**Capture and Preservation of Born Digital News**

*March 5-6*

Universidad Nacional Autónoma de México, México City

New Newspapers: Changes in Acquisitions and Access at the U.S. Library of Congress

Robin Butterhof

Serial and Government Publications Division, Library of Congress, Washington, DC, United States of America

Creative Commons LicenseThis work is made available under the terms of the Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0>

Abstract:

*The U.S. Library of Congress collects both domestic and foreign newspapers. While foreign newspapers are typically acquired via gift or purchase, domestic newspapers are acquired primarily via legal deposit. During the past two years, new legal deposit provisions for newspapers have led to rapid changes in acquisitions and access at the Library of Congress.*

*In terms of acquisitions, the Library of Congress does not now—nor has historically sought to—comprehensively collect all newspapers published in the United States. Rather, it generally seeks representative coverage of newspaper publishing in each state and territory based on criteria laid out in the collection development policy. In practice, this list includes approximately 400 current American newspaper titles.*

*While the collecting policy has not changed over the past two years, the technical methods and formats for acquiring these publications are starting to change. In early 2018, U.S. copyright regulations were amended to require online registration and digital deposits for group registration of newspapers, the primary method of registering newspapers for copyright protection in the United States. The first digital copyright deposits under these regulations arrived in March 2018. During the summer of 2019, a second digital delivery method was created for newspaper publishers that did not choose to register their titles, but whose publications were selected by the Library for its collections under legal deposit regulations.*

*To provide access to these copyright-restricted collections, an onsite access system for this content as well as other restricted non-newspaper content was built in 2019 and is now available to researchers at the Library. In addition to legal deposit for print newspapers, the Library of Congress has used web archiving as a technique to capture online-only news sites since 2014.*

**Keywords:** PDF, newspapers, legal deposit

# Copyright and Legal Deposit Requirements in the United States

In the United States, the first federal legislation for copyright protection dates to 1790. The Copyright Act of 1790 covered “maps, charts, book[s]” and required the copyrighted items to be deposited with the Secretary of State. The law then had several minor revisions and its first general revision in 1831. After the 1831 revision, minor provisions were added (1846) and later removed (1859) for copyright deposits to be delivered to the Library of Congress and the Smithsonian.

In the second major revision of 1870, copyright administration was centralized at the Library of Congress. Registrants were also required to send deposits to the Library of Congress, a significant boon to the Library’s collections. In 1897, the Copyright Office became a separate department within the Library of Congress, and a Register of Copyrights was appointed.

With the third major revision in 1909, deposits were required to be the “best edition.” The requirement of best edition was further clarified in the fourth revision of the law, the Copyright Act of 1976. In that, the best edition was defined as “the edition, published in the United States at any time before the date of deposit, that the Library of Congress determines to be most suitable for its purposes.” The best edition is specific to the type of work, with preferred formats for books, photographs, movies and so on. This list of preferences was codified in the “Best Edition of Published Copyrighted Works for the Collections of the Library of Congress,” also known as the best edition statement.

The best edition statement only applies if a work is published in multiple formats. If a work is unpublished or only published in one format, the existing format is the best edition. For example, an unpublished novel could be registered in the form of a word processing document as that is the only version of the work. For a novel published only in paperback, a paperback copy could be used as the registration deposit as that was the only format published. If that novel was published in both hardcover and paperback format, however, the hardcover version would need to deposited as that is preferred as per the best edition statement.

The Copyright Act of 1976 also distinguished between the activities of legal deposit (17 U.S.C. § 407) and registration (17 U.S.C. § 408). With the Copyright Act of 1976, copyright protection was established at the point a work was fixed in tangible form; formal registration or notice were no longer necessary. With that change, registration became an opt-in activity that provides certain legal benefits, creates a registration record, and requires a fee.

Legal deposit, on the other hand, does not require a fee, does not provide certain legal benefits, and is not part of the registration record. It is typically initiated at the request of the Library rather than the publisher. Materials published in the United States are subject to legal deposit unless specifically exempted by the Copyright Office; this exemption is known as a grant of special relief. For instance, a publisher could apply for special relief if sending the deposits would cause undue hardship. While registration and legal deposit are considered separate activities, deposits sent in for registration purposes can be used to satisfy legal deposit requirements.

