Creating eBooks From Archival Material in a Digital Commons Space

Introduction

Digital Initiatives in the James P. Adams Library at Rhode Island College began creating eBooks from archival collections during the spring semester of 2013. Over the course of several months tools were identified and best practices were determined to assist staff in this process. While the collective content output is valuable, the endeavor gave the office vital experience in developing and editing a content format that should prove increasingly important to the scholarly community in the future. In addition, it provided an opportunity for exposure to new modes of production that should prove useful going forward.

The advent of touchscreen devices for reading text content began in the first decade of the twenty-first century with the introduction of various devices ranging from the iPhone and iPad to the Kindle and Sony Reader. These devices have been followed by a litany of other products ranging from the Nook and Kobo Reader personal reading devices (PRDs) to other multimedia tablets in various form factors (e.g., color screens, processing capabilities, etc.) such as the Asus Nexus 7.

The ePub conversion project was not undertaken solely to present the archives in a new way or as an artifice to highlight collections. The tremendous changes that are occurring in regards to information production and distribution will have a profound impact upon academia. In order to engage with the shifting landscape and provide valuable services to the community in an underserved area, librarians at RIC need to have an understanding of the nature of digital content alongside the devices it will be engaged with upon. A central feature of libraries lies in preserving knowledge and the digital means of preservation present an excellent tactic to use in performing this function. The findings that resulted from the process of converting content to eBooks will supply a foundation that will inform the Digital Initiatives workflow in the digital milieu.

Components of Digital Text

Digital content has three components: content, container, and context. The information contained in a scholarly work is the content. This is not something that should be altered in the act of archiving and preserving material. Context is the metadata, ephemera, audio, or visual material that supplement the content and enhance

user experience. Archivists can provide additional context and material that enriches the material to be preserved. Exciting work is being done in academia using new digital tools to build scholarship around vast data stores. In short, content and context are separate from the focus of this paper. Container represents the means of engagement with content. There are innumerable possible containers, yet presently academic content is consumed most commonly in three ways: print, Internet, and through a touchscreen device. Additional modes of content presentation are necessary in order to accommodate disabilities, such as the audible recording of books for people with visual disabilities. This is important to note this because in creating eBooks, digital commons will be presenting a means to convey content that is more accessible to individuals with disabilities.

Priorities

There is a tremendous opportunity present in working with material to make it fit containers. One of the goals in RIC Digital Initiatives is to make materials as accessible as possible. This entails issuing content in the three forms discussed above and ensuring that all materials are treated in accordance with best practices in preservation. A significant factor in effective preservation entails usability. In undertaking the project to create eBooks, priority was placed on ensuring that work was done with open standards in mind. To ensure long-term preservation and access, material must be adaptable to newly developing technologies. The more closely aligned work in RIC Digital Initiatives is with open digital standards, the more likely it is to be usable in the future with new ways of communicating content.

The issue of fair use is widely considered one that is far too complex and nuanced to be considered in absolute terms. This is the likely reason why legislation surrounding content and copyright in the digital milieu did not seek to address fair use. This point is important to emphasize when approaching the issue. If the rights guaranteed to us in the 1976 Copyright Act, derived from Constitutional clauses, are to be preserved new methods of sharing information that does not rely on technological restrictions will need to be developed so that the right to fair use is preserved. If material is only disseminated with TPM then the anti-circumvention provisions of the DMCA will have the net effect of rendering fair use impossible. By making content available in unrestricted form, Digital Initiatives at RIC is providing a test case for fair use in the digital environment that should hopefully serve to support copy rights.

Text for reading is most commonly delivered digitally as eBooks, web pages, or PDFs. Text should be optimized for delivery in print, on touchscreen devices, and on the Internet. In order to serve all of these areas, Digital Initiatives needed to identify the best ways to deliver content while also ensuring that access would not be limited by financial considerations. This focus on free content distribution has the additional benefit of alleviating many

copyright concerns. Each format has distinct advantages but also presents problems. Increasing access was the primary goal of the project, so knowledge of the various file formats and devices was necessary to determine how to proceed.

The three means of delivering digital text each present unique strengths and weaknesses.

