Library of Congress and News Media Alliance (NMA) Technical Pilot project

Report on ePrints

3/8/2017

Contents

Overview

About the Project

In the last decade, newspapers increasingly have turned to online publishing, either in addition to or as a substitute for print publishing. The Library of Congress, in its quest to sustain and preserve a universal historical record, is considering how to incorporate these online products into its newspaper collecting and archiving responsibilities. Historically, much of the Library's newspaper collection has been developed through transfers of physical materials from publishers to the Copyright Office in fulfillment of registration requirements. The purpose of this pilot was to explore the technical feasibility of accepting electronic copies of newspapers in lieu of physical deposits. Copyright regulatory changes were deliberately excluded from consideration, with the expectation that those issues could be pursued if technical concerns were satisfied. The hope of all pilot participants was that it would be possible to take advantage of digital technologies already in use by publishers, reducing the burden on publishers currently submitting print or microfilm copies for Copyright registration purposes, increasing the opportunity for timely and complete deposits, and increasing functionality of the collection through improved searchability.

The News Media Alliance (NMA), known as the Newspaper Association of America (NAA) until 2016, is a non-profit representing more than 2,000 U.S. newspapers. In its August 31, 2009 response to the Copyright Office's Notice of Proposed Rulemaking concerning mandatory deposit of online-only content, NAA expressed interest in working with the Library to "help identify means of providing mandatory and registration deposit copies to the Copyright Office and Library of Congress…" In 2012, NAA and the Library of Congress Office of Strategic Initiatives executed a pilot project to explore deposit scenarios for both ePrint and newspaper website content. In connection with ePrints, NAA members provided feedback on the Library's proposed PDF (Portable Document Format) specifications and sent sample PDF files. The Library project team recorded observations over the course of several months, and concluded its ePrint investigation with a report issued in October 2012.

The news industry and the Library, with particular regard to electronic publishing, have matured in the five years since the initial pilot, while the microfilm industry has waned. In 2016, at the request of NAA, the Library started a new pilot project to re-examine PDFs and their metadata and test technical delivery methods. For the purpose of this new pilot, the Library re-assessed its collection requirements and publisher capabilities for generating and transferring digital newspaper content to the Library.

As the unit of the Library responsible for collection management and access, Library Services focused on exploring what news print publishers are currently producing for online distribution, what PDF characteristics would be practical for publishers to implement in support of efficient and sustainable collection management and access, and the technical mechanisms for establishing regular transfers of digital data.

This report focuses on one type of electronic news content, the "ePrint" or "edition." An ePrint is an electronic facsimile of the newspaper print edition, typically in PDF format. It is often provided on a newspaper's website for use with e-readers or to encourage online browsing. NMA solicited its membership to find participants interested and willing to provide feedback to the Library on ePrints as a possible future deposit copy format.

About the Team

Library Services assembled a team of technical and content experts from the Serial & Government Publications Division (SER) and the Technology Policy Directorate to work with Danielle Coffey, the Vice President of Public Policy at NMA. The seven-member Library team included: Teresa Sierra, Deborah Thomas, Robin Butterhof, and Nathan Yarasavage from SER; Beth Dulabahn, Kate Murray, and Melissa Hire from Technology Policy. The goal of this group was to explore various types of PDFs and metadata generated by newspaper publishers, test a range of technical methods whereby newspaper publishers might deliver content electronically to the Library, and inform future decisions about recommended formats for newspapers and group registration scenarios.

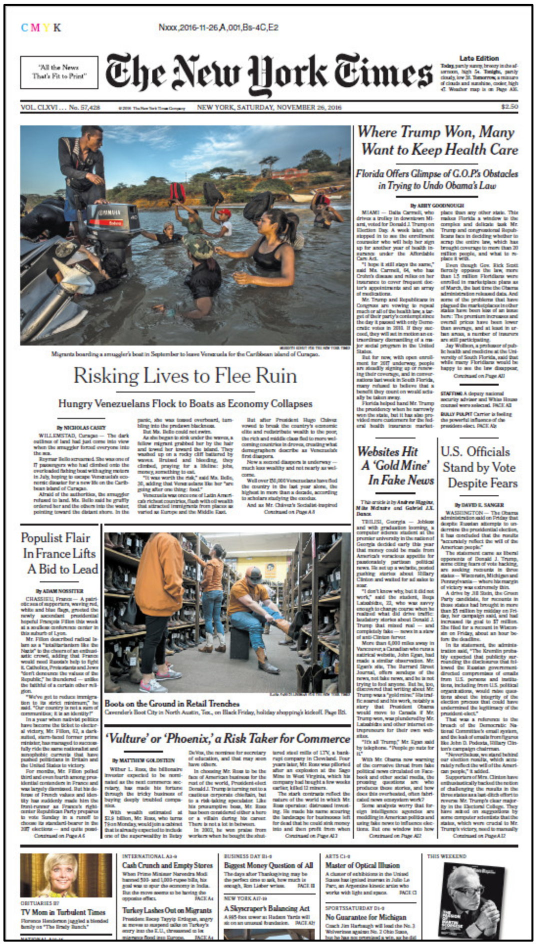
Purpose

The purpose of this report is to share conclusions from the 2016-2017 investigation on ePrints, and to provide background and context for the conclusions.