# Newspapers and Copyright Registration

For the first half of the 20th century, very few publishers chose to register their newspapers for copyright protection. At that time, publishers had to register each individual issue for a separate fee. The University of Pennsylvania has compiled information on newspapers and other periodicals that did choose to register from the Copyright Office’s Catalog of Copyright Entries (CCE) books. Although whole issues were not registered by the newspaper publishers, newspapers often included licensed content such as comic strips and syndicated columns that were independently registered for copyright protection.

In 1990, the Copyright Office created a new form of registration that allowed for 3 months of a weekly or monthly serial title to be registered at one time for a single fee; this registration specifically excluded daily titles. The Newspaper Association of America (NAA) then sought the creation of a similar group registration option for daily newspapers.

“The Newspaper Association of America approached the Copyright Office and the Library about establishing a group registration privilege for daily newspapers. Because of the expense of registering individual issues, currently most daily newspapers are not registered in the Copyright Office. While the lack of registration does not invalidate copyright, it does limit the availability of certain remedies should infringement occur. The Library of Congress studied carefully the advantages of group registration from the standpoint of securing acquisitions to the collections. Currently, the Library purchases microfilm copies of selected daily newspapers. If these microfilm copies can be secured on a timely basis through copyright deposit, a substantial cost savings will accrue to the Library.”

In September 1992, the “Group Registration of Daily Newspapers” registration option was established, and a publisher could deposit a month of issues on microfilm for one registration fee. There were several requirements for this option.

* Only daily newspapers were eligible.
* The microfilm needed to be “positive, 35mm silver halide microfilm meeting the Library’s best edition criteria that reproduces in their entirety all issues published as final editions with issue dates in the designated calendar month.”
* Registration was required “within three months from the date of publication of the last issue included in the group registration application.”
* Only serials defined as “newspapers” by the Library of Congress were permitted to use this option.

Adding this option for news publishers led to a dramatic increase in the quantity and efficiency of registering daily newspapers for copyright protection. In addition, much of the deposited content could be used to fulfil legal deposit requirements and augment the Library’s collections. As the microfilming industry contracted, however, the Library began to explore alternate approaches to collecting newspapers. At the same time, the news and publishing industry began to shift to more dynamic and complex forms of distribution.

# Technical Pilot Projects

In response to the changing news landscape, the Library ran a technical pilot project in 2012 to 1) examine PDF files produced by newspaper publishers and 2) explore web archiving as a way to capture news sites. The pilot tested web archiving techniques on several news sites and assessed the capture of those sites. It also examined files from five newspaper publishers and listed preferred and required file characteristics for online-only deposits. The pilot concluded with a report on the pilot process and outcomes.

In 2016, the Newspaper Association of America (NNA) changed its name to the News Media Alliance (NMA) and opened membership to digital-only news sites. As digital publishing had matured since the 2012 pilot project, the NMA expressed interest in a new technical pilot. This pilot ran from April 2016-March 2017. In the first phase of the pilot, a few publishers supplied the files used for online/electronic distribution or printing purposes. In the second phase of the process, the Library provided a list of preferred characteristics for the files, and publishers were asked to assess how difficult it would be to adopt the preferred characteristics. If the characteristic was regarded as easy, publishers were to adopt those characteristics for the next round of file deliveries. The pilot concluded with the publishers delivering files to the Library. At the end of the pilot, it was determined that the PDFs of print issues as produced by publishers “could be a viable technical alternative to microfilm deposits for newspapers.”

# Changing Group Registration for Newspapers

After the 2016 pilot had determined that PDFs could be an acceptable alternative to microfilm, discussions with the Copyright Office began on how to change the current requirements to permit digital deposits for newspapers. Unlike in other countries, the shift to digital deposits was not a single sweeping change for all copyrighted items; other kinds of works, such as books and photographs would fall under separate rules.

To update the group registration of newspapers, technical requirements for the files were based on the pilot outcomes, and policy changes were discussed by Library collections staff and Copyright Office staff. The revised requirements for registration were posted for public review on November 6, 2017, and the new rule went into effect January 30, 2018.

The rule included several key provisions.

