add eBooks to the formal objects that are received through the Cataloguing-in-Publication (CIP) program. CIP has a set of initial publisher participants, to whom LOC has communicated format requirements. Sample files have been received, and the Library is beginning development of its ingest pipeline. Initially at least, titles submitted through CIP will be those published simultaneously as print and eBook.

LOC cannot demand eBooks through Mandatory Legal Deposit, because the current interim regulation solely requires deposit of electronic-only serial publications. They anticipate, however, that there will be agreements in future that will enable some publishers to deposit eBooks instead of print books when the publisher's print version go out of print. LOC anticipates all categories of eBooks (mass-market, trade, scholarly, text, and reference) will be submitted.

Identifiers in eBooks include ISBN and DOI. LOC catalogues books using ILS/MARC,[76] and also maintains Electronic Records Management System (ERMS) records for licensed books.

In addition to eBooks received through CIP, LOC also digitizes volumes from its collections, and receives PDFs of volumes through other collaborative collection building efforts, such as the World Digital Library.[77] Collections include files in PDF, HTML/XHTML, XML/TEI, and EPUB2 formats (LOC expects soon to have EPUB3 as well). LOC has requested DRM-free EPUB files for the CIP eBook program, and so far no publishers have said no, except those that do not produce EPUBs.

LOC has not yet explored the issue of embedded components, and is still in the process of developing the pipeline for validation, characterization, and assessment of eBooks.

The Library has multiple eBook licence agreements in place, but no purchased ownership agreements. LOC provides access through browser only; there are no device downloads.

6.3. HathiTrust

Founded in 2008, the HathiTrust is a partnership of over 90 research institutions and libraries, mostly from the United States, but including institutions in Canada, Spain, and Australia. The HathiTrust Digital Library 'provides long-term preservation and access services for public domain and in copyright content from a variety of sources, including Google, the Internet Archive, Microsoft, and in-house partner institution initiatives.'[78] In addition, the HathiTrust Research Center (HTRC) 'enables computational access for non-profit and educational users to published works in the public domain and, in the future, on limited terms to works in-copyright' in the HathiTrust Digital Library.[79] As of 01 February 2014, HathiTrust collection contains 10,978,052 volumes of which 3,615,823 are in the public domain.[80] There are no legal requirements that would result in deposit in HathiTrust. Works are deposited by member libraries and the scanned images of works are licensed for deposit.

Its collection consists primarily of digitized books and journals, but it does contain some born-digital volumes, and expects to contain more in the future. The collection includes all types of eBooks – mass-market, trade, scholarly, textbooks, and reference – with a preponderance of the latter three types.

[75] http://www.loc.gov/about/mission.html
[76] http://www.loc.gov/marc/
[77] http://www.loc.gov/wdl/
[78] http://www.hathitrust.org/digital_library
[79] http://www.hathitrust.org/htrc
[80] http://www.hathitrust.org/hathitrust_updates

HathiTrust uses MARC bibliographic metadata supplied by depositing partners. Inconsistencies in cataloguing practices of depositing partners, including authority control practices, have posed challenges to HathiTrust's management of partner records. These MARC records describe the original of the digitized image. They are not considered definitive; rather, the definitive catalogue record is that in the institutions' own systems or in WorldCat[81] (Rothman, 2014). Cataloguing records for volumes may include ISBN, ISSN, LCCN, OCLC, or other standard identifiers. HathiTrust identifies items using a unique identifier, frequently derived from the barcode of the originally scanned item. In addition to challenges in managing bibliographic data from depositing institutions, desires to expand the size and scope of HathiTrust's collections have necessitated the development of new tools and processes, and new ways of handling various submission information package (SIP) structures (Beers *et al*., 2010).

Primary preservation formats in HathiTrust are ITU G4 TIFF, JP2, and Unicode; PDF is not a primary preservation format for them. Some coordinate optical character recognition (OCR) information for many digitized volumes is encoded in ALTO[82] XML, DjVu XML, or hOCR[83]. They generate derivative versions —EPUB and PDF — as requested by users from OCR text, and are currently working toward ingest of born-digital journal articles, which will be in NISO Journal Article Tag Suite (JATS)[84] XML. HathiTrust uses JHOVE, along with custom-written programs to validate and assess content. Tools they have made available to assist in the ingest of locally digitized materials are available at http://www.hathitrust.org/ingest_tools.