About ePrints

The ePrint is an electronic facsimile of the newspaper print edition, usually in PDF format, and is typically created for online subscribers of the newspaper and/or commercial printing purposes. ePrint versions are readily available for tablets and smart phones from popular newspapers such as the Wall Street Journal, New York Times, and USA Today. As pictured on the right, the ePrint contains text, graphics, and images. NMA reports that ePrints are a tracked category in circulation figures captured by the Audit Bureau of Circulation, the industry-trusted standard for audited circulation information.

ePrints are not merely for online browsing; they serve other purposes as well. They may be sent to a third-party vendor for printing or used as source files for the creation of microfilm mandatory deposit and registration copies. Newspapers may create ePrints themselves or use a third party vendor to create, host, and deliver the ePrint. Examples of third party vendors include Tecnavia, Libre Digital, Olive, and Newz Group. In addition, PDF has become a common format of Web-based electronic document exchange. A wide variety of both desktop and production-level software applications can both produce and display PDF files, making it a useful format for long-term management and archiving.



C M Y K Nxxx,2016-11-26,A,001,Bs-4C,E2

"All the News That's Fit to Print"

The New York Times

Late Edition

Today, partly sunny, breezy in the afternoon, high 54. Tonight, partly cloudy, low 38. Tomorrow, a mixture of clouds and sunshine, cooler, high 47. Weather map is on Page A16.

VOL. CLXVI ... No. 57,428 © 2016 The New York Times Company NEW YORK, SATURDAY, NOVEMBER 26, 2016 $2.50

MERIDITH KOHUT FOR THE NEW YORK TIMES

Migrants boarding a smuggler's boat in September to leave Venezuela for the Caribbean island of Curaçao.

Risking Lives to Flee Ruin

Hungry Venezuelans Flock to Boats as Economy Collapses

By NICHOLAS CASEY

WILLEMSTAD, Curaçao — The dark outlines of land had just come into view when the smuggler forced everyone into the sea.

Roymar Bello screamed. She was one of 17 passengers who had climbed onto the overloaded fishing boat with aging motors in July, hoping to escape Venezuela's economic disaster for a new life on the Caribbean island of Curaçao.

Afraid of the authorities, the smuggler refused to land. Ms. Bello said he gruffly ordered her and the others into the water, pointing toward the distant shore. In the panic, she was tossed overboard, tumbling into the predawn blackness.

But Ms. Bello could not swim.

As she began to sink under the waves, a fellow migrant grabbed her by the hair and towed her toward the island. They washed up on a rocky cliff battered by waves. Bruised and bleeding, they climbed, praying for a lifeline: jobs, money, something to eat.

"It was worth the risk," said Ms. Bello, 30, adding that Venezuelans like her "are going after one thing: food."

Venezuela was once one of Latin America's richest countries, flush with oil wealth that attracted immigrants from places as varied as Europe and the Middle East.

But after President Hugo Chávez vowed to break the country's economic elite and redistribute wealth to the poor, the rich and middle class fled to more welcoming countries in droves, creating what demographers describe as Venezuela's first diaspora.

Now a second diaspora is underway — much less wealthy and not nearly as welcome.

Well over 150,000 Venezuelans have fled the country in the last year alone, the highest in more than a decade, according to scholars studying the exodus.

And as Mr. Chávez's Socialist-inspired

Continued on Page A8

Where Trump Won, Many Want to Keep Health Care

Florida Offers Glimpse of G.O.P.'s Obstacles in Trying to Undo Obama's Law

By ABBY GOODNOUGH

MIAMI — Dalia Carmell, who drives a trolley in downtown Miami, voted for Donald J. Trump on Election Day. A week later, she stopped in to see the enrollment counselor who will help her sign up for another year of health insurance under the Affordable Care Act.

"I hope it still stays the same," said Ms. Carmell, 64, who has Crohn's disease and relies on her insurance to cover frequent doctor's appointments and an array of medications.

Mr. Trump and Republicans in Congress are vowing to repeal much or all of the health law, a target of their party's contempt since the day it passed with only Democratic votes in 2010. If they succeed, they will set in motion an extraordinary dismantling of a major social program in the United States.

But for now, with open enrollment for 2017 underway, people are steadily signing up or renewing their coverage, and in conversations last week in South Florida, many refused to believe that a benefit they count on would actually be taken away.

Florida helped hand Mr. Trump the presidency when he narrowly won the state, but it has also provided more customers for the federal health insurance marketplace than any other state. This makes Florida a window to the complex and delicate task Mr. Trump and congressional Republicans face in deciding whether to scrap the entire law, which has brought coverage to more than 20 million people, and what to replace it with.

Even though Gov. Rick Scott fiercely opposes the law, more than 1.5 million Floridians were enrolled in marketplace plans as of March, the last time the Obama administration released data. And some of the problems that have plagued the marketplaces in other states have been less of an issue here: The premium increases and overall prices have been lower than average, and at least in urban areas, a number of insurers are still participating.