The unique considerations when working with digital content in libraries relate to the Digital Millennium Copyright Act of 1998 (DMCA). This act purposely does not address fair use concerns. It only relates to the circumvention of digital rights management (DRM) and technological protection measures (TPM). Any undertaking that compromises DRM in any way violates the DMCA. Whether or not this circumvention was related to fair use rights is of no consequence to the DMCA. There is no way to justify any violation of DRM under the provisions of the DMCA. Bearing this in mind, Digital Initiatives at RIC had to ensure that any measures were not violated in the conversion process. This meant that any PDFs protected by passwords could not be converted. As far as any other issues with conversion, precedent still needs to be set concerning digital content. Fair use is an essential right that allows libraries to take advantage of copy rights.

PDF

PDFs are written in the PostScript language. The other formats (web and eBooks) are written and based in elements of the browser stack (HTML, CSS, and JavaScript). PDF is written in the PostScript language and this presents a severe disadvantage for the PDF format and may prevent it from remaining relevant on emerging technologies. The language does present an excellent format for print on demand (POD) purposes but lacks the functionality present in eBooks and webpages to account for variation in screen size. Print documents may well be where the future lies for PDF files.

The same properties that make PDF an excellent format for print present distinct problems for use with touchscreen devices. The fundamental reason PDFs are not a suitable e-book format relates to text sizing and the static nature of content in PDFs. The pages cannot adjust to container screen size and require effort on the part of the user to view. PDF is much like an image of a particular size placed in a picture frame that has no necessary relation in size to

that of the original image. In order for a section of the picture to be viewed it has to be manipulated within the frame to be viewed. PDF files require resizing and zooming depending on varying device characteristics. This is a problem that can be dealt with when reading shorter texts, yet it can be averted in longer works by placing texts in file format such as MOBI or EPUB, as will be detailed in the following paragraphs.

Portable document format (PDF) is the most common form of digital scholarly content presently. An open format developed by the Adobe Corporation in the early 1990s, PDF is an excellent format for print-on-demand (POD) documents. It is ubiquitous in academia, as scholarly journal articles are typically made available through databases and digital repositories in this form. It represents the standard when making digital text available to print. The simple nature of the file type and the reliability of a good print product make the barrier to access extremely low. For many learners, in particular adult scholars who are not as familiar with digital content as younger learners, it represents an excellent format and should remain relevant in academia for the foreseeable future. The same characteristics that make PDF an excellent format for POD are those that make it suboptimal for touchscreen devices. PDF is written using the PostScript language, an old language mainly used for printing. It is a static language and does not have a large user base.

eBooks

When addressing the issue of creating eBooks, a high priority was placed on creating content in the open ePub standard rather than the .azw or .ibooks format. If .azw or .ibooks could then be created from the ePub file that would be a bonus. The ePub format is open and reliant upon the markup language HTML. eBooks are generally delivered in either an ePub or MOBI file format. The Apple iBooks format is simply an ePub with proprietary tags and coding that limit its functionality to devices operating on iOS.

The MOBI format is the open version of the Amazon Kindle format (.azw). It was initially developed by the French entity MOBIpocket the early 2000s. Amazon purchased the format in 2005. Amazon has added some formatting and proprietary tags for DRM purposes, yet fundamentally the file type remains the same. Development of the MOBI format has ceased but Amazon Kindle devices are still able to properly render MOBI files. While remaining distinct formats, MOBI and ePub are similar enough to be viewed as the pertinent formats librarians need to be cognizant of when engaging with the eBook format.

The ePub file format is the primary open file format for electronic books. Nearly all e-readers and apps accept this format, regardless of the source of acquisition. If a device fails to read a particular ePub it is most likely due to

internal qualities of the file. The file may contain some form of DRM or other attribute that makes it unreadable by specific programs. In addition, specific programs may be incapable of reading specific files due to some TPM.

The creation of a standardized format was first begun in 1999 by the Open e-book Authoring Group. In 2007 the group, now known as the International Digital Publishing Forum (IDPF) released a version named ePub 2. It is designed so that text and images can be resized depending on container characteristics to create a more fluid reading experience for the user. This has particular value when reading on devices with very small screens, such as smartphones. This is presently the format used by most works, but an ePub3 standard has just been developed and introduced that should allow the format to remain very useful with the new and developing tablet technologies and digital textbooks.

