Practical Technology for Archives

Volume 1 | Issue 3 Article 4

11-1-2014

Doc to PDF and HTML

Eric Willey
Illinois State University

Doc to PDF and HTML

Description
[Excerpt] Max J. Evans notes the paradox the digital age presents to archivists: the explosion of information and budget cuts means increasing backlogs and less time to gain detailed subject knowledge of collections, while users believe that all information is "quickly and easily available" if not already digitized and on the web. For some institutions the lack of digitization extends not only to collections, but also to access points such as finding aids. In a 2004 survey of seventeen institutional repositories Christina J. Hostetter found that "in most cases, archives have approximately 10 percent or less of their descriptions to holdings online." In a 2010 paper Christopher J. Prom found that among surveyed institutions "the 'average' institution makes descriptive information at any level of completeness available on the Internet for a paltry 50% of its processed collections and 15% of its unprocessed collections." While these statistics include information regarding processed collections not available in any form (online or off), Prom notes many respondents to his survey identified a strong need for "better tools to do their descriptive work" including a "streamlined process for creating finding aids in an open source format that can be viewed on the web." Prom concludes, in part, that "it is currently beyond the capacity of many institutions to implement MARC and EAD in a cost-effective fashion" and more economical means of providing online access points are needed. The current article provides one means of batch creating HTML or PDF documents from existing word processing documents. The method described has a relatively low barrier to entry, and is particularly targeted at smaller institutions which might face challenges in creating online access points due to lack of funding and specialized training.

Keywords
digital archives, file types, conversion, word processing

Doc to PDF and HTML

Eric Willey

Illinois State University

Introduction

Max J. Evans notes the paradox the digital age presents to archivists: the explosion of information and budget cuts means increasing backlogs and less time to gain detailed subject knowledge of collections, while users believe that all information is "quickly and easily available" if not already digitized and on the web.[i] For some institutions the lack of digitization extends not only to collections, but also to access points such as finding aids. In a 2004 survey of seventeen institutional repositories Christina J. Hostetter found that "in most cases, archives have approximately 10 percent or less of their descriptions to holdings online."[ii] In a 2010 paper Christopher J. Prom found that among surveyed institutions "the 'average' institution makes descriptive information at any level of completeness available on the Internet for a paltry 50% of its processed collections and 15% of its unprocessed collections."[iii] While these statistics include information regarding processed collections not available in any form (online or off), Prom notes many respondents to his survey identified a strong need for "better tools to do their descriptive work" including a "streamlined process for creating finding aids in an open source format that can be viewed on the web."[iv] Prom concludes, in part, that "it is currently beyond the capacity of many institutions to implement MARC and EAD in a cost-effective fashion" and more economical means of providing online access points are needed.[v] The current article provides one means of batch creating HTML or PDF documents from existing word processing documents. The method described has a relatively low barrier to entry, and is particularly targeted at smaller

institutions which might face challenges in creating online access points due to lack of funding and specialized training.

While this paper specifically focuses on converting finding aids in Microsoft Word .doc format, the overall principles would likely generalize to any word processing document without unusual formatting requirements or images. Briefly, finding aids are uploaded to Google Drive and converted to Google Doc format, and then downloaded into either a PDF file with full text search capabilities, or an HTML file with an internal CSS stylesheet for formatting. Free software is then used to edit the HTML finding aids and finalize their appearance before uploading to the web. It should be stressed that this process has only been tested on finding aids with text and minimal formatting. No attempt has been made by the author to apply it to documents with tables, charts, graphs, or images. Archivists implementing this technology should understand HTML and CSS sufficiently that they feel comfortable editing existing code, and be comfortable uploading and downloading large numbers of documents.

Preparation

As a first step in this or any project involving the transformation of a large number of files it is vital to create separate working and backup versions of the files by copying them to separate folders or even separate computers. In the event that undesirable and irrevocable changes are made to the files this will ensure that an uncorrupted copy of the original information remains. For an additional layer of protection archivists may find it useful to work on the copied files at an entirely separate computer from where the originals are stored. If the original files cannot be accessed from the work station where transformation is occurring, there is little danger of them being accidentally altered. In any case, under no circumstances should archivists work on the only existing copy of their files.