* Issues were to be deposited within 3 months after the publication of the earliest issue in the group
* Weekly and semiweekly newspapers could use this registration option in addition to dailies
* Each registration would include 1 month of newspaper issues
* The application and deposit files would be uploaded via the Copyright Office’s Electronic Copyright Office (eCO) website
* The Library of Congress could provide public access to content with the following stipulations
  + Two Library of Congress authorized users via a secure server over a secure network that serves Library of Congress premises.
  + (d) ‘‘Authorized user’’ means Library of Congress staff, contractors, and registered researchers, and Members, staff and officers of the U.S. House of Representatives and the U.S. Senate for the purposes of this section.
  + (e) ‘‘Library of Congress premises’’ means all Library of Congress premises in Washington, DC, and the Library of Congress Packard Campus for Audio– Visual Conservation in Culpeper, VA
* File naming and structure
  + Files must be named according to specifications, namely GRNP\_[ISSN]\_[NISOdate].
    - Ex. GRNP\_12345678\_20200101 for an issue published on January 1, 2020 with International Standard Serial Number (ISSN) 1234-5678.
  + Whole issues must be submitted instead of individual page-level PDFs
* File characteristics
  + Must be viewable and searchable
  + Must contain embedded fonts
  + Be free from any access restrictions (such as those implemented through Digital Rights Management (DRM)).

ISSNs were chosen as the standard filename identifier as ISSNs are free to obtain via an established publisher customer service channel, and ISSNs are commonly used by many publishers.

In addition, the new rule noted that newspaper websites cannot be registered under this application option.

“As noted in the NPRM, the final rule may be used to register a newspaper that is distributed in an electronic format, such as a PDF version of a physical publication. To do so, the publisher would have to demonstrate that each issue contains a fixed selection of content, each issue is distributed as a collective work, and the content of each issue does not change once it has been distributed to the public. To the extent the commenter is referring to newspaper websites, the Office reiterates that a website would not be considered a ‘‘newspaper’’ for purposes of this group registration option, for the reasons stated in the NPRM.”

In February 2019, the 3-month requirement was dropped at the request of publishers. Publishers that had missed the 3-month window would have to register single issues for individual fees after that point, a much more expensive and arduous process. When the 3-month rule was lifted, some publishers registered content going back to 2015 as they could now do so. Notably, while the 3-month requirement was dropped from the rule, there are still significant benefits to registering within the 3-month window, as noted in the rule change.

“The Office still encourages publishers to submit their claims within three months of publication, because it may provide certain legal benefits. To seek statutory damages and attorney’s fees in an infringement action, publishers must register their issues in a timely manner. Specifically, a publisher typically may seek these remedies if a newspaper issue was registered (i) before the infringement commenced or (ii) within three months after the first publication of that work. See 17 U.S.C. 412.”

# Acquisitions and Review of Group Registration Deposits

Registration applications and deposits are reviewed by copyright examiners to determine that the content is indeed eligible for registration and that non-deposit materials are not included.

After the content has been reviewed and the registration certificate issued by the Copyright Office, the original deposit is retained by the Copyright Office for official record, and a copy is transferred to the Library’s collections.

In addition to the PDF files, some metadata from the copyright application is transferred with the files. This includes the copyright tracking number, the month of the delivery, and the title as supplied by the publisher. This information is particularly helpful to collections staff in checking for partial deliveries, content inconsistencies, or transfer problems. For example, one publisher registered a deposit under X title, but all of the PDF files were for Y title.

# Changes with Legal Deposits

When group registration changed, the publisher community expressed an interest in submitting PDFs to the Library to satisfy legal deposit requirements. Due to the best edition requirement, legal deposits still had to be submitted in microfilm or print format. There is fluidity between registration and legal deposit as a publisher opts to register an individual title on a month-by-month basis. If a publisher decides not to register a particular month of a title, it may then be eligible for acquisition under legal deposit.

In 2019, the Copyright Office began accepting applications for special relief, allowing qualified publishers to deposit PDFs. One large newspaper publisher qualified for special relief and signed a special relief agreement with the Copyright Office. As part of the agreement, the publisher sent a test day of files following the same file naming and other requirements of the group registration option. After the test files were approved, the publisher set up automated delivery for its titles. This process for obtaining legal deposits is slightly different than the process for registration deposits in that legal deposits are sent via Secure File Transfer Protocol (sftp) daily deliveries instead of Copyright Office monthly uploads. However, both types of deposits are copied to the Library’s preservation storage in a similar method.