Jay Wolfson, a professor of public health and medicine at the University of South Florida, said that while many Floridians would be happy to see the law disappear,

Continued on Page A11

STAFFING A deputy national security adviser and White House counsel were selected. PAGE A11

BULLY PULPIT Carrier is feeling the powerful influence of the president-elect. PAGE A15

Populist Flair In France Lifts A Bid to Lead

By ADAM NOSSITER

CHASSIEU, France — A patriotic sea of supporters, waving red, white and blue flags, greeted the newly ascendant presidential hopeful François Fillon this week at a soulless conference center in this suburb of Lyon.

Mr. Fillon described radical Islam as a "totalitarianism like the Nazis" to the cheers of an enthusiastic crowd, adding that France would need Russia's help to fight it. Catholics, Protestants and Jews "don't denounce the values of the Republic," he thundered — unlike the faithful of a certain other religion.

"We've got to reduce immigration to its strict minimum," he said. "Our country is not a sum of communities; it is an identity!"

In a year when nativist politics have become the ticket to electoral victory, Mr. Fillon, 62, a dark-suited, stern-faced former prime minister, has managed to successfully ride the same nationalist and xenophobic currents that have pushed politicians in Britain and the United States to victory.

For months, Mr. Fillon polled third and even fourth among presidential contenders in France and was largely dismissed. But his defense of French values and identity has suddenly made him the front-runner as France's right-center Republican Party prepares to vote Sunday in a runoff to choose its standard-bearer in the 2017 elections — and quite possi-

Continued on Page A6

[illegible]

Boots on the Ground in Retail Trenches

Cavender's Boot City in North Austin, Tex., on Black Friday, holiday shopping's kickoff. Page B5.

'Vulture' or 'Phoenix,' a Risk Taker for Commerce

By MATTHEW GOLDSTEIN

Wilbur L. Ross, the billionaire investor expected to be nominated as the next commerce secretary, has made his fortune through the tricky business of buying deeply troubled companies.

With wealth estimated at $2.9 billion, Mr. Ross, who turns 79 on Monday, would join a cabinet that is already expected to include one of the superwealthy in Betsy DeVos, the nominee for secretary of education, and that may soon have others.

In choosing Mr. Ross to be the face of American business for the rest of the world, President-elect Donald J. Trump is turning not to a cautious corporate chieftain, but to a risk-taking speculator. Like his presumptive boss, Mr. Ross has been considered either a hero or a villain during his career. There is not a lot in between.

In 2002, he won praise from workers when he bought the shuttered steel mills of LTV, a bankrupt company in Cleveland. Four years later, Mr. Ross was pilloried after an explosion at the Sago Mine in West Virginia, which his company had bought a few weeks earlier, killed 12 miners.

The stark contrasts reflect the nature of the world in which Mr. Ross operates: distressed investing. He made his name scouring the landscape for businesses left for dead that he could sink money into and then profit from when

Continued on Page A13

Websites Hit A 'Gold Mine' In Fake News

This article is by Andrew Higgins, Mike McIntire and Gabriel J.X. Dance.

TBILISI, Georgia — Jobless and with graduation looming, a computer science student at the premier university in the nation of Georgia decided early this year that money could be made from America's voracious appetite for passionately partisan political news. He set up a website, posted gushing stories about Hillary Clinton and waited for ad sales to soar.

"I don't know why, but it did not work," said the student, Beqa Latsabidze, 22, who was savvy enough to change course when he realized what did drive traffic: laudatory stories about Donald J. Trump that mixed real — and completely fake — news in a stew of anti-Clinton fervor.

More than 6,000 miles away in Vancouver, a Canadian who runs a satirical website, John Egan, had made a similar observation. Mr. Egan's site, The Burrard Street Journal, offers sendups of the news, not fake news, and he is not trying to fool anyone. But he, too, discovered that writing about Mr. Trump was a "gold mine." His traffic soared and his work, notably a story that President Obama would move to Canada if Mr. Trump won, was plundered by Mr. Latsabidze and other internet entrepreneurs for their own websites.

"It's all Trump," Mr. Egan said by telephone. "People go nuts for it."

With Mr. Obama now warning of the corrosive threat from fake political news circulated on Facebook and other social media, the pressing questions are: who produces these stories, and how does this overheated, often fabricated news ecosystem work?

Some analysts worry that foreign intelligence agencies are meddling in American politics and using fake news to influence elections. But one window into how

Continued on Page A12

U.S. Officials Stand by Vote Despite Fears

By DAVID E. SANGER

WASHINGTON — The Obama administration said on Friday that despite Russian attempts to undermine the presidential election, it has concluded that the results "accurately reflect the will of the American people."

The statement came as liberal opponents of Donald J. Trump, some citing fears of vote hacking, are seeking recounts in three states — Wisconsin, Michigan and Pennsylvania — where his margin of victory was extremely thin.