A significant challenge for HathiTrust has been ensuring appropriate access to the more than 70 per cent of the corpus that is composed of materials that are in copyright, or whose copyright status is unknown. These materials are searchable, but are only available for reading at partner institutions by users who have reading disabilities, consistent with the Chaffee Amendment and Fair Use[85] (York, 2012). The presence of 'inserts' or embedded objects is a consideration in the copyright review that partners conduct on materials in the collection, and may result in a work that is otherwise in the public domain being restricted. Some works are available to requests from the United States only, as HathiTrust only has information about its US copyright provisions; others are available worldwide because they are in the public domain worldwide. The organization also makes works available to read and download where the rights holders have granted permission.

Simultaneous access to works that are made available for uses under Section 108 is restriced to the number of print copies of the work held by the user's institution. Works so restricted are made available for 24-hour digital 'checkout' periods.

Appropriate agreements related to the submission of materials from depositing institutions with respect to preservation of those materials are obtained. HathiTrust intends to provide perpetual access to works in the repository, subject to copyright constraints.

Deletions from the HathiTrust repository are rare, and occur only in instances where

- a volume is either wholly unusable due to quality problems, or a superior copy of the volume is available in HathiTrust (such deletions are authorized by the depositing institution); or
- removal is requested by the rights holder.

Any removal must be approved by the Executive Director. HathiTrust's takedown policy is published at http://www.hathitrust.org/deletion.

[81] http://www.worldcat.org/
[82] http://www.loc.gov/standards/alto/
[83] https://docs.google.com/a/umich.edu/document/d/1QQnIQtvdAC_8n92-LhwPcjtAUFwBlzE8EWnKAxlgVf0/preview
[84] http://jats.nlm.nih.gov/
[85] The limited terms of access under which copyright materials are made available can be found at http://www.hathitrust.org/access_use.

6.4. Portico

Portico is a digital preservation service provided by ITHAKA, a not-for-profit organization based in the United States with a mission to help the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. It was created in 2002 to provide a sustainable digital archive to serve the academic community and its original focus was on scholarly electronic journal content. Since 2008, it has also preserved scholarly eBooks and digitized collections.

Deposit of eBooks into Portico is voluntary on the part of publishers, and is not part of any legal deposit requirement. Portico's licence agreements with participating publishers stipulate that all rights, including preservation rights, to digital objects embedded in deposited eBooks must be cleared by the publisher before deposit with Portico. Portico will only withdraw content under a court order or similar requirement. In such cases, Portico prefers to work with publishers to ‘darken’ the content, but keep it in the archive, rather than removing it entirely from the archive. Licence agreements further stipulate that DRM must not be embedded in deposited content.

Portico does not distinguish digitized from born-digital books in its collection. The collection includes trade eBooks, scholarly monographs, and some books that are used as text books or reference books (though Portico does not target these categories of books for preservation at this time).

The 165,000 books ingested so far into the archive come to Portico principally as ‘header-only’ XML in a variety of XML formats (NLM/BITS, proprietary XML, and ONIX). One publisher sends both the header and the full text of the article, marked up in a proprietary XML vocabulary. Portico preserves all of these with a normalized version of the publisher-provided information in NLM/BITS and Dublin Core. The XML files are provided along with page image files in PDF. Some publishers provide a single PDF file for an entire book; some provide a PDF file for each chapter in a book; some provide both. Portico has not yet received any content in EPUB. Publishers typically provide ISBN and DOI identifiers with each book. If separate PDF files are provided for each chapter, then publishers usually also provide a DOI for each chapter.

Other than for audit purposes by contributing publishers and staff at participating libraries, Portico does not make books available for use except in the case of what are called trigger events (cessation of a publisher’s operations; discontinuation of a title by a publisher; back content no longer offered by a publisher; catastrophic and sustained failure of a publisher’s delivery platform) or a post-cancellation access (PCA) or perpetual access claim by a participating library. (Portico has fulfilled several perpetual access claims for eBooks where the publisher is still making the books available, but the small collection that originally offered the books has been encompassed within a much larger and more expensive collection, a collection which the library did not want to license or purchase.) Normalized descriptive metadata, as well as the PDF files, are then made available via web access; no specialized eBook reader hardware or software is required. As the content is not delivered via proprietary publisher platforms such as Amazon, Portico has not had to deal with any of the privacy issues raised by the use of those platforms.

Portico uses JHOVE and the UNIX file utility to identify, characterize, and validate publisher files. They have found that metadata that comes in ONIX format is particularly challenging to handle; there are many ways to encode the same data, which complicates the development of transformation tools. The requirement to preserve books that come with both a single PDF for the whole book and individual chapter PDFs as well necessitated a refinement of the content model that underpins the archive.

