IFLA WLIC 2016 COLUMBUS

Submitted on: 22.11.2017

2016 Satellite meeting - *News, new roles & preservation advocacy: moving libraries into action*
10 – 11 August 2016
Lexington, Kentucky USA, USA

Exploratory Analysis of Born-Digital Newspaper Content

Mark E. Phillips
Digital Libraries, University of North Texas, Denton, Texas USA
E-mail address: mark.phillips@unt.edu

Ana Krahmer
Digital Libraries, University of North Texas, Denton, Texas USA
E-mail address: ana.krahmer@unt.edu

Abstract:

The Texas Digital Newspaper Program, operated by the University of North Texas Libraries, actively works to digitally preserve news in the form of print and born digital newspaper content via The Portal to Texas History. For two years, TDNP has partnered with the Texas Press Association to preserve born-digital newspaper titles from its member institutions. These PDF-based print masters total more than 3 million pages from over 500 titles across the state and allow UNT Libraries to explore significant metrics associated with born-digital newspaper content at a scale that previously had been impossible. This paper reports on exploratory investigations by the TDNP to understand aggregate patterns in the generation of born-digital news editions by analyzing technical metadata extracted from the 3 million pages currently in the preservation collection. While this research is still in its early stages, the goal is to provide an overview of current publishing practices of the more than 500 newspaper publishers across Texas. Furthermore, this research can enhance libraries' understanding about current publishing trends as they plan digital preservation policies and practices in support of publisher preservation needs.

Keywords: newspapers, PDFs, metadata, preservation, press associations

1 WHAT CAN WE LEARN FROM A STATEWIDE COLLECTION OF NEWSPAPER PDFS?

The Texas Digital Newspaper Program (TDNP), operated by the University of North Texas Libraries, actively works to digitally preserve news in the form of print and born digital newspaper content via The Portal to Texas History. For two years, TDNP has partnered with the Texas Press Association to preserve born-digital newspaper titles from its member institutions. These PDF-based print masters total more than 3 million pages from over 500 titles across the state and allow UNT Libraries to explore significant metrics associated with born-digital newspaper content at a scale that previously had been impossible. This paper reports on exploratory investigations by the TDNP to understand aggregate patterns in the generation of born-digital news editions by analyzing technical metadata extracted from the 3 million pages currently in the preservation collection. While this research is still in its early stages, the goal is to provide an overview of current publishing practices of the more than 500 newspaper publishers across Texas. Furthermore, this research can enhance libraries' understanding about current publishing trends as they plan digital preservation policies and practices in support of publisher preservation needs.

At UNT Libraries (UNT), the Texas Digital Newspaper Program has partnered with the Texas Press Association to gather and digitally preserve the PDFs submitted to NewzGroup, an electronic clipping service by the association's publisher members. To plan for the preservation of this wide range of PDF content from publishers it is important to understand the characteristics of the PDF files we are interested in preserving. As part of this initiative, UNT has to plan for long-term collection sustainability to uphold our preservation commitment, and we wanted to gather as much data as possible from the raw PDF set to address sustainability needs.

1.1 Review of Literature

A review of the literature about newspaper PDF preservation is difficult to write when the primary audience for the paper is the same group who has been investigating and writing about PDF preservation since the late 2000s. Patrick Fleming of the British Library (2011) wrote, "All national and regional daily newspapers and an ever growing number of weekly newspapers are produced digitally with an output in PDF format. Within five years, the entire UK and Irish newspaper publishing industry will be produced using digital technology. This, together with the growth in colour presses, has enabled newspapers to increase the number of pages and products that they offer" (Fleming, p. 22). Fleming's tacit point is that the cultural value of the newspapers is rising even while the cost to produce them electronically is dropping, and the wealth of PDF content available in 2016 is such that we can now tackle tangible preservation questions about PDF files themselves.

UNT Libraries first started working with PDF newspaper editions in 2011, when a publisher contacted them to help her preserve an entire run of newspapers (Krahmer & Phillips, 2013). Groups such as the Center for Research Libraries and the National Digital Stewardship Alliance's (NDSA) Content Working Group have looked closely at newspaper PDF preservation as a valuable means to preserve current news content. According to the NDSA Content Working Group, actionable steps to encourage publishers to participate in preservation efforts should include educating them "on use of web archiving technologies" through online tutorials with the goal of "encourag[ing] them to archive and preserve news

content" (NDSA, p. 3, 2013). As electronic legal deposit has become more widely adopted in Europe, libraries have begun to tackle a variety of questions, such as those of Par Nilsson of the National Library of Sweden (2014): with what news content to preserve, when in production the content should be preserved, and what are the implications of quality problems in the electronically deposited content (Nillson, 2014, p. 4-5). It is our hope that data from the Texas PDF set will help to address questions that libraries are beginning to ask about file-level preservation within the wider dataset.