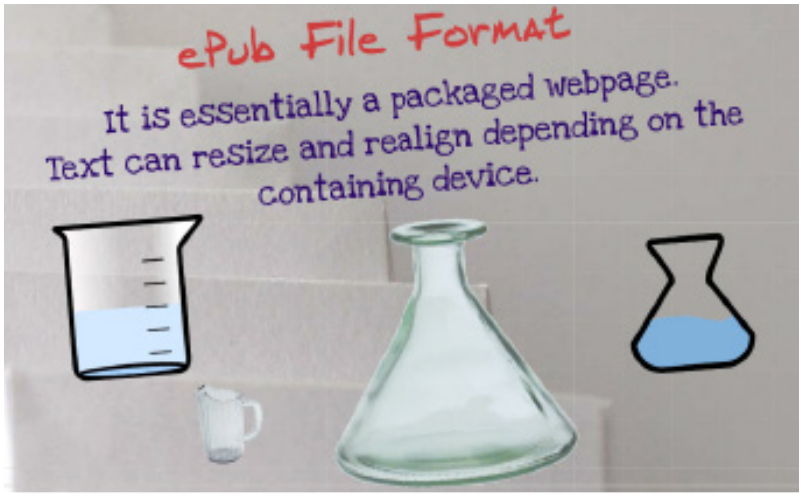
The IDPF is "the global trade and standards organization dedicated to the development of electronic publishing and content consumption" ("About Us," n.d.). It develops and maintains the ePub format. IDPF members include publishers in all formats, booksellers, authors, and software developers involved in e-content. Among the hundreds of members are disparate entities like Apple, Aptara, the Open University, McGraw-Hill, Google, and the University of Michigan Library.

The ePub format, at its core, "was always intended to be a general-purpose document format" (Garrish, 2011, p. 12). Fortunately, a cursory knowledge of HTML, XML, or other web-standards will provide a familiarity with the basis of ePub. Working in the format and creating a file does not require the knowledge of another set of protocols or languages specific to the format. ePub is built on the browser stack, a combination of HTML, CSS, and JavaScript. This is a tremendous advantage in several ways. For developers, authors, and publishers these formats do not require any particular knowledge of computer languages apart from a basic familiarity with the structure of the World Wide Web. Tying the format to these languages ensures their currency and relevance in the technology world. In addition, it allows them to develop along with the dominant form of information dissemination present in the world. From a practical standpoint, the browser stack is optimized for legibility and functionality across screen size. It has evolved over time to be accessible on a vast proliferation of devices. It is the ideal means for delivering content to e-reading devices. When discussing webpages, this is something of a trinity that combines to create a whole, functioning entity. HTML is a means for delivering content. This content will then be formatted and its appearance will be determined by CSS. Through the course of use and interaction, a webpages behavior will be manifested by the JavaScript code it is linked to. While ePub has been tied to web standards since its inception, much of the updating present in ePub3 pertains to behavior.

The ePub 2 format is comprised of three main parts. The three standards, in the form of files and folders, are delivered in an Open Container Format (OCF) in the form of a zip file archive. All files related to DRM and encryption, of primary concern to librarians when providing access to materials, are present in the first folder. The other folder, the Open e-book Structure (OEBPS) contains all the content of the book. The IDPF announced the formation of a working group to create an ePub 3 standard. The decision to update the standard was mainly driven by three considerations:

1. A desire to make the ePub a global format by ensuring that all written languages rendered properly (specific problems related to Arabic and many Asian languages);
2. The market for e-book devices changed considerably since the introduction of ePub 2 with the advent of smart-phones and tablets;
3. The format did not make allowances for content integration, multimedia, and linking capabilities.