7. Conclusions

Though previously lagging somewhat behind, book publishing is rapidly paralleling the evolution of the print-to-digital transition that has transformed, and continues to transform, journal publishing. After a somewhat delayed beginning, commercial and other pressures have perhaps caused this transformation to accelerate at an even greater rate than that for journals. The large-scale investment in eBook technologies – massive scanning and digitization, refinement of purpose-built eBook reader devices, sophisticated supply chains and search engine technologies – are creating pressures on memory institutions to handle complex new content, with multiple technical and legal complications, at very large scales, well in advance of any stabilization of standards for formats, for workflows, for tools, or for best practices in contracting with producers of this content.

The uses to which the eBook is put, amongst competitors actively seeking to create disruptive and exclusive uses of new technologies, is furthermore rapidly stretching the boundary of what is meant by a book in digital format. They have become complex digital objects, with other potentially complex digital objects embedded within them, which might be dependent for their semantic integrity on extrinsic objects to which they contain hypertext links.

Ownership of the responsibility for the preservation of different large categories of digital artefacts that fall under the rubric of eBooks is not clearly established. Nor are the costs for carrying out the preservation, and establishing sufficient permanent funding to meet those costs.

The right to permanent possession, including perpetual access and preservation rights, is the exception rather than the norm in eBook licensing. The threat to permanent access and preservation is further increased by the widespread use of digital rights management technologies in licensed artefacts. The stability of this leased content, even in libraries, is not assured: content can be modified, and even withdrawn under the terms of use of most eBook licences.

There are sometimes quite serious preservation risks associated with the formats in which eBooks are created. This is particularly the case for proprietary formats, particularly those closely tied to a particular commercial vendor’s hardware platform and distribution system. The variants, deliberate and inadvertent, on even standard eBook format such PDF and EPUB call to mind the ‘browser wars’ of the early days of the Internet. There is also as yet no well-established scheme for uniquely and consistently identifying eBooks. Nor are the tools for validating and characterizing all these formats fully mature.

Large-scale digitization of print books has created valuable and widely used digital surrogates for those books that are being put to uses (rapid search and discovery, text mining across corpora of collections) impossible with print books. However the scale of digitization has introduced quality assurance issues both for the digitized images and for the creation of descriptive metadata for these books. They have also, in some cases, embroiled institutions in legal entanglements arising from both the eBook’s similarity to, and difference from, its print source.

8. Recommended Actions

What steps will help to stabilize the complexities of the eBook environment and help ensure effective sustainable preservation of these complex digital objects?

Publishers

- Ensure explicit comprehensive rights to all objects embedded within a book;
- Negotiate explicit protocols for propagating updates to eBook content. This would include negotiation for removal or darkening of collection items;
- Conform to open standards for eBook formats, metadata, and identifiers. This would include, at a minimum, articulation and documentation of usage for standard formats like ONIX;
- Use standards such as EPUB which include requirements on packaging of the content, as well as the encoding of bibliographic metadata;
- Implement ITSC widely, so that all members of the community can easily identify the intellectual work that is represented by any given eBook;
- Move away from digital rights technologies for restricting access to eBook content; and
- Deliberately engage with memory institutions to ensure the preservation of, and long-term access to, published eBooks.

Libraries and Cultural Heritage Institutions

- Articulate policies for categories of eBook content for whose preservation they will be responsible;
- Co-ordinate with other institutions, both to establish there are no unintentional preservation gaps, and that there is no needless duplication of preservation efforts;
- When acquiring or licensing eBook content, ensure the acquisition includes preservation rights, and prohibits DRM technologies in the preservation copy acquired from the vendor;
- Consider and understand what preservation rights are provided when eBooks are licensed and exactly how long-term access will be ensured by the publisher;
- Articulate preservation policies for the handling of embedded objects, including articulation of legal rights to the content, and workflow requirements to ascertain preservation risks for that embedded content;
- Encourage publishers to participate in preservation institutions to ensure the long-term viability of their eBook content; and
- Invest in maturing existing characterization tools, and extending the toolset. Establish whether there is a preservation requirement somehow to maintain the hardware, or to emulate form factor effects, for those eBook formats closely tied to reader hardware.

9. Glossary

ASCII

American Standard Code for Information Interchange is a character encoding scheme. Developed by the precursor organization of the American National Standards Institute, it encodes 128 characters: the digits 0-9, the letters a–z and A–Z in the Latin alphabet, plus additional punctuation characters and special machine control codes. The term is sometimes used to refer to content in plain (Latin) text.

ASIN

Amazon Standard Identification Numbers (ASINs) are 10-digit unique alphanumeric codes that identify items sold by Amazon and its international affiliates. For books, the ASIN is the same as the ISBN. Books with 13-digit ISBNS are also assigned 10-digit ASINs.