# Processing Deposits

The registration deposits and legal deposits are automatically picked up by the Library’s content transfer services (CTS). Upon pickup, the content is bagged according to the BagIt specification; this process creates checksums and generates a manifest of the files. The bag is then copied to long-term storage in the Library’s digital repository.

After the copy to long-term storage, the content is sorted into items that are retained under Library’s collection development policy and those that are not. As noted in the Collection Policy Statement for US Newspapers,

“The Library is not comprehensive in its collecting of United States newspapers. Instead it seeks representative coverage of newspaper publishing in each state and territory to reflect the news and reading habits of the public over time and to support the information needs of Congress and researchers. The Library relies on state libraries and other local and regional repositories to acquire and preserve local newspapers and make them available for research.”

In practice, the overarching guidelines of the collection policy are translated into a list of approximately 400 titles called “US Newspapers Currently Received.” This list includes a representative coverage of newspaper publishing in each state or territory. Some registered newspapers fall outside the collection development policy guidelines, and those titles are not retained for the Library’s collection. Those titles are retained by the Copyright Office under their retention policies.

For titles that are on the “US Newspapers Currently Received” list, a copy is moved to processing storage, and duplicate deliveries for a single day are reviewed. While the filename structure was intended to prevent duplicate same-day deliveries, a timestamp appended to the filename by the registration system allows for duplicate deliveries. With the legal deposit daily deliveries, multiple issues for the same day could appear as the files are picked up and bagged on a daily basis. For the legal deposit daily deliveries, individual issues are combined into 1-month bags to match the format of the registration deposits.

After the files have been organized, the International Standard Serial Number (ISSN) is reviewed to ensure that the ISSN is the correct ISSN for that particular title. Publishers sometimes used the print ISSN in their filename structure for their registration deposits. While the PDFs delivered to the Library are often used for the production of the print newspaper, the ISSN Center and cataloging policy determined that an online ISSN is the appropriate type of ISSN for these digital deposits. The online ISSN also requires a separate bibliographic record; the digital deposit cannot be attached to the print bibliographic record.

After the initial review, the ISSNs are converted to Library of Congress Control Numbers (LCCNs) for access purposes, and the filenames are normalized. After the normalization and correction work is complete, a metadata file and the corrected bags enter a second workflow. This workflow converts the metadata file into individual bag-info.txt files that accompany each bag.

**Example** - el Nuevo Herald (Miami, FL) - November 2019 issues

*Bag name*: 2019201720\_201911

*bag-info.txt file*

LC-Keep: Yes

LC-Project: sgpnpemanpr

External-Identifier: 2019201720 (LCCN)

LC-LCCN: 2019201720

LC-ISSN: 2688-8793

LC-Content-Type: textual

Internal-Sender-Identifier: 2688-8793 (ISSN)

LC-Items: 30 newspaper issues

LC-Title: el Nuevo Herald

LC-Content-Custodian: Serial And Government Publications Division

LC-Content-Acquisition: copyright\_deposit

LC-Content-Process: content\_transfer

LC-Content-Date: 201911

*manifest-md5.txt file*

[checksum] data/2019201720\_20191101.pdf

[checksum] data/2019201720\_20191102.pdf

[checksum] data/2019201720\_20191103.pdf

. . .

(Some of the fields in the bag-info.txt field are duplicative as certain fields are indexed in the digital repository and other local fields are not.)

During the second workflow, content is reviewed in a local quality review tool that supports automated sampling of images and metadata. One common discovery is duplicate pages. These duplicate pages can be due to corrections made to a page over the course of a day or if the newspaper published several local editions with ads unique to that local area. Similarly, missing pages from issues are identified. Any problems are noted, and the content is then copied to long-term storage and access storage.

# Providing Access to Deposits

In October 2018, the Library published the FY2019-FY2023 Digital Strategic Plan. The plan noted:

“We will expand our ability to provide secure on-site access to rights-restricted digital material . . . Additionally, we will build systems that will allow users the maximum authorized access to content, based on their role (e.g., congressional user, staff, or researcher) and on or off-site location.”