In addition, publishers wanted to use the format for other media than literary books, notably periodicals, newspapers, and text books. In the creation of the ePub3 standard care was taken to incorporate the existing technologies incorporated into ePub and add to them with the use of widely used web-based utilities. The step to incorporate new improvements and functionality is very prescient on the part of the IDPF. In order to maintain its viability as a file format in the fast evolving digital environment, the IDPF wisely made sure that ePub was able to adapt and utilize new abilities. Learning from the lessons provided by such obsolete technology products as WordPerfect, MySpace, and Netscape Navigator, the IDPF has worked to ensure that the ePub format will remain vibrant.



eBooks are fundamentally a collection of web pages packaged in a zip file and delivered to a device for consumption. The ePub file format was initially developed by the International Digital Publishing Forum and has gone through three major revisions since inception. Versions 1.0 and 2.0 essentially focused on text delivery and were composed of XML and XHTML files. The newest version, ePub3, is a major step forward and was released in 2011. ePub3 brings the full incorporation of the browser stack (HTML, CSS, and JavaScript) to the eBook

realm. Interactive text, multimedia content, game based learning, and several features allowing for more accessibility regardless of disabilities present. Much of this potential was displayed when Apple introduced iBooks 2.0 in January of 2012. The launch of fully interactive textbooks with dazzling features was greeted with considerable fanfare. It presents educators with a hint of the possibilities present in digital course material. Due to the limited adoption of the ePub3 format at this time, librarians can develop the skills to create eBooks in ePub2 formats that can eventually be enhanced by the functionality allowed for in ePub3. As of this writing, no reading application fully supports the standard and development remains very dynamic. Therefore, when Digital Initiatives began assessing the creation of eBooks the ability to devise content that possessed valid structure according to the ePub2 standards was emphasized. Because the ePub3 standards are so new and adoption is extremely limited, this format is only marginally applicable to present practice yet presents future potentialities content creators should be cognizant of. The revolutionary characteristic pertaining to digital content transmission present in eBooks lies in a structure designed so that text and images can be resized depending on container characteristics to create a more fluid reading experience for the user.

eBooks represent the only way to deliver content customized for varying screen sizes not necessitating Internet access presently. This will change as the developments in HTML5 proliferate, but pragmatically speaking, eBooks are the format to be concerned with currently when creating content for touchscreen devices. The analogy of a pitcher of water best applies to eBooks and touchscreen devices. The water represents the content. Touchscreen devices can be considered pitchers of varying shapes and sizes. The content is enabled to be reformatted by the underlying organizational structure of the browser stack and poured into any container, regardless of physical properties. This allows for convenience when dealing with content and touchscreen devices. However, it should be noted that the eBook file format is relatively young and is not standardized to a high degree like HTML across platforms. While there has been a good deal of consistency when rendering books across ereading applications, there was not absolute uniformity so at times Digital Initiatives had to make judgments when producing content. It was decided that legibility and a faithful reproduction of print characteristics should be the primary considerations when editorial decisions needed to be made.

Web

This brings us to the third prominent way that content is delivered digitally, webpages. Like eBooks, webpages are built using HTML, CSS, and JavaScript. Webpages have the additional benefit of representing a mature medium. Content is conveyed in the same way across all browsers at this point due to the adoption of standards. These standards

have developed and authority has been recognized as the Internet has matured. The W3C working group now writes and develops standardized language in conjunction with the computer industry. This makes the Internet an extraordinarily reliable place to publish content with the desired appearance. In addition, Internet access is available for free at nearly all libraries.