A drive by Jill Stein, the Green Party candidate, for recounts in those states had brought in more than $5 million by midday on Friday, her campaign said, and had increased its goal to $7 million. She filed for a recount in Wisconsin on Friday, about an hour before the deadline.

In its statement, the administration said, "The Kremlin probably expected that publicity surrounding the disclosures that followed the Russian government-directed compromises of emails from U.S. persons and institutions, including from U.S. political organizations, would raise questions about the integrity of the election process that could have undermined the legitimacy of the president-elect."

That was a reference to the breach of the Democratic National Committee's email system, and the leak of emails from figures like John D. Podesta, Hillary Clinton's campaign chairman.

"Nevertheless, we stand behind our election results, which accurately reflect the will of the American people," it added.

Supporters of Mrs. Clinton have enthusiastically backed the notion of challenging the results in the three states as a last-ditch effort to reverse Mr. Trump's clear majority in the Electoral College. They have seized on suggestions by some computer scientists that the states, which were crucial to Mr. Trump's victory, need to manually

Continued on Page A12

OBITUARIES B7

TV Mom in Turbulent Times

Florence Henderson juggled a blended family on "The Brady Bunch."

INTERNATIONAL A3-9

Cash Crunch and Empty Stores

When Prime Minister Narendra Modi banned 500- and 1,000-rupee bills, his goal was to spur the economy in India. But the move seems to be having the opposite effect. PAGE A4

Turkey Lashes Out on Migrants

President Recep Tayyip Erdogan, angry at moves to suspend talks on Turkey's entry into the E.U., threatened to let migrants flood into Europe. PAGE A4

BUSINESS DAY B1-6

Biggest Money Question of All

The days after Thanksgiving may be the perfect time to ask, how much is enough, Ron Lieber writes. PAGE B1

NEW YORK A17-19

A Skyscraper's Balancing Act

A 985-foot tower at Hudson Yards will sit on an unusual foundation. PAGE A17

ARTS C1-8

Master of Optical Illusion

A cluster of exhibitions in the United States has ignited interest in Julio Le Parc, an Argentine kinetic artist who works with light and space. PAGE C1

SPORTSSATURDAY D1-8

No Guarantee for Michigan

Coach Jim Harbaugh will lead the No. 3 Wolverines against No. 2 Ohio State, but he has not promised a win, as he did

THIS WEEKEND

Phase 1: Project Initiation and Planning

The first phase of the project entailed planning the project phases and onboarding the NMA-selected participants. The Library Services project team requested a diverse participant group with differing circulation volumes, production and distribution methods in order to have a valid sample of the market. The four publishers selected by NMA – the Wall Street Journal, The New York Times, The Washington Post, and the Gannett Company (USA Today) – produce some of the largest circulated newspapers in the United States. To round out the sample, two small to medium titles were selected including the Des Moines Register and the Iowa City Press-Citizen, both published by the Gannett Company.

A kickoff meeting with the pilot participants was held on April 28, 2016. In the kickoff meeting, the Library Services team shared the goals of pilot, reviewed the instructions for preparing and delivering the sample files to the Library, and fielded questions.

Phase 2: Sample Analysis of Current ("As-is") Files and ePrint PDF Specification Development

The goal of Phase 2 was for Library Services to gather and analyze ePrint PDF files in order to form a baseline from which to discuss future deposit specifications. In this phase, the pilot participants were asked to send two weeks' worth of published content as it was natively produced, without any modifications to the production process. The publishers submitted a combined total 14,500 samples (24GB). These as-is samples were sent via postal carrier to the Library on physical media, such as flash/thumb drives and external hard drives, since the automated mechanisms for receiving files were not yet fully operational.

Sampling Methodology and Analysis of As-is ePrint Sample Files

A minimum of 15 samples from each publisher was examined, purposefully selected to represent a variety of publication dates and days of the week, editions and content sections. The technical characteristics of these files were reviewed and analyzed using validation tools including Adobe Acrobat XI Pro, PreFlight, JHOVE, DROID, as well as text and photo viewer tools such as NotePad++, HxD and IrfanView.

The samples exhibited varying technical characteristics, primarily with regard to format, metadata, and completeness. To assess format, reviewers checked the PDF version and verified whether the files had searchable text, embedded fonts, device-independent colorspaces, or content tagging for accessibility. The images were also examined for overall visual quality, format, and compression. In terms of metadata, reviewers examined the files for embedded metadata as well as any information included in the filename. To assess completeness, the PDF files were compared to original print copies as well as any comparable microfilm deposits held by the Library. In addition to checking section and total page counts, particular attention was paid to supplemental materials, such as weekly Sunday magazines, and content provided by a third-party, such as comics or ads.