1.2 Research Approach

The data we describe here represents fifteen years of newspapers, from 2002 until 2016 encompassing 513 titles, and 2,974,271 PDF files. For this presentation and paper, we have accumulated the data and have just begun analysis. This research project is still in its nascent stages in regard to applying the data to any specific purpose.

At the outset of the project, we identified three research questions:

R1. What PDF versions are present in this set of newspaper content?

R2. What are the primary authoring tools used to create these PDF?

R3. What is the most common page size of newspapers in this dataset?

The goal of answering these research questions was to help us at UNT with preservation and access planning at UNT Libraries' TDNP for the PDF newspaper pages in this collection. R1 and R2 support long-term preservation decisions while R3 helps us calculate future storage requirements for the newspaper content when delivered by The Portal to Texas History. All of these questions support UNT Libraries' commitment to digital preservation and digital-preservation planning, particularly in terms of helping us revise Digital Libraries' preservation and policies in future iterations.

2 DATA GATHERED

For this study, we used `pdfinfo`, a common command line tool for extracting metadata from PDF files, and using this, we processed each file create a metadata output file displaying metadata embedded in the PDF. This extracted metadata comprises the dataset used in this project. After the metadata extraction was complete, we separated valid from invalid PDF files. Valid PDF files, of course, allowed for successful metadata extraction. From the total of 2,974,271 PDFs in the dataset, 2,692 were invalid PDF files, and we thus removed these files for a different investigation process to determine why they were corrupt and did not create valid metadata files during the conversion process. The remaining 2,971,579 PDF metadata instances are what we evaluate for the remainder of this paper.

While there are fifteen years of PDF files present in the dataset, the majority (96%) of the files were created from 2012 to 2016. Figure 1 shows the distribution of PDF files for each year. Note that this research was conducted mid-year in 2016 so the PDF counts for that year are lower than previous complete years.