AZW

AZW is an eBook format used exclusively on the Amazon Kindle eBook reader device. It is compatible with Kindle software on personal computers or Apple's iPhones. It is based on the MOBI format. Files in AZW might or might not employ DRM. AZW files are sometimes identified as KF7 (Kindle Format version 7), to distinguish them from the later Amazon eBook format, KF8.

ANGLO-AMERICAN CATALOGUING RULES (AACR2)

The Anglo-American Cataloguing Rules (AACR) is a library cataloguing code published jointly by the American Library Association (ALA), the Canadian Library Association (CLA), and the UK's Chartered Institute of Library and Information Professionals (CILIP). AACR rules cover the description of, and the provision of access points for, all library materials commonly collected at the present time. See also RDA.

CASCADING STYLE SHEETS (CSS)

Cascading Style Sheets (CSS) is a mechanism for adding styles, or display information (such as font characteristics, colours, or spacing) to Web documents. The specifications for CSS are maintained by the CSS Working Group of the World Wide Web Consortium (W3C).

DAISY

DAISY (Digital Accessible Information System) is a technical standard for digital audio books, periodicals and computerized text. It is intended to create an accessible alternative to text and other display materials for those with impaired vision or reading disorders. The format employs MP3 and XML to create a searchable, navigable audio alternative to text**.**

DIGITAL OBJECT IDENTIFIER (DOI)

The digital object identifier (DOI) system provides a technical and social infrastructure for the registration and use of persistent interoperable identifiers for use on digital networks. A DOI is a unique alphanumeric string assigned by a registration agency (the International DOI Foundation) to identify content and provide a persistent link to that objects location on the Internet. A DOI can be assigned to any entity (physical, digital, or abstract) that a user wishes uniquely to identify. It has been standardized as ISO 26324.

Digital Rights Managment (DRM)

Digital Rights Management (DRM) is a set of technologies employed to protect commercial intellectual property in digital content, whether that content is an eBook, a piece of recorded music, a video, or any other digital artefact. It enforces the use of digital licences, which restrict a consumer's access to a digital object in certain ways, including frequency and duration of access to the object, as well as restriction of rights of transfer, or of the ability to copy the object. DRM can also be used to warrant the tamper-resistance of the object. DRM typically entails encryption of the bit stream comprising the object, though it may be confined to implanting a digital watermark in the object to trace users of the object, thus making digital piracy detectable. There are a number of DRM schemes in use, including Adobe Adept DRM, Apple FairPlay, and proprietary DRM employed by Amazon and by the now-defunct Microsoft Reader's proprietary format.

DUBLIN CORE METADATA INITIATIVE

Dublin Core is a metadata specification maintained by the Dublin Core Metadata Initiative (DCMI). It is intended to provide a common vocabulary, or 'core metadata', for simple and generic resource descriptions.

EPUB

EPUB is an interchange format for digital publications. It defines a means of representing, packaging and encoding structured and semantically enhanced web content – including XHTML, CSS, SVG, images, and other resources – for distribution in a single-file format. It is maintained by the International Digital Publishing Forum (IDPF). EPUB3 is the latest version of this format.

FIXED LAYOUT

A fixed layout format is one in which the structure of a page, including the relative position of a page's components, remains fixed across the devices used to display it. Contrast this with reflowable text formats.

FORM FACTOR

Form factor refers to the physical size and shape of a device.

HYPERTEXT MARKUP LANGUAGE (HTML)

HyperText Markup Language is a set of annotations (markup) for creating documents and other information that can be displayed in a web browser. It was developed by Tim Berners-Lee in 1989 to enable document linking and sharing by physicists working at CERN (the European Organization for Nuclear Research). Its syntax and definition comprise an SGML vocabulary. Since 1996, versions of the HTML specification have been maintained by the World Wide Web Consortium (W3C). The latest version of HTML is HTML5. (*See also* XHTML)

ISBN

The International Standard Book Number (ISBN) is a unique commercial book identifier. Each ISBN code uniquely identifies a book. ISBNs assigned before January 1, 2007 are 10 digits long. Those assigned after that date are 13 digits long. Each country has its own agency which assigns ISBNs to publishers located there.

KF8

KF8 is Amazon's proprietary file format, employed by the Kindle Fire eBook reader device, released by Amazon in 2011. It is based on the MOBI format, but it is not entirely backwardly compatible with it. It includes support for some features of HTML5 and CSS3. It is only partially compatible with EPUB 3.

OEBPS

Open eBook Publication Structure (OEBPS) is an XML-based specification for the content, structure, and presentation of electronic books. OEBPS was developed by the Open eBook Forum, a group of organizations involved in electronic publishing and now known as the International Digital Publishing Forum. It has been superseded by EPUB.