Format Analysis of Sample Files

In general, all sample files could be opened with common PDF readers, and the text and images were viewable. The team made the following observations about the format of the sample files:

- The samples provided were PDF version 1.3, 1.4, and 1.5. None of the submitted samples was PDF/A, a form of PDF with enhanced features for archiving.
- While the majority of text was searchable, not all text was searchable. In general, headline and bylines were full-text searchable, but supplementary information such as comics and television listings were not searchable. Nor was text within advertisements when the advertisement was submitted as an image.
- In addition, each of the publishers embedded all fonts, which included a mix of TrueType and Type 1 fonts, thereby facilitating accurate text display.
- All publishers submitted images with device-independent color information, which is required for accurate color display.
- Text and other content was not tagged for accessibility, preventing assistive technology (such as screen readers for the visually impaired) to understand the correct reading order of the text as well as the presence and meaning of significant elements such as tables, figures, captions and lists.
- The images in the files were JPEGs with varying degrees of compression; most were 150-250 dpi with the JPEG quality between 50 and 98.

Metadata and File Naming Analysis of Sample Files

PDF files have the capacity to contain metadata embedded into the file itself. Some embedded metadata is auto-generated, conforming to presets or templates in the creating software and often contains information such as the file creation date and software name. PDF files can also contain more contextual information structured as Dublin Core or XMP data such as publisher name, title and publication date. This information must be purposefully embedded in the file through customized templates or automated scripts. Without good embedded contextual metadata, filenames and metadata outside of the file are critical; with such embedded metadata, the file is self-describing and more robust.

The composition of embedded metadata in the Phase 2 samples was sparse, and inconsistent when implemented. Most samples contained only auto-generated data and no enhanced data to describe the creation of the file, publisher information or publication dates. In two notable cases, the metadata differed between editions of the same title, and personal names of publisher staff members were included as part of the provenance data. One publisher created all the PDF files for the submitted batch on the same date; this date was stored as the file creation date in embedded metadata. The lack of an embedded publication date to provide this date information meant that the only place the publication date was known was in the masthead.

File names were in some cases problematic. While file names from some publishers contained structured data useful for understanding the structure of the issue, other file names were highly complex, requiring a detailed key to understand. In one case, the file names included special characters (+ , *), which are problematic in many computer systems.

Completeness Analysis of Sample Files

In the context of this project, a complete copy contains all of the pages of the issue as printed. To assess completeness, the PDF samples were compared to original print copies as well as any comparable microfilm deposits held by the Library. The composition of the issue varied. In Phase 2, most publishers submitted individual page-level PDFs within a date folder; one publisher submitted a single, multi-page PDF issue for each day. While rare, there were some instances in which a page was completely missing from an issue or only part of the published page was provided (such as one column of a multi-column page). In terms of supplemental material, two publishers included weekly Sunday magazine supplements in the Phase 2 samples.

There were two noticeable differences between the Phase 2 PDF deliveries and the comparable microfilm. Microfilm typically mimicked the issue as delivered to a customer in print form, often including Sunday third-party circulars and syndicated magazine supplements; the PDF deliveries lacked this content. Similarly, some PDF samples included multiple versions of the same page, sometimes for different geographical markets or sometimes as the page updated with breaking news. As microfilm reflects a single print issue, it does not include multiple versions of the same page.

Developing the ePrint PDF Specification

After reviewing the as-is samples, the Library Services project team created a tiered chart of Preferences for PDF Delivery (see *Appendix A: Preferences for PDF Delivery*) for the Phase 3 set of samples. These preferences were designed to encourage structure, uniformity and enhanced features for access and preservation of submitted files. Publishers were asked to review the preferences chart, review the output options for their local PDF production workflows, and generate a new set of Phase 3 "to-be" samples that incorporated more of the PDF preferences. These samples would then be delivered to the Library digitally via SFTP transfers.

The preferences are outlined below with the accompanying rationale for the preferences.

Format Preferences

- There should be no security measures such as digital rights management (DRM), passwords, encryption, etc. which would limit or prevent future access.
- PDF files need to be readable by Acrobat 5.0 or later versions, but ideally would conform to PDF/A standards to include enhanced archival features.
- As a default, all text, including but not limited to bylines, articles, classifieds, television schedules, obituaries and ads, should be searchable for easier discovery and access.
- Text should be structured in logical column-reading order to facilitate accessibility/Section 508 compliance.
- All fonts should be embedded to lessen the chance of dropped or garbled text due to substituted fonts.
- Fast-view or optimize for the web should not be implemented.
- Interactive content such external bookmarks, hyperlinks, named destinations, comments, forms, Javascript actions, external cross references, alternate images, embedded thumbnails,

annotations, or private data are not permitted because these can potentially compromise the security of the file and decrease its long-term viability.

- Multimedia content such as video or audio should not be embedded in the file because it increases rendering complexity.
- Each PDF will include all necessary embedded raster-based photographic images such that the file will open without error and all content is visible.
- Color information should be device independent to maintain accurate color display.

Metadata and File Naming Preferences

- Filenames must exhibit a uniform structure and include basic bibliographic information to facilitate scalable processing for many publishers.
- A limited required set of embedded metadata is essential to maintaining the document's provenance information as well as contributing to enhanced data description.