Once working files are created, archivists will need to determine if they will convert the files to HTML or PDF format. When choosing which format to convert files

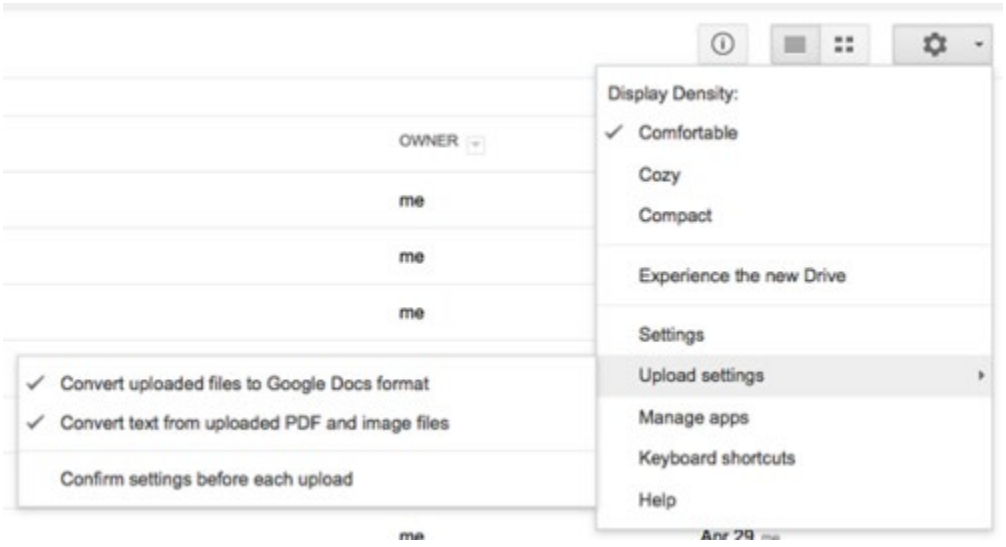
to, archivists should consider the needs of their users and their own technology needs and capabilities. PDF is a simpler conversion process because files can likely be converted without extensive reformatting, and the resulting files will be easier for users to print and download; however, PDF files will be larger in size, less likely to display well on the web (especially on mobile devices or devices with small screens), and will take longer for users to download than HTML files. Conversely, HTML files will be smaller, display on the web better (provided they are formatted to do so before conversion and edited), and will likely display better on mobile devices, but will also likely require more reformatting of the original documents and editing; therefore, they will require a greater time investment than PDF files. A third option for archivists wishing to invest the time is to convert the files to both formats, using PDF for patrons wishing to print or save the document, and HTML for web display. Which of these options is best for an individual archivist or institution will depend largely on the resources available and goals of the project. It may also be advisable to run a small test batch of approximately ten percent of the files to be converted in both PDF and HTML format to get a general idea of how much time will be involved in the conversion process, and what the end results will look like before making final decisions regarding file conversion.

When working copies of the files have been created and a decision between PDF and HTML format has been made, archivists should make any formatting changes desired to enhance web display using their word processing program. If possible changes should be made prior to conversion, as it will likely be easier in most cases to make changes to a word processing document than a PDF or HTML file. For PDF files, no further changes will be possible once the file format has been changed without repeating the entire process; however, as PDF is intended to be a print format it is likely only minimal changes to the original document will be necessary. For documents converted to HTML, which is a web display format, archivists may wish to consider eliminating large amounts of white space, justifying to the left margin rather than centering text, deleting indents, and adding URLs and email addresses to the contact information in the file. Editing files converted to HTML post-conversion is possible, but

for most archivists it will likely be easier to change the original word processing document prior to conversion.