ONIX

ONIX for Books Product Information Message is an XML-based standard for rich book metadata for supply chain information to be transmitted between publishers, retailers, and others in the book-selling supply chain. It is maintained by EDItEUR, an international group coordinating development of the standards infrastructure for electronic commerce in the book, eBook and serials sectors.

PORTABLE DOCUMENT FORMAT (PDF)

PDF, originally developed by Adobe Systems, is a format designed to maintain and communicate device-independent page description information. The original page description format has been elaborated over successive versions to enable the embedding of such complex objects as image, audio, and moving image files, hyperlinks, embedded XML metadata, and updatable forms. Specification for various versions and profiles of the format are now maintained by the International Standards Organization.

PATRON-DRIVEN ACQUSITION (PDA)

Patron-driven acquisition is an on-demand model of collection development, in which libraries only purchase new materials when those materials are explicitly requested by patrons, usually after reaching some pre-set threshold number of requests for the object.

RESOURCE DESCRIPTION AND ACCESS (RDA)

RDA provides a set of guidelines and instructions on formulating data to support resource discovery. It provides a comprehensive set of guidelines and instructions covering all types of content and media, and is intended as a successor to AACR2 (Anglo American Cataloguing Rules). It is published by the American Library Association (ALA), the Canadian Library Association (CLA), and the UK's Chartered Institute of Library and Information Professionals (CILIP).

REFLOWABLE TEXT

Reflowable text refers to an electronic document designed to adapt its presentation to the form factor of the device on which it is displayed. It employs a flexible layout scheme that can adapt to different screen and font sizes. Print page boundaries and other display and formatting structures give way to the size boundaries of the device on which a document is read. Contrast this with fixed layout.

STANDARD GENERALIZED MARKUP LANGUAGE (SGML)

SGML (ISO 8879:1896) is a meta-language: a standard for specifying document markup languages. Markup languages based on SGML are declarative descriptions of a document's structure and other attributes. Developed out of IBM's Generalized Markup Language (GML), it informed the development of XML. HTML is an SGML vocabulary.

TEXT ENCODING INITIATIVE (TEI)

The TEI is an international organization founded in 1987 to develop guidelines for encoding machine-readable texts in the humanities and social sciences. The set of guidelines and schemas it has produced are also often referred to as TEI.

Extensible Hyper Text Markup Language (XHTML)

XHTML (Extensible HyperText Markup Language) is a family of XML markup languages designed to mirror HTML, but based on XML rather than SGML, which can therefore be parsed and manipulated by standard XML tools, rather than by HTML-specific parsers such as web browsers. It is maintained by the World Wide Web Consortium (W3C).

Extensible Markup Language(XML)

Extensible Markup Language (XML) is a simple text-based format, derived from SGML, for representing structured information, including documents, data, configuration, books, and transactions. It is maintained by the W3C.

10. Further Reading

Anderson, David, 2013. Preserving Europe's Digital Cultural Heritage: A Legal Perspective. *New Review of Information Networking*, 18: 1, 2013. DOI: 10.1080/13614576.2013.775836. Available at http://www.tandfonline.com/doi/full/10.1080/13614576.2013.775836 [Accessed February 24, 2014].

Arms, Caroline, Chalfant, Don, DeVorsey, Kevin, Dietrich, Chris, Fleischhauer, Carl, Lazorchak, Butch, Morrissey, Sheila and Murray, Kate, 2014. *The Benefits and Risks of the PDF/A-3 File Format for Archival Institutions: An NDSA Report*. Available at http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_PDF_A3_report_final022014.pdf [Accessed February 24, 2014].

Bailey, Timothy P., Scott, Amanda L, Best, Rickey D., 2010. Cost Differentials between E-Books and Print in Academic Libraries. *College & Research Libraries*. Available at http://crl.acrl.org/content/early/2013/10/23/crl13-542.abstract [Accessed February 24, 2014].

Baye, Michael R., De los Santos, Baburand. Wildenbeest, Matthijs R.,2013. Searching for Physical and Digital Media: The Evolution of Platforms for Finding Books. *Economics of Digitization*. National Bureau of Economic Research, Inc. Available at http://www.nber.org/chapters/c12989.pdf [Accessed February 24, 2014].

Beagrie, Neil, 2013. *Preservation, Trust and Continuing Access for e-Journals. DPC Technology Watch Report* 13-04 September 2013. Available at http://www.dpconline.org/advice/technology-watch-reports [Accessed February 24, 2014].