Completeness Preferences

- Issues are requested to be assembled in book form or with a single unifying XML file providing identification and logical page sequence information, rather than individual pages, to ensure the complete issue is delivered.
- PDF files will open without error, and all content is visible to assure access to the complete contents of the file.
- Newspaper issues submitted will contain the complete latest published edition, including all sections represented in the printed version (e.g. advertisements, magazines, etc.)

Phase 2 Summary

The Library Services team concluded Phase 2 with an understanding of the characteristics of ePrint PDFs commonly generated by large publishers. This knowledge shaped the content, technical, and delivery properties listed in the *Preferences for PDF Delivery*, the guiding document for Phase 3 samples.

Phase 3: ePrint Sample Analysis of Specification-compliant ("To-be") Files

Guided by the *Preferences* document (see *Appendix A: Preferences for PDF Delivery*), publishers were asked to create one month's worth of "to-be" samples and send to the Library via Secure File Transfer Protocol (SFTP). One publisher requested retrieval from their servers instead of depositing content on the Library's servers. To move forward with the pilot, the Library Services project team pursued this option and fetched the content from the publisher's servers. In this phase, the team aimed to test its automated transfer process for incoming sample files, and used a watch folder service to identify new files deposited by publishers each 24-hour period. When new files were found, automated processes at the Library were invoked to inventory and scan files for malware. Files were then bagged using the Bag-it specification and copied to staging servers for review.

Sampling Methodology and Analysis of To-be ePrint Sample Files

The publishers submitted a combined total 12,532 sample files (35 GB) for Phase 3. The evaluation method was the same method used in Phase 2, with a minimum of 15 samples from each publisher examined representing a variety of sections, content, publication dates and days of the week. Regarding delivery packaging, one publisher submitted two weeks' worth of issues (each issue named with an identifying date) as a single delivery. The other three publishers delivered issues daily.

Format Analysis of Sample Files

Most of the publishers appeared to have made modest changes in order to follow the LC preferences provided at the end of Phase 2, especially in regard to file format standard and embedded metadata. No publishers submitted PDF/A files, with most submitting PDF versions 1.3, 1.4 and 1.6.

More text was searchable including comics and television listings. Text stored as images in ads was not searchable (as expected). There were two unusual items found with regard to embedded text. The first was also a minor problem introduced in the switch from Phase 2 to Phase 3 samples; in the Phase 3 sample, the lowercase L's and uppercase i's in static text images (such as the publisher's block) are slightly oversized in comparison to the other text. This was not a problem in the Phase 2 sample. In another sample, sometimes the embedded text was concatenated. For example, "in Woodstock" on the page became "inWoodstock" in the embedded text.

As in Phase 2, the files did not include structured text for improved accessibility. Similarly, all publishers used device-independent color information embedded all fonts, although the specific types of TrueType and Type 1 fonts changed for each publisher.

Metadata and File Naming Analysis of Sample Files

No publisher added metadata beyond what is auto-generated in the file creation process, and in some cases provided even less metadata than in Phase 2. In the case of one publisher, issues with incorrect use of embedded metadata fields were corrected, but in other cases useful information such as embedded filenames was removed. The lack of key dates (namely publication date and file creation date) embedded as metadata continued to be an issue. Without this embedded metadata, the information must be gleaned from other sources or, in the case of publication date, derived from the masthead or the filename if the date is included there.

In one sample, photos could be extracted in uncropped form, and they included very specific and detailed captions and credit information in the photo metadata. The publisher would need to assess if such information was desirable to include in the copyright deposit.

File naming varied greatly between publishers assessed during the pilot. Two of the four publishers included an ISSN number as part of the name. Two publishers did not include edition number and title as part of the file name. One publisher, submitting multiple titles, used a complex naming convention (not including ISSN) requiring a key or map to parse and delivered all titles in a single directory. The team was glad to see that no special characters were included in file names.

Completeness Analysis of Sample Files

While three publishers submitted issues as book PDFs, one submitted single PDF pages without logical page sequence information. Furthermore, the delivered files of this publisher did not represent the final print issues, and instead represented the daily production output including multiple titles, editions, dates, partial issues, and replacement files.

There was a significant problem with rendering pages from one publisher. In one sample, some pages rendered in Adobe Acrobat, but not Adobe Reader; other pages rendered in Adobe Reader, but not Adobe Acrobat. This was primarily a problem with two-page spreads.

Another significant problem was the omission of some sections. For example, in one sample from Phase 3, the Classifieds section (typically 3 pages or so) were omitted from some issues. This was not a problem in the Phase 2 sample submitted.

Phase 3 Summary

The Library Services project team successfully tested an automated transfer process for incoming sample files. Additional workflow steps were tested to simulate the quality review of a sample batch and preservation of the files on long-term storage. Analysis of the Phase 3 samples revealed that some publishers were willing and able to make nominal modifications to their files in support of the pilot project. In some cases, however, the files received in Phase 3 exhibited problems that had not surfaced in the Phase 2 samples examined.

Conclusions

During the course of this pilot, two things were evaluated: ePrint PDF file characteristics as produced by publishers and reliable regular transfer methods.