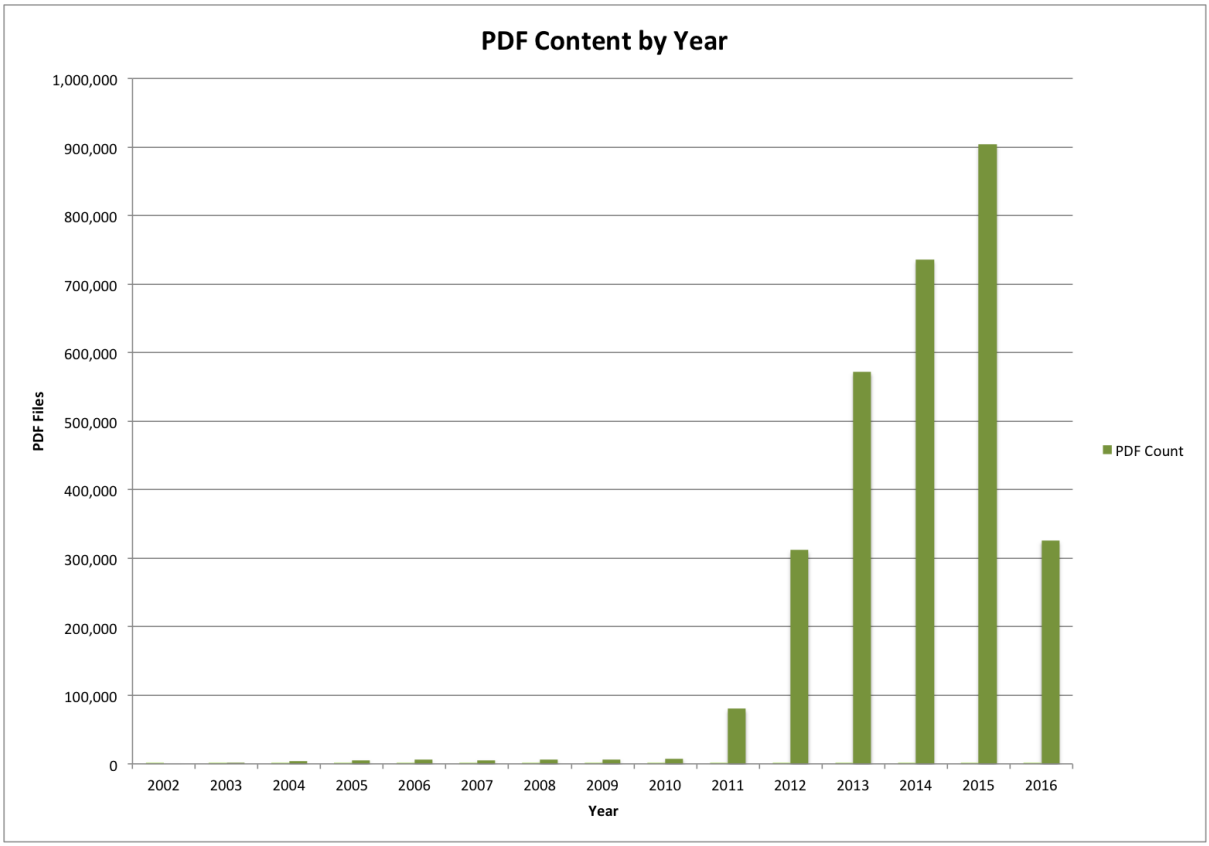

Figure 1: Distribution of PDF files for each year.

Table 1 represents all extracted and derived metadata from individual PDF files. The total number of fields extracted is a great deal larger than what we explored for the purposes of our study because we deliberately narrowed the field analysis to those fields we need for our purposes.

Table 1: Extracted and derived metadata for a single PDF file.

Field	**Example Value**
Valid Meta File	True
Optimized	True
Producer	Adobe PDF Library 9.9
Orientation	portrait
Meta File Path	./90121/2016-03-21/90121_NAXXX-0321-A-A@03.pdf.meta
Creation Date	2016-03-20T21:37:18Z
Page Height Rounded	23.25
Creator	Adobe InDesign CS5.5 (7.5.3)

File Size	3124156
Page Width Rounded	12.75
Page Width Inches	12.7
Page Dim Rounded	12.75x23.25
Page Count	1
Title ID	90121
Tagged	False
Encrypted	False
Version	1.4
Page Height Inches	23.3
Page Rotation	0
ID	90121/2016-03-21/90121_NAXXX-0321-A-A@03.pdf.meta

2.1.1 PDF Versions

170 (33%) of the titles only have one PDF version, 140 (27%) have two PDF versions and 98 titles (19%) have as many as three PDF versions per title. The remaining 105 titles (21%) have four or more PDF versions in their respective PDFs. PDF versions are often passively set by authoring software or in some situations are dependent upon what original publishers needed to do with their PDF pages, the version can have implications on preservation planning because support for certain types of functionality changes with different versions. The versions speak to functionality and features within the PDF.

2.1.2 Producer

"Producer" is the engine that converted or created the PDF file, and is sometimes a third-party tool. The Producer information can be used to help explain file-level problems if they occur. 258 distinct values are present in the Producer field of the PDF dataset. 2,942,165 (98.9%) of the PDFs in the dataset have data present in the Producer field. Of these 258 values, the top ten values represent the values present in 78% of the total PDFs. Table 2 represents the top ten "Producer" entries derived from the dataset.

Table 2: "Producer" field, representing the engine that converted or created the PDF file.

Producer	**PDF Count**	**% of Total**
iTextSharp 4.0.2 (based on iText 2.0.1)	899,689	30.2%
5,0,0,298 *	440,100	14.8%

Adobe PDF Library 9.9	273,339	9.2%
Adobe PDF Library 8.0	140,586	4.7%
OneVision PDFengine (Windows 64bit Build 24.072.S)	124,361	4.2%
Jaws PDF Library 3.63.3591	112,222	3.8%
Adobe PDF Library 9.0	94,159	3.2%
Adobe PDF Library 10.0.1	85,214	2.9%
OneVision PDFengine (Windows 64bit Build 25.080.S)	80,782	2.7%
OneVision PDFengine (Windows 64bit Build 23.224.S)	36,351	1.2%

*the value 5,0,0,298 is the I.R.I.S software package.