Bläsi, Christoph, and Rothlauf, Franz, 2013. *On the Interoperability of eBook Formats*. European and International Booksellers Federation. Available at http://eibf-booksellers.org/sites/default/files/press_release/2013-05-16/interoperability_ebooks_formats_pdf_13599.pdf [Accessed February 24, 2014].

Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*. Available at http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf [Accessed February 24, 2014].

Castro, Clarice and de Queiroz, Ruy, 2013. The Song of the Sirens. *Information, Communication, and Society,* 16:9, 1441–1455, DOI: 10.1080/1369118X.2012.681678. Available at http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.681678 [Accessed February 24, 2014].

Christenson, Heather, 2011. Hathitrust: A Research Library at Web Scale. *Library Resources & Technical Services,*55:2 (April 2011). Available at http://www.hathitrust.org/documents/christenson-lrts-201104.pdf [Accessed February 24, 2014].

JISC, 2012. Preparing for Effective Adoption and Use of Ebooks in Education. *JISC Observatory Tech Watch Series,* Report No. 4, Final Version, December 2012. Available at http://blog.observatory.jisc.ac.uk/techwatch-reports/ebooks-in-education [Accessed February 24, 2014].

Kirchhoff, Amy, 2011. *Preservation of Digitized Books and Other Digital Content Held by Cultural Heritage Organizations*. A report for the NEH and IMLS resulting from a grant from the Advancing Knowledge: The IMLS/NEH Digital Partnership given to Portico and Cornell University Library. Available at http://www.portico.org/digital-preservation/wp-content/uploads/2010/01/NEH-IMLS-D-book-model.pdf [Accessed February 24, 2014].

Kirchhoff, Amy, 2012. E-book Preservation: Business and Content Challenges. In *No Shelf Required 2: Use and Management of Electronic Books.* Sue Polanka, Ed. ALA Editions 2012.

Mount, Dan, 2014. *E-lending Landscape Report 2014*. Canberra ACT, Australian Library and Information Association, April 2014. Available at https://www.alia.org.au/sites/default/files/publishing/ALIA-Elending-Landscape-Report-2014_0.pdf [Accessed April 29, 2014]

Muir, Laura, Graeme Hawes, 2013. The Case for e-Book Literacy: Undergraduate Students' Experience with e-Books for Course Work, *The Journal of Academic Librarianship*, 39: 3, 260–274, May 2013., DOI: 10.1016/j.acalib.2013.01.002 Available at http://www.sciencedirect.com/science/article/pii/S0099133313000049 [Accessed February 24, 2014].

OCLC, 2012. *The Big Shift: Public Library Strategies for Access to Information in Any Format*. Available at http://www.oclc.org/content/go/en/thebigshift.html?urlm=168727 [Accessed February 24, 2014].

Office of the Register of Copyrights, 2011. *Legal Issues in Mass Digitization: A Preliminary Analysis and Discussion Document*. October 2011. Available at http://www.copyright.gov/docs/massdigitization/USCOMassDigitization_October2011.pdf [Accessed February 24, 2014].

Wischenbart, Ruediger, 2013. *The Global eBook Market: Current Conditions & Future Projections 2013*. O'Reilly Media, Inc. Available at http://shop.oreilly.com/basket.do?nav=%2Fproduct%2Fid%2F114721&from=detail [Accessed February 24, 2014].

Zarins, Uldis, 2013. *Thought Leaders in Latvia: libraries and e-books*. Available at http://www.eifl.net/case-study-libraries-and-ebooks-in-latvia [Accessed February 24, 2014].

Zhang, Laurina, 2013. *Intellectual Property Strategy and the Long Tail: Evidence from the Recorded Music Industry*. Available at http://inside.rotman.utoronto.ca/laurinazhang/files/2013/11/laurina_zhang_jmp_nov4.pdf [Accessed February 24, 2014].

11. References

Amazon, 2012. *Kindle Store Terms of Use*. Available at: http://www.amazon.com/gp/help/customer/display.html/ref=hp_200699130_storeTOU1?nodeId=201014950 [Accessed February 24, 2014].

Baye, Michael R., De los Santos, Babur and Wildenbeest, Matthijs R., 2013. *Searching for Physical and Digital Media: The Evolution of Platforms For Finding Books in Economics of Digitization*, National Bureau of Economic Research, Inc. Available at http://www.nber.org/chapters/c12989.pdf. [Accessed February 24, 2014].

Beers, Shane, York, Jeremy, and Mardesich, Andy, 2010. *Adding New Content Types to a Large-scale Shared Digital Repository*. Presented at iPRES 2010.(September 2010). Available at http://www.hathitrust.org/documents/hathitrust-ipres-201009.pdf [Accessed February 24, 2014].