Files

The PDF files created by publishers were, for the most part, useable for archival purposes, i.e. identifiable, renderable, reproducible, and searchable. Four areas of concern include partially renderable pages, missing pages, problematic file names, and a lack of logical page order.

Recommendation

The critical technical traits for files include:

- A screen-renderable and valid PDF according to the ISO 32000 family of specifications for PDF 1.7
- *Issue*-level file (rather than *page*-level files). Each PDF should present all pages in logical page sequence.
- Filename containing ISSN, publication date, and edition enumeration in the following format: ISSN_yyyymmdd_xx
- Embedded metadata must contain ISSN (in dc:identifier or xmp:identifier) and publisher date (in dc:date)
- Structured text (sometimes referred to as "tagged text") in support of Section 508 compliance
- Searchable text to the greatest extent possible (articles, advertisements, obituaries, etc.)
- All fonts must be embedded
- No proprietary access restrictions such as those implemented through Digital Rights Management schema or software.

Most sample files exhibited these traits. However, in order to be useable by the Library, the files must have easily understood file names, and all pages for the day's issue must be delivered in logical order, preferably encapsulated as one object. Without logical order and easily understood file names, it is difficult for Library staff to determine if all of the pages for the day's issue have been delivered. In practice, this means a single, multi-page PDF issue per day, with reliable identity information included in the filenaming structure or embedded metadata.

In terms of content characteristics, Library Services would recommend deposited ePrint files conform to similar characteristics as described in the current Group Registration for Newspapers/Newsletters circular (https://www.copyright.gov/circs/circ62a.pdf), regarding frequency of publication (daily), latest edition, and completeness.

Transfers

For the pilot, two transfer methods were tested.

Physical Delivery:

The first method tested was physical media delivery. Due to security restrictions on Capitol Hill, standard deliveries undergo significant processing delays as packages must be inspected off-site before delivery. For this pilot, an expedited method was used to circumvent such delays. Because the Library had not yet established a path for electronic transfer of files from publishers, physical delivery was used for Phase 2. The team does not consider physical delivery of files to be a viable long-term option for a variety of reasons, including processing delays and potential for mislabeling.

Digital Delivery:

The second method pursued was SFTP transfer: both "fetch," wherein the files are picked up from the publisher's server, and "catch," wherein the files are delivered to the Library's servers. The first thing to note was the SFTP transfer process. Currently, the Library's servers are configured to accept deliveries via SFTP only. One publisher mentioned that an FTP/FTPS option would be preferable, and lacking that, would rather have the Library fetch content from their servers.

For the content transfers, the Library provided SFTP account setup forms to the publishers to transfer files to the Library's servers. The account request form was provided in mid-October, and creating all of the accounts required some troubleshooting, from minor to major. This configuration process spanned from 1.5 months to 3 months. Two caveats include the fact that the troubleshooting process spanned the holidays, and that one publisher did not complete the SFTP account setup, but instead provided their server credentials to the Library.

ePrints representing the printed issue are generally large-format, full-color, with high quality illustrations and photographs resulting in multi-page files of significant data size. While file sizes were not a problem in the 4-week pilot delivery process using direct SFTP (deliveries were daily or weekly by publisher preference), by extrapolation, group deposits of this type of material would entail significant quantities of data delivered to the Library depending on the frequency of deposit expected. In the pilot, daily issue files averaged approx. 160 MB each (range 35 MB-265 MB per issue), and zipping the files does not reduce the file size significantly (about a 2% reduction on files tested). Similarly, the number of titles received could be expected to increase significantly beyond the small number submitted for the pilot. Decisions regarding the desired means of data transfer to the Library should take this into consideration.

In conclusion, this pilot project has achieved its goals of identifying current publisher practices for newspaper ePrints, evaluating technical file characteristics that can support Library Service's mission of sustaining a modern newspaper collection, and exploring the challenges of establishing possible future regular data transfers. The results give the team confidence that ePrints could be a viable technical alternative to microfilm deposits for newspapers. The team is grateful to the News Media Alliance and participating newspaper publishers for collaborating with the Library to explore this option.

Appendix A: Preferences for PDF Delivery

In Phase 3 of this project, the goal was for publishers to produce digital objects with as many Level 2 characteristics as possible. Please note that the digital objects did not need to be wholly Level 1 or Level 2, but could have some characteristics in either column; in addition, as indicated by the asterisk (*), some individual characteristics are actually part of the PDF-A specification and therefore included only as reference points.

Level 1 describes basic technical characteristics for files that the Library would like to acquire. Providing e-print PDFs that exhibit these characteristics will allow the Library to implement basic collection practices, relative to long-term management concerns.

Level 2 describes optimal characteristics for preservation, and content with those characteristics would be more likely to meet the Library's collection and long-term preservation needs over time with the least resource impact.

Not permitted means not allowable by desired format specification.

Not acceptable means not acceptable to the Library of Congress.