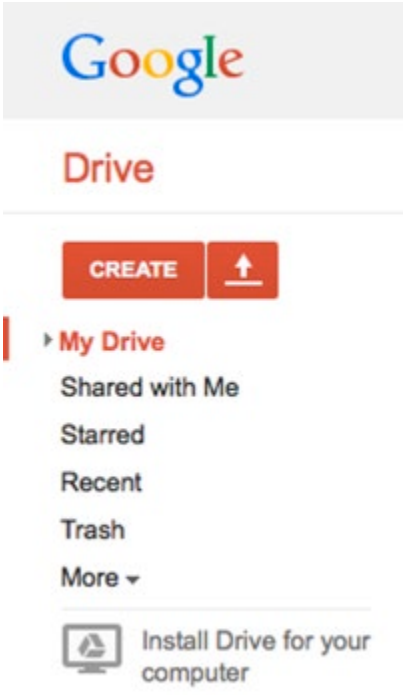
Conversion to PDF Format

When the files are formatted as they will appear in the PDF document, open or create a Google Drive account. This account will be used to upload the files and then to convert and download them in the new PDF format. An account can be created by visiting https://drive.google.com/ and using either an existing Google account and password, or by following the prompts to create a new account. Once the Google Drive account has been logged into, create a folder for the documents being worked on, and click on the "Settings" icon in the upper right corner (it is not labelled, but resembles a gear). Then select "Upload Settings" from the drop-down menu. Select "Convert uploaded files to Google Docs format." This step is critical, as files must be converted to Google Docs format when they are uploaded in order to be converted to PDF when downloaded.



Click on "Settings," the gear shaped icon in the upper right corner, select "Upload Settings," and check "Convert uploaded files to Google Docs format."

When the appropriate settings are selected, click on the Upload icon (which appears as an orange arrow pointing upwards on the left side of the screen), follow the menu to select the word processing files or folder you wish to upload, and upload the relevant files. This process is very similar to moving files to another directory.



After formatting files and selecting options, click the Upload icon which appears as an arrow pointing upwards in the top left hand corner to upload files.

Due to the selection of the "Convert uploaded files to Google Docs format" setting in the previous step the files will automatically be converted to the Google Drive format allowing them to be downloaded as PDF files. When all files have been uploaded (which may take some time depending on the number of files, their size, and the speed of the archivist's internet connection), right-click on the folder containing your finding aids, select download, choose "PDF" from the drop-down menu of file formats, and finally choose the new folder or destination on your computer where you would like to download the converted files. Google Drive will zip the files, move them, and notify you when the download is complete. Click on the zip file and the new PDF files will extract

into the folder, and conversion is complete. Checking at least a few of the files to make sure they are formatted correctly is highly recommended; however, if there are formatting or content issues it will likely be easier to edit the original working file in a word processing program and convert it again than it will be to modify the PDF file itself.

HTML Conversion

The process for HTML conversion is nearly identical to that for PDF conversion. Archivists should create a Google Drive account and set their preferences by clicking on the "Settings" icon in the upper right corner, create a folder, and then upload their files by clicking on the Upload icon (the upward pointing arrow in the upper left corner) and selecting the files. After they are uploaded the files are then downloaded to a local folder by selecting the files, right clicking, and choosing the HTML format. The zipped files are downloaded to the computer, where they can be unzipped by double clicking the file, and then viewed in a web browser (Firefox, Chrome, Safari, Explorer, etc.).

If the format of the files is acceptable, the conversion process is complete. If changes are desired, archivists have two choices: They can either make changes in the word processing documents and run the conversion process again, or edit the HTML and internal CSS stylesheet in a text editor such as Notepad++. This decision will likely depend on how comfortable an archivist is with editing HTML and CSS, and the scope of the changes to be made. It should be noted that each of these files have their own internal CSS stylesheet. Any changes made in one HTML file will not appear in other files, as they would if these were HTML pages operating through a common external stylesheet.

Unfortunately, there may be occasional errors in the formatting of your finding aids. Non-printing characters, mistakes, and general errors in the conversion process can all result in documents that look less than ideal. For archivists who are completely unfamiliar with HTML and CSS and wish to learn there are several online resources that offer excellent and often free tutorials. During editing, archivists might notice that some

undesirable code appears repeatedly. A good portion of this code was likely generated because the HTML and CSS were trying to mimic the layout of the word processor page (which is designed to look like a printed page) in a web page (which isn't necessarily designed to look like a printed page). Archivists can open multiple files in Notepad++ and do search and replace across all of them, which can quickly eliminate this recurring code. Simply open Notepad++, select "Open," and hold down shift while using the arrow keys to select multiple files. The "Replace" function then has an option to "Replace All in All Opened Documents." Some code that may come up in multiple converted files includes:

max-width:432pt: This sets the maximum width of the web page so that it appears to be on 8.5×11 inch paper. This can result in text that stops partway across a users screen, leaving a large amount of undesirable white space on the right hand side of the document. Replacing this with max-width:none; should allow the file to fill the entire broswer window.

font-family:"Georgia": and **font-family:"Calibri";** Font families may change across or even within documents. Archivist may wish to search for and replace font-families with one choice (for example, **font-family:"Times New Roman";**) for the sake of consistency.

line-height:1.1500000000000001: Line height default settings in the word processing program can lead to spacing issues in the HTML document. Again for the sake of consistency archivists may wish to scan HTML documents and replace common line heights with line-height:1;.