CDNL, 2010. *British Library: International Survey on Electronic Legal Deposit*. Available at http://www.cdnl.info/Legal_Deposit/CDNL_2010_-_BL_international_survey_on_e-Legal_Deposit.pdf [Accessed February 24, 2014].

Christenson, Heather, 2011. HathiTrust. *Library Resources & Technical Services*, 55:2, 93–102. DOI: 10.5860/lrts.55n2.93. Available at http://alcts.metapress.com/content/Q7720VR01V980266 [Accessed February 24, 2014].

Clarke, M., 2009. Professional and Scholarly Publishing Leads the Market for Ebooks by a Wide Margin. *the scholarly kitchen*. Available at: http://scholarlykitchen.sspnet.org/2009/11/24/professional-and-scholarly-publishing-leads-the-market-for-e-books-by-a-wide-margin/ [Accessed February 24, 2014].

Conway, Paul, 2013. Preserving Imperfection: Assessing the Incidence of Digital Imaging Error in HathiTrust. *Preservation, Digital Technology & Culture*,42: 1, 17–30. DOI: 10.1515/pdtc-2013-0003 Available at http://www.degruyter.com/view/j/pdtc.2013.42.issue-1/pdtc-2013-0003/pdtc-2013-0003.xml?format=INT [Accessed February 24, 2014].

DAISY Planet, The, 2010. *DAISY = Accessibility, Will EPUB 3 = Accessibility? The Daisy Planet, The DAISY Consortium's Monthly Newsletter*. Available at: http://www.daisy.org/planet-2010-11#a2 [Accessed [Accessed February 24, 2014].

Dixler, E., 2008. Browsing Books. *The New York Times Sunday Book Review*. Available at: http://www.nytimes.com/2008/03/16/books/review/PaperRow-t.html?_r=0 [Accessed February 24, 2014].

Fryer, J., 2013. Electronic Lending and Public Libraries: Folding Shelves. *The Economist*. Available at: http://www.economist.com/news/international/21573966-e-books-mean-plot-twist-public-libraries-and-publishers-folding-shelves. [Accessed February 24, 2014].

Gardiner, E. & Musto, R.G., 2010. The Electronic Book. In *The Oxford Companion to the Book,* Michael F. Suarez, SJ and H. R. Woudhuysen, eds. Oxford University Press.

Goldfarb, Charles F., 1996. *The Roots of SGML – A Personal Recollection*. Available at http://www.sgmlsource.com/history/roots.htm [Accessed February 24, 2014].

Goyal, K., 2013. Editing Ebook Metadata. In *Calibre Manual*. Available at: http://manual.calibre-ebook.com/metadata.html [Accessed February 24, 2014].

Hart, Michael, 1992. *The History and Philosophy of Project Gutenberg*. Available at http://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart [Accessed February 24, 2014].

International DOI Foundation, 2012. *DOI® System and the ISBN System: Factsheet*. Available at: http://www.doi.org/factsheets/ISBN-A.html [Accessed February 24, 2014].

International ISBN Agency, 2010. *E-Books and ISBNs: a position paper and action points from the International ISBN Agency*. Available at:

http://www.isbn.org/sites/default/files/images/isbn_agency_e-books_position_paper.pdf [Accessed December 22, 2013].

International Standards Organization (ISO), 1986. *Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML)*, ISO 8879:1986. http://www.iso.org/iso/catalogue_detail.htm?csnumber=16387 [Accessed February 24, 2014].

Kindle Direct Publishing, *Publishing FAQ*. Available at: https://kdp.amazon.com/help?topicId=A36BYK5S7AJ2NQ [Accessed February 24, 2014].

Klems, B., 2013. How can the average writer make money self-publishing e-books? *Writer's Digest*, June 25, 2013. http://www.writersdigest.com/online-editor/how-can-the-average-writer-make-money-self-publishing-e-books [Accessed May 7, 2014]

Liu, Qiong, Safavi-Naini, Reihaneh and Sheppard, Nicholas Paul, 2003. Digital rights management for content distribution. In P*roceedings of the Australasian information security workshop conference on ACSW frontiers 2003 – Volume 21,* 49–58 (ACSW Frontiers '03), Chris Johnson, Paul Montague, and Chris Steketee (Eds.), Australian Computer Society, Inc., Darlinghurst, Australia, Australia. Available at http://dl.acm.org/citation.cfm?id=827994 [Accessed February 24, 2014].

Lynch, Clifford A., 2013. Ebooks in 2013: Promises Broken, Promises Kept, and Faustian Bargains, *Digital Content: What's Next,* supplement to *American Libraries* (May 2013). Available at http://www.cni.org/wp-content/uploads/2013/05/ALA-Ebooks-Paper.pdf. [Accessed February 24, 2014].