** Indicates required by PDF/A specification.*

Property	**Level 1 – Compliance to basic non-archival implementations of PDF.**	**Level 2 – Enhanced preservation ready files with compliance to more structured and data rich implementations of PDF for archiving and preservation**
TECHNICAL Properties		
Format compliance	PDF will be readable by Acrobat 5.0 or later versions	• PDF/A-1b (ISO 19005-1:2005) • PDF/A-2b (ISO 19005-2:2010) • PDF/A-1a (ISO 19005-1:2005) • PDF/A-2a (ISO 19005-2:2010)
File structure and filename	Publisher, title, and publication date information incorporated into file name and directory structures	ISSN and publication date information must be incorporated into the filename and directory structures (ISSN/ISSN_yyyy-mm-dd_edition number).

Property	**Level 1 – Compliance to basic non-archival implementations of PDF.**	**Level 2 – Enhanced preservation ready files with compliance to more structured and data rich implementations of PDF for archiving and preservation**
Issues	Multiple page-level PDFs must sort in page sequence order or include a single unifying XML document providing identification and logical page sequence information	One multi-page PDF per issue with pages presented in a logical page sequence
Security measures (DRM, passwords, encryption, etc.)	Not acceptable	Not permitted and not acceptable*
Searchable text	Selected text is searchable; at a minimum article text, titles and bylines.	All text is searchable including but not limited to article text, titles and bylines, classified ads, commercial advertisements (if submitted as text, not images), community calendars and TV programming listings.
Structured text	Not implemented	All content is tagged, structured in logical column-reading order during creation*
Fonts	Selected fonts embedded	All fonts must be embedded and also must be legally embeddable for unlimited, universal rendering*
”Fast View” / Linearization	Permitted	Not implemented (not strictly incompatible with PDF/A but many PDF/A tools will ignore it if present.)

Property	**Level 1 – Compliance to basic non-archival implementations of PDF.**	**Level 2 – Enhanced preservation ready files with compliance to more structured and data rich implementations of PDF for archiving and preservation**
Embedded metadata	XMP preferred but not required: • publisher [dc:publisher/no XMP alternative] • title [dc:title/no xmp alternative] • ISSN [dc:identifier/xmp:identifier] • published date [dc:date/no xmp alternative] • geographic coverage of edition if specialized [dc:coverage]	XMP required* <pdfaid:part> and <pdfaid:conformance> Preferred but not required: • publisher [dc:publisher/no XMP alternative] • title [dc:title/no xmp alternative] • ISSN [dc:identifier/xmp:identifier] • published date [dc:date/no xmp alternative] • geographic coverage of edition if specialized [dc:coverage]
Interactive content (bookmarks, hyperlinks, etc.)	Permitted	No bookmarks, hyperlinks, named destinations, comments, forms, Javascript actions, external cross references, alternate images, embedded thumbnails, annotations, or private data*
Multimedia content (audio, video)	Permitted	Not permitted*
Image information	Each PDF will include all necessary embedded raster-based photographic images and vector-based graphics.	Each PDF will include all necessary embedded raster-based photographic images and vector-based graphics.*

Property	**Level 1 – Compliance to basic non-archival implementations of PDF.**	**Level 2 – Enhanced preservation ready files with compliance to more structured and data rich implementations of PDF for archiving and preservation**
Behavior upon opening / rendering	PDF file will open without error and content is visible.	• PDF file will open without error and content is visible. • The PDF will open to Fit Page sizing. • The PDF will open to single page layout. • The PDF will open with neither document outline nor thumbnail images available. • The PDF will open with the tool bar, menu bar, and user interface elements visible. • The PDF will not open centered in the screen.
Color information	Color space, such as CIE and ICC, may be defined in a device independent manner	Color spaces must be specified in a device-independent manner*
CONTENT Properties		
Completeness	Complete edition (latest, preferred)	All published content (advertisements, magazines, etc.)
Resolution	Any	High
DELIVERY Properties		
Transfer process	Signiant/Email	SFTP to LC Servers
Frequency	Daily or less	Weekly

Appendix B: Summary of Publisher Phase 3 Samples

Publisher	Publisher A	Publisher B	Publisher C	Publisher D
Delivery	Fetch*	Catch**	Catch	Catch
Troubleshooting Required?	Publisher delay in server setup	no issues	no response until January	Publisher tool configuration problems
Delivery frequency	daily	weekly	daily	daily
Single Page or Multipage PDF	multi	multi	multi	single
Average file size (Mb)	67	269	25	2
Average # of Pages	39	90	49	Title A-185 Title B-108 Title C-17
PDF version	1.3	1.7	1.6	1.4
Embedded Metadata	auto-generated only	auto-generated only	auto-generated only	auto-generated only
Filename structure	issn_date_edition	title_issn_date	title_date	lengthy code
ISSN included in filename	yes	yes	no	no
DPI	250	200	150	225
Renderability	yes	yes, but some spreads missing	yes	yes
Embedded Fonts	yes	yes	yes	yes
Text Searchability	yes	yes	yes	yes
Accessibility (508)	no	no	no	no
Security measures in files	no	no	no	no

** "Fetch" = Automated LC retrieval from publisher's server*

*** "Catch" = Publish delivery to LC server (SFTP)*