After Conversion

After conversion and editing, documents will need to be placed on the web. While that is beyond the scope of this paper archivists may wish to take note of ArchiveGrid, the OCLC Research discovery service which will freely harvest and make searchable finding aids in HTML and PDF (as well as EAD) format. This can offer an

additional access point for no charge, and no additional effort on the part of the archivist (all that is required at this time is for archivists to supply a link for the finding aids, and a contact link for their institution). Even institutions which do not contribute MARC records to OCLC can join ArchiveGrid. In practical terms this means that an institution would not need to create catalog records of any sort to make their finding aids searchable through ArchiveGrid. ArchiveGrid can be contacted at: http://beta.worldcat.org/archivegrid/collections/

Limitations and Conclusion

It should be stressed that conversion of finding aids to HTML and PDF is not necessarily considered best practice at many institutions. A fully encoded EAD finding aid would offer greater opportunities for data exchange with other institutions and search capability for patrons; however, the process to convert documents to EAD is much more involved than what is described here. Some knowledge of XML would be required, and even when the finding aids were converted EAD is not a display language and further work with the file (likely through XSLT transformation) would be required to create a document which could be displayed on the web. This process is intended as a quick and dirty way to make a large number of documents available, not as a recommendation for best practice or even long term strategy.

This process would likely work for any collection of word processing documents an institution might care to place online, including transcriptions, calendars of letters, or collection guides. It is also certainly conceivable that collections in a born digital word processing format could be converted to HTML and placed online using this format. It should also again be stressed that this process may behave unpredictably for documents containing tables, spreadsheets, or images (particularly for HTML documents, and likely less so for PDF documents). While documents and finding aids which are simply lengthy should not not pose special challenges or problems, moving beyond simple text with minimal formatting may require more tweaking of the initial or converted file. Finally, please bear in mind that Google is continually changing services, and options and

capabilities may change over time. Currently, the Google Drive Help Center is located at https://support.google.com/drive/ with the majority of topics related to this article under the "Google Drive on the web" tab. Despite these limitations, it is hoped that archivists and librarians will find this process useful in making their finding aids and research documents available for their patrons.

About the Author

Eric Willey is a Special Formats Cataloger at Milner Library, Illinois State University in Normal, Illinois. He has previously worked as an Associate Curator of Special Collections at the Filson Historical Society in Louisville, Kentucky, an intern at the Illinois Regional Archives Depository (IRAD) site at Western Illinois University, and a project assistant with the McCormick-International Harvester Collection at the Wisconsin Historical Society.

Bibliography

Evans, Max J. “Archives of the People, by the People, for the People.” The American Archivist 70 (Fall/Winter 2007): 387-400.

Hostetter, Christina J. “Online Finding Aids: Are They Practical?” Journal of Archival Organization 2(1/2) (2004): 117-145.

Prom, Christopher J. “Optimum Access? Processing in College and University Archives.” The American Archivist 73 (Spring/Summer 2010): 146-174.

Walch, Vicki. “Where History Begins: A Report on Historical Records Repositories in the United States.” Council of State Historical Records Coordinators, 1998, http://www.statearchivists.org/reports/HRRS/HRRSALL.PDF [accessed March 12, 2011].

Notes:

[i] Max J. Evans, “Archives of the People, by the People, for the People,” The American Archivist 70 (Fall/Winter 2007): 388.

[ii] Christina J. Hostetter, “Online Finding Aids: Are They Practical?,” Journal of Archival Organization 2(1/2) (2004): 123.

[iii] Christopher J. Prom, “Optimum Access? Processing in College and University Archives,” The American Archivist 73 (Spring/Summer 2010): 162.

[iv] Prom, 165.

[v] Prom, 168.