Manley, Laura and Holley, Robert P., 2012. History of the Ebook: The Changing Face of Books. *Technical Services Quarterly,* 29:4, 292–311. DOI: 10.1080/07317131.2012.705731. Available at http://www.tandfonline.com/doi/abs/10.1080/07317131.2012.705731#.UvFWCLRO2vA [Accessed February 24, 2014].

mobipocket.com, *Mobipocket Developer Center*. Available at: http://www.mobipocket.com/dev/article.asp?BaseFolder=prcgen&File=building.htm [Accessed February 24, 2014]..

Morrissey, Sheila M., 2012. The Network is the Format: PDF and the Long-term Use of Digital Content, *Archiving 2012*, 200–203. Available at http://www.portico.org/digital-preservation/wp-content/uploads/2012/12/Archiving2012TheNetworkIsTheFormat.pdf [Accessed February 24, 2014].

NOOK Press, *Frequently Asked Questions About NOOK Press*. Available at: https://www.nookpress.com/support/faq [Accessed February 24, 2014].

Nunberg, G., 2009. Google's Book Search: A Disaster for Scholars. *The Chronicle of Higher Education*. Available at: http://chronicle.com/article/Googles-Book-Search-A/48245/[Accessed February 24, 2014].

Polanka, Sue, 2011. *No Shelf Required: E-books in Libraries*. American Library Association, Chicago.

Romano, Frank, 2002. E-Books and the Challenged of Preservation. B*uilding a National Strategy for Preservation: Issues in Digital Media Archiving.* Council on Library and Information Resources and Library of Congress. Available at http://www.clir.org/pubs/reports/pub106/contents.html/ebooks.html [Accessed February 24, 2014].

Rothman, Jon, 2014. *Bibliographic Metadata and HathiTrust*. ALCTS CaMMS Catalog Management Interest Group Meeting,American Library Association MidWinter Convention Philadelphia, Pennsylvania, January 25, 2014. Available at http://www.hathitrust.org/documents/HathiTrust-ALAMidwinter-20140125.pptx[Accessed February 24, 2014].

Schonfeld, R., Housewright, R. & Wulfosn, K., 2013. *ITHAKA S+R US Faculty Survey 2012*, New York, NY: ITHAKA S+R. Available at: http://www.sr.ithaka.org/research-publications/us-faculty-survey-2012 [Accessed February 24, 2014].

St. Clair, Glorianna, 2008. The Million Book Project in Relation to Google. *Journal of Library Administration*, 47:1–2, 151–163. DOI: 10.1080/01930820802111041. Available at http://www.tandfonline.com/doi/abs/10.1080/01930820802111041 [Accessed February 24, 2014].

van der Knijff, Johan, 2012. *EPUB for Archival Preservation*. National Library of the Netherlands. Available at http://www.openplanetsfoundation.org/system/files/epubForArchivalPreservation20072012ExternalDistribution.pdf

van der Knijff, Johan, 2013. *EPUB for Archival Preservation: an update*. Available at http://www.openplanetsfoundation.org/blogs/2013-05-23-epub-archival-preservation-update[Accessed February 24, 2014].

van Dormolen, Hans, 2012. *Metamorfoze Preservation Imaging Guidelines: Image Quality*. Version 1.0, January 2012. Available at http://www.metamorfoze.nl/sites/metamorfoze.nl/files/publicatie_documenten/Metamorfoze_Preservation_Imaging_Guidelines_1.0.pdf [Accessed April 29, 2014].

van Trier, Gerard, 2006. Permanent Access to the Records of Science – The International Role of the e-Depot at the Koninklijke Bibliotheek, National Library of the Netherlands, *LIBER Quarterly*, 16: 3/4. Available at https://liber.library.uu.nl/index.php/lq/article/view/URN%3ANBN%3ANL%3AUI%3A10-1-113470/8062 [Accessed February 24, 2014].

Warnock, John, 1991. *The Camelot Project*. Available at http://www.planetpdf.com/planetpdf/pdfs/warnock_camelot.pdf [Accessed February 24, 2014].

Working Group of the Conference of Directors of National Libraries (CDNL), 1996. *The Legal Deposit of Electronic Publications*. Available at http://www.unesco.org/webworld/memory/legaldep.htm [Accessed February 24, 2014].

York, Jeremy, 2012. HathiTrust: The Elephant in the Library. *Library Issues,* 32: 3 (January 2012). Available at http://www.hathitrust.org/documents/HathiTrust-LI-201201.pdf [Accessed February 24, 2014].