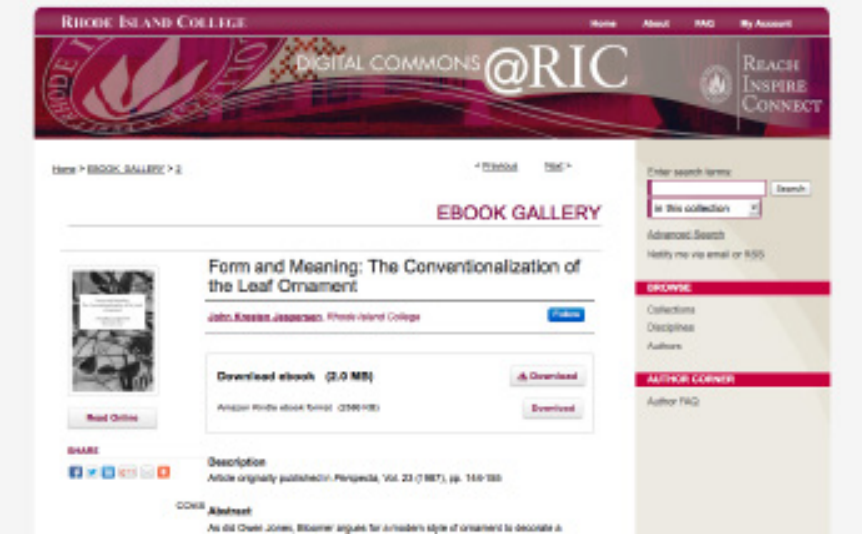
The accessibility of webpages at present rests in an Internet connection. This limits the portability and use in all locations. It should be noted that many of the developments present in HTML5 will eliminate this difficulty. HTML5 allows for a storage cache of up to 1GB. This is plenty of cached storage to account for nearly all eBooks. It is important to remain cognizant of this characteristic of HTML5 while at the same time recognizing that widespread adoption is not imminent. While this is a development that is coming, it is uncertain when exactly or how prevalent it will be in the short term. Bearing this in mind, providing content in ways that are accessible to users now requires the RIC Digital Initiatives office to provide digital content in other forms as well. Creating content as eBooks for the near term has the potential to address the eventualities present in HTML5. When the qualities of HTML5 have proliferated enough that they become the standard means for eBook distribution, eBooks created as in either the ePub or MOBI file format can be repurposed with minimal effort. The file remains the same and the packaging characteristics are simply removed. Perhaps a basic line of markup needs to be added at the beginning of the file so that it can be properly rendered.

Each form of digital content has benefits and drawbacks. If the three defining barriers to access of digital content are related to finance, technological aptitude, and the presence (or lack thereof) of Internet access then a strategy for providing content in PDF, eBook, and webpage form would mitigate theses obstacles. Keeping this in mind, the optimal way for RIC Digital Initiatives to move forward in the conversion process would rest in a solution that would allow for the production of content in all of these forms.

Priorities

In evaluating tools for use Digital Initiatives determined that the priorities associated with eBook conversion were, in order of importance: ease of use; cost-effectiveness; cross-platform applicability; and quality of output. The requirement that any strategy adopted require limited skill was a function of staffing. The Digital Initiatives office relies upon student workers, part-time adjunct faculty, part-time professional staff, and graduate students from URI's GSLIS engaged in professional field experience (PFE) to accomplish the task of collecting and archiving scholarly works from RIC faculty and students. A tool for use should require minimal training so that as much time as possible can be spent

engaged in the conversion process rather than in training. Additionally, rather than relying on expensive software that required licenses or support fees, a solution that necessitated minimal financial investment was sought. Preferably, a program that could work on Apple, Microsoft, and Linux operating systems would remove cost barriers and also lower the level of training necessary to engage in the process. Ideally, the strategy to process eBooks should allow Digital Initiatives to build a workflow for all incoming materials around it saving financial resources in the form of staffing required and in simplifying the multitude of file formats that Digital Initiatives would like to make content available in.



PressBooks

PressBooks is an open source plugin developed for the WordPress content management system (CMS). The founder, Hugh McGuire, edited an excellent collection of essays entitled *Book: A Futurist's Manifesto* that was published in 2012. It compiles writing from publishers, librarians, technologists, and scholars in the present environment. The potentialities of digital text are examined throughout the work. The book served as a test case for PressBooks as a publishing platform. The print version and eBook were published and distributed by O'Reilly Media. It was released in three digital parts and feedback was solicited through web and digital publishing. The entire work is available for free at http://book.pressbooks.com and readers are encouraged to leave comments and ask for further clarification on sections. While the work was in the formative stages users and contributors were able to engage with and discuss it. This is a model that has the potential to revolutionize the way written text is consumed in academia. Faculty could issue text that could be better used by students with a tool like PressBooks. It could be more contemporaneous and also tailored to student needs when examining subjects. Further clarification on topics could be provided according to student input.