When the Digital Strategic Plan was published, newspaper PDFs had been collected for six months, and it was important to the reference staff to provide access to those PDFs as quickly as possible. Access to recent microfilm and print issues was an established service in the reading room, and some titles were not available via other means, such as subscription databases or print subscriptions. This was particularly important for situations where the patron required the page image and not just the article text.

With the publication of the Digital Strategic Plan, the Library’s legal team and collections staff investigated how to best maintain research access while respecting intellectual property rights. This resulted in guidelines for rights-restricted digital collection items in multiple formats, including books and newspapers. While the new registration rule had established guidelines on who could access the files and where the files could be accessed, the legal team laid out the functional restrictions. For the newspaper deposits, two users could access the same issue at the same time, and printing was allowed as the takeaway method. Downloads or emailing were not permitted. Similar restrictions were established for other items depending on the format type and method of acquisition. Some purchased foreign newspapers are accessible on-site to an unlimited number of users. Meanwhile, books have a simultaneous user limit for the whole book, rather than chapters or pages.

Once the legal team had provided a framework, a technical project was chartered for the access system in early 2019. The project started with a kickoff, a review of the current bag structure and content types, a cognitive walkthrough, and the creation of user stories. Due to technical constraints, patrons can access the content via specified user terminals in reading rooms; staff can access the content at their local workstations via authentication. The public access terminals are identified as a single patron account; patrons are not required to separately log onto the machines.

Although isolated from the main loc.gov site for security reasons, the user experience (UX) for the new access system was largely based on UX work on the main site. This provides a consistent search experience for researchers.

* Basic and advanced full-text search options
* Faceting by location, date, format, and language
* Keyword in context snippets in the search results

The access restrictions for this content did lead to some unique design features. For a newspaper issue restricted to two simultaneous users, the third user will get an error message to try again later. Similarly, a user will get a timeout message after ten minutes so an issue isn’t locked for a whole day. There is also site redirect for unauthorized users with information on how to get to the restricted access terminals.

To provide access to the digital files and metadata, the access application employs an extract, transform, and load (ETL) sequence. The system uses Apache Tika for file type detection and text extraction, Compact Language Detector 2 for language detection, Elasticsearch for indexing, and pdf.js for the PDF viewer. At the start of the process, Tika attempts to extract any text in the file. If no content can be extracted, the file is indexed as “no language content.” Of the 15,000+ deposit files, eleven issues encountered text extraction difficulties. Remediation options such as optical character recognition (OCR) or manual reprocessing may be used for these issues in the future. Similarly, there can be some imperfect outcomes with the text extraction depending on the design of the document. For example, “nominations added” in a particular issue was extracted as “nomina-tions added.”

After the text has been extracted, Compact Language Detector 2 detects the language; the content is then indexed according to that language. If the issue is multilingual, the issue is indexed as both languages. There is a 2% cutoff for language content to ignore loan words from other languages and erroneous content. This prevents a few words from changing the categorization of an issue to be multilingual.

One current requirement for inclusion in the access application is a catalog record with a Library of Congress Control Number (LCCN). The application first checks for an LCCN in the bag-info.txt file. If there is an LCCN, it checks the LCCN permalink service (lccn.loc.gov) to see if there is a public record for that LCCN. It then pulls the metadata associated with that LCCN and applies that metadata to all of the issues in the bag. At regular intervals, the application checks for changes in the records, pulls the new information from the record, and re-indexes the content if necessary. This system allows catalogers to make corrections to the bibliographic record without having to update the bag metadata. There is a side benefit to making the content accessible; namely, it surfaces inaccurate or incomplete information in the catalog records that can then be fixed. For example, a title will show up as having “no language” in the language facets if a catalog record is missing the language code in the 008.

For display, the full PDF files are currently made available without image compression or splitting the documents into individual pages. With this, larger files can be slow to load in the pdf.js viewer, and it can cause problems with printing. Other PDF viewers were investigated, but the customization possible with pdf.js was important for the project. The large PDF problem is also unique to newspapers and less of a problem for other format types, like books and legal gazettes.

For reporting purposes, Kibana is the primary tool for access statistics as well as ETL information. For ETL, Kibana reports on failed processes, such as incorrect filenames, no LCCN assigned, no extractable text, etc. It also reports on incomplete metadata, such as a lack of language code in the 008. This reporting is important as the content is omitted from search results without this information. For access, Kibana also provides reports on staff and public access sessions.