2.1.3 Creator

PDF "Creator," generally refers to the software version used to generate the PDF. This information can be used to answer questions relating to processing and file interoperability errors. Distinct from the Producer field, there are 207 distinct values for the Creator field in the PDF dataset. 1,962,563 (66%) of the PDFs in the dataset have a value set for the Creator field. The top Creator field values are listed in Table 3.

Table 3: PDF "Creator," representing the software version used to generate the PDF. The "Creator" field is not to be confused with the "Producer" field.

Creator	PDF Count	% of Total
I.R.I.S.	474,665	16.0%
Adobe InDesign CS5.5 (7.5.3)	135,829	4.6%
Newsway	128,078	4.3%
Adobe InDesign CS4 (6.0.6)	117,972	4.0%
Adobe InDesign CS6 (Macintosh)	84,435	2.8%
Adobe InDesign CS3 (5.0.4)	76,965	2.6%
CCI Europe	76,707	2.6%
Adobe InDesign CS5 (7.0)	72,175	2.4%
Fred 3.0	67,839	2.3%
Adobe InDesign Server CS3 (5.0.5)	59,776	2.0%

2.1.4 Dimensions

We took the width and height for each of the PDF files and used it to calculate the most common page sizes present in the dataset. To reduce the number of unique values into something more manageable, we rounded each value to the nearest quarter of an inch. This resulting number is stored in the dataset in the field "Page Dim Rounded" and is stored in the format of *widthxheight* (11.0x22.0). This is consistent with how we display sizes in the physical description field for online access. We can use the width and height of a PDF file for quality control and calculating storage requirements of rasterized derivatives (JPEG or TIF). Present in the PDF dataset are a total of 820 unique size values. 2,971,579 (99%) of the PDFs in the dataset have a value set for this generated Rounded Dimensions field. Table 4 represents the ten most common page sizes, which account for 44% of the total pages sizes in the dataset.

Table 4: Top ten page dimension occurrences in the dataset.

Rounded Dimensions	PDF Count	% of Total
11.0x22.0	246,568	8.29%
11.0x22.25	188,279	6.33%
12.0x22.0	153,629	5.17%
11.0x22.5	141,599	4.76%
10.5x21.5	137,946	4.64%
11.0x21.5	119,230	4.01%
11.5x22.0	103,021	3.46%
11.0x11.0	91,705	3.08%
13.5x23.5	66,450	2.23%
10.0x21.0	64,327	2.16%

3 CONCLUSION

The data we have generated here is descriptive of the PDF newspaper set, and we will use it for a few purposes. At UNT, we archive the PDF as the master file because this is exactly what was contributed from the publisher. As such, we can use this data to help us with long-term preservation planning. In addition, the information from the metadata helps us understand and troubleshoot file-level processing errors when we create derivative files from these PDFs.

These PDFs represent the print masters of newspapers, that are sent to the printing office in preparation for delivery to people's doorsteps. Since these are born-digital, but are also intended to be printed into an analog medium, the costs to digitally preserve are reduced—no scanning is required for these newspaper issues. We have little to no ability to return to the original producers of the PDF content, and as a result, this type of information

gives us the best possible in-depth comprehension of these files. We have only begun to consider how this metadata will support preservation planning, and we encourage other institutions working in newspaper PDF preservation to utilize this data for their own institutional planning purposes. The many different applications and file characteristics that are involved in creating newspaper PDFs reveal the work we as preservation technologists much put into planning and responding to file needs on an institutional level.

Acknowledgments

We would like to thank the Texas Press Association and NewzGroup for their involvement in newspaper preservation. We would also like to thank the News Media Section of the International Federation of Library Associations for their continuing support and interest in news preservation worldwide.

References

Fleming, P. (2011). "The British Library newspaper strategy." In Hartmut Walravens (ed.), *Newspapers: Legal Deposite and Research in the Digital Era.* International Federation of Library Associations Publications 150. The Hague, The Netherlands: Walter de Gruyter.

Krahmer, A. & Phillips, M. (2014). "Texas newspaper PDF preservation: A Low-cost solution with tremendous value." International Federation of Library Associations Conference Proceedings. Lyon, France: News Media Section.

National Digital Stewardship Alliance. (2012.) Case Study: Newspaper e-prints. Retrieved from http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_CaseStudy_NewspaperEPrints.pdf

Nilsson, Par. (2014). "Collecting bits and pieces." International Federation of Library Associations Conference Proceedings. Lyon, France: News Media Section.