PressBooks was identified as the best tool to use for Digital Initiatives to use for the project for several reasons, the primary being ease of use. The process of creating eBooks with PressBooks is extremely simple. The interface is exactly the same as a WordPress blog and content is entered in a simple WYSIWYG editor. A basic familiarity with text documents and the simple cutting and pasting computer functions are the only skills that are required to create

an eBook. This was of major benefit to Digital Initiatives because the work of creating eBooks was to be completed by a graduate student in a professional field experience internship, a librarian serving as adjunct faculty in the Digital Initiatives and Reference departments, and a part-time professional staff person. In addition, this cohort was ensuring that normal department activities continued while the head librarian was absent due to medical leave. In order to undertake this project successfully, the process needed to require limited time spent training in new tasks.

The training necessary to begin the eBook conversion process was minimal and only required guided experience and thorough documentation. The interface of PressBooks allows for multiple editors. A staffer in Digital Initiatives at RIC could be given a brief introduction to PressBooks and then be allowed to enter a text. Upon completion, the resulting eBook could then be assessed and corrected (if necessary) before being released. The final product could then be downloaded as a PDF, ePub, and MOBI. PressBooks provides the additional benefit of allowing for eBook styling consistent with user desires. Digital Initiatives at RIC was able to create a stylesheet and copy and paste less than two pages of text into a field. With this CSS inserted, all Digital Initiatives Press at RIC publication was given a uniform look. Once these files were available Digital Initiatives at RIC worked closely with Sarah Rodlund at bepress to create a book gallery for the files. The page displaying individual eBooks was customized so that file formats would be apparent and a link to the web version of the book was present.

PressBooks presented several other benefits that were discovered over the course of the project. PressBooks operates in a browser so there are no issues hindering functionality relating to platform or hardware being used. In addition, presently storage is free in a PressBooks cloud, so there are no limitations relating to workstation location or database connection. Full use only requires an Internet connection.

Adobe InDesign

Workflow in the print publishing industry is typically structured around Adobe InDesign. For publishers this software has styling and typesetting functionality that make it well suited to print. Use of InDesign requires a degree of familiarity and skills that necessitate training lasting for several hours or days, depending on the task. The newest version of InDesign, contained within Adobe Creative Suites 6 (CS6), incorporates features that enable users to produce eBooks and enhance them with the multimedia attributes present in ePub3. This product costs hundreds of dollars for licenses per workstation and is the final CS version that will be released as a software package. Future editions will only be available by subscribing to the Adobe Creative Cloud. While CS6 functions on both the Windows and Mac operating systems, use is not uniform. Proficiency in each system requires a familiarity with different features of the

particular operating system edition. The working files for InDesign are a proprietary format that can only be edited using an Adobe program. For these reasons, InDesign was not a practical solution to be integrated into a standard Digital Initiatives workflow. InDesign does have excellent stylizing properties and proves valuable when working with content that contains tables or multiple images. This proves valuable for the conversion process of works in the fields of art and natural sciences. At RIC in Digital Initiatives most of the archival content is text-based and can be formatted with PressBooks yet in the unusual event where more complex formatting is required a staff member with the necessary skills can prepare the content using InDesign. Beautiful material can be produced, yet for much of the work at Digital Initiatives using InDesign is excessive. Using it for most work in Digital Initiatives is not unlike using a Ferrari to go the grocery store. It will get you there but the skills required to drive it present a barrier to entry and the performance difference is negligible for such a mundane task.

Conclusion

The use of PressBooks at RIC Digital Initiatives in converting digital archives into eBooks provided us with the opportunity to assess and learn a new software that can have profound implications for the academic community at RIC. The process was undertaken with the hope to develop a workflow that would enable Digital Initiatives at RIC to become a publishing hub for the school, supporting faculty and students in preparing and publishing works to support scholarship and enrich education.

In the process of engaging in the process, Digital Initiatives at RIC acquired first-hand knowledge of the various file formats for digital content presently used. Using PressBooks as a tool for publishing allows Digital Initiatives at RIC to produce high-quality content in multiple formats while presenting software that is easy to learn, inexpensive to operate, and compatible with any computer possessing an updated web browser and internet connection.

References

Garrish, M. (2011). *What is EPUB 3?* O'Reilly Media. Retrieved from http://shop.oreilly.com/product/0636920022442.do

International Digital Publishing Forum. (nd). About Us | International Digital Publishing Forum. *International Digital Publishing Forum*. Retrieved from http://idpf.org/about-us