# Access Statistics

Since the debut of the access application, there have been 2,299 user sessions via the public access terminal and by staff. Over 180 unique staff members have accessed the site via their local workstations. As of February 2020, 12,972 registration deposit issues and 2,031 legal deposit issues from 63 newspaper titles are available. Legal deposit issues are available through January 2020. As publishers can register content when they wish after the removal of the 3-month requirement, the dates of the registration deposits vary depending on when the content was registered with the Copyright Office. Staff members have sometimes mistaken these variations as content embargos as that is typically the case for subscription databases.

# Collecting News Websites

Outside of the content being collected via registration and legal deposit, the Library also has collected online-only news websites since 2014. Publications include the *Huffington Post*, the *Texas Tribune*, the *Drudge Report*, the *Onion*, and others.

This content is not harvested under legal deposit provisions, but by notifying publishers of the intent to harvest and requesting permission to make the archive publicly accessible. If the publisher agrees to public access, the site is made publicly available approximately one year after the initial capture. If the publisher declines to provide public access, the title is only available on-site. As publishers have cut back from publishing daily issues to publishing weekly or semi-weekly, the Library has contacted those publishers about web archiving to fill in the gaps. One recent example is the *Pittsburgh Post-Gazette.* In September 2019, print publication was cut back to Thursday, Friday, and Sunday; the remainder of the week is published digitally.

Some news sites are difficult to capture via standard web archiving methods. A standard crawl typically harvests a few links past the home page, and the site may be captured daily. News sites frequently refresh the home page, and new articles are often several links away from the home page. When possible, the Library harvests the site’s RSS feeds to capture articles as they are being published. These RSS feeds are then integrated back into the main capture for a seamless experience for users. Jones’s article “Using RSS to Improve Web Harvest Results for News Web Sites” has more information on this web archiving method.

As newspapers shift to new distribution methods, new methods of acquiring this important content are required. Over the past two years, the Library has significantly changed its workflows for collecting and providing access to newspapers to keep pace with these changes.

References

“Best Edition” of Published Copyrighted Works for the Collections of the Library of Congress. 37 CFR § 202 Appendix B. <https://www.copyright.gov/title37/202/37cfr202-appb.html>

Group Registration of Daily Newspapers, 37 CFR § 202 (1992). <https://www.govinfo.gov/content/pkg/FR-1992-09-01/pdf/FR-1992-09-01.pdf>

Group Registration of Newspapers, 37 CFR § 201 and 202 (2018). <https://www.govinfo.gov/content/pkg/FR-2018-01-30/pdf/2018-01838.pdf>

Group Registration of Daily Newspapers, 37 CFR § 202 (2019). <https://www.govinfo.gov/content/pkg/FR-2019-02-13/pdf/2019-02186.pdf>

Jones, G. & Neubert, M. (2017). Using RSS to Improve Web Harvest Results for News Web Sites. *Journal of Western Archives. 8*(2), Article 3. <https://digitalcommons.usu.edu/westernarchives/vol8/iss2/3>

Library of Congress. (2020). *General News on the Internet Web Archive.* <https://www.loc.gov/collections/general-news-on-the-internet-web-archive/>.

Library of Congress. (2018, October 1). *FY2019-2023 Digital Strategic Plan of the Library of Congress.* <https://www.loc.gov/static/portals/digital-strategy/documents/Library-of-Congress-Digital-Strategy-v1.1.1.pdf>

Library of Congress and News Media Alliance (NMA). (2017, March 8). *Technical Pilot Project: Report on ePrints.* <https://www.newsmediaalliance.org/wp-content/uploads/2017/11/LC_NMA_ePrint-Report_2017-3-8-1.pdf>

Library of Congress. (2020). *US Newspapers Currently Received.*

<https://www.loc.gov/rr/news/ncr_list.php>

Library of Congress. (2008). *Collection Policy Statements: Newspapers – United States.* <https://www.loc.gov/acq/devpol/neu.pdf>

Ockerbloom, J. M. (2020). *First copyright renewals for periodicals.* The Online Books Page. <https://onlinebooks.library.upenn.edu/cce/firstperiod.html>