**A Survey of Born Digital Newspaper Collection Practices in the United States**

**Donnie Summerlin**

**Digital Library of Georgia**

**Background**

Since 1953, the Georgia Newspaper Project (GNP) has been microfilming historic newspapers from across the state. These materials have been a valuable primary source of the information for researchers across Georgia and beyond. Over time, however, increasing supply costs, equipment obsolescence, and user demands for online access to materials have necessitated a reassessment of how to approach the collection, preservation, and distribution of Georgia’s newspaper titles. As a result, the GNP, which is now a division of the Digital Library of Georgia (DLG) and is headquartered at the University of Georgia Main Library, began investigating alternative methods for collecting newspapers statewide.

One of the most promising options for change is the collection of print master (sometimes referred to as facsimile, digital edition, e-print, or print-ready) born digital newspaper files. Based on interactions with colleagues, the collection of these materials appears to be a cost-effective and scalable alternative to microfilming. Before the DLG was comfortable with making a transition away from microfilming, however, we felt that a more detailed assessment of file collection practices was necessary. We sought to understand the practical and technical considerations involved in establishing a project of this nature and decided that a national survey of born digital newspaper collection practices would be a valuable approach to seeking out that information.

**Literature Review**

Before preparing the survey, the DLG examined the professional literature on current born digital newspaper collection practices in the United States. Scholars and practitioners have published case studies, surveys, and standards guidelines related to born digital newspaper collection and preservation over the last decade. These works were invaluable in helping the DLG to establish a framework for a survey of born digital newspaper collection practices.

Several leading institutions have published research and case studies on born digital newspaper collection practices in the United States. Lois J. Widmer, Laurie N. Taylor, and Mark V. Sullivan detail in “Born Digital Newspaper Preservation Workflows for the Florida Digital Newspaper Library and the Caribbean Newspaper Digital Library” the institutions’ early adoption of born digital practices, including file transfers from publishers and website harvesting.[[1]](#footnote-1) Ana Krahmer, Mark Phillips and Ron Larson published “Two states provide models for newspaper preservation” in 2015 in the *Newspaper Research Journal*. The article describes the development of the newspaper collection programs at the University of North Texas Libraries and the Wisconsin Historical Society, where both institutions adopted workflows to collect and make available born digital newspaper content.[[2]](#footnote-2) In “Preserving Kentucky’s Newspapers: Analogue Beginnings to Digital Frontier,” Kopana Terry and Eric Weig detail the University of Kentucky’s work to harvest born digital newspapers from an aggregator service and prepare those materials for preservation and distribution.[[3]](#footnote-3) Sheila Rabun, in “Microfilm to Born-Digital for Current Newspapers: A Case Study from the University of Oregon Libraries,” documents the organization’s transition to born digital collection and provides a detailed description of their workflow.[[4]](#footnote-4)

Researchers have additionally reported on surveys of the born digital landscape. In “Born Digital Legal Deposit Policies and Practices,” Frederick Zarndt, Dorothy Carner, and Edward McCain update a previous survey to report on international born digital deposit procedures and regulations.[[5]](#footnote-5) Jennifer Salamon, in her paper, “Born-Digital News in Ohio: What Is (and Isn’t) Happening,” takes an intrastate survey approach and reports on the Ohio History Connection’s efforts to survey their state’s born digital newspaper collection activities and the practices of newspaper publishers in Ohio. These works were informative in helping the DLG shape its own national survey.[[6]](#footnote-6)

There is also a small body of literature produced on digital newspaper collection and preservation standards. The most prominent work on this subject is the “Guidelines for Digital Newspaper Preservation Readiness” by Katherine Skinner and Matt Schultz, which establishes a detailed set of recommendations for documenting, organizing and preserving digital newspapers, including born digital newspaper content. The report is essential reading for any institution interested in preserving their digital newspaper collections.[[7]](#footnote-7)

**A National Survey of State-wide Born Digital Newspaper Collection Practices**

After completing an examination of scholarly writings on the subject, the DLG drafted an online survey that included questions related to file types, collection methodologies, staff commitments, approaches to distribution, copyright permissions, and microfilming practices. The purpose of the survey was to collect practical information that would inform our preparations to establish a print master born digital newspaper collection project. We also hoped to share our findings with others interested in establishing or modifying their own collection programs.

The DLG conducted the online survey via Google Forms and publicized it on listservs and/or social media pages related to newspaper digitization and preservation, including those associated with the National Digital Newspaper Program (NDNP), the IFLA News Media Section, and Dodging the Memory Hole. We also contacted institutions and individuals directly, based on prior interactions and recommendations from other participants. The survey, which ran from July to November 2018, received nine responses from institutions that collect born digital newspaper files on a statewide basis. An additional institution responded with information on their intention to collect born digital newspaper files in the future. For the purposes of reporting, that response was excluded from the results.

Respondents were also given the option to provide their contact information or to remain anonymous. All of the participants chose to identify themselves and granted the DLG permission to contact them to ask additional questions. After the closing of the poll, the DLG spent early 2019 following up with participants to clarify responses and collect additional information on their workflows and collection practices. Based on these continued investigations, the list of survey participants appears to fairly represent the majority of institutions collecting born digital newspaper files statewide in the United States.

The results of the survey are summarized below. The questions have been grouped by subject for analysis purposes and do not necessarily reflect their order in the survey form.

**File Types**

With the myriad of file formats present in the twenty-first century, the DLG was interested to examine what types of newspaper files were produced by newspaper publishers and collected by institutions in the United States. In that vein, we asked participants “What file types do you harvest or collect?” They were offered choices of the most common file formats and were also given an “Other” option. The response was nearly unanimous. All nine participants selected PDFs, with a single institution also selecting JPEGs.

Follow up communications with these institutions confirm this uniformity to the point where “PDF” is often used as an eponym in place of terms like “born digital files.” This conformity can at least partially be explained by the prominence of publishing software that utilizes PDFs as the default file format for exporting. This assertion was later supported by a survey conducted by the DLG of newspaper publisher practices in Georgia, which confirmed that most of the publishers in the state were producing print master PDFs from publishing software associated with that specific file format.

**Collection Methods**

Although the survey demonstrated uniformity in the file types collected, there is no such homogeneity in regard to how those files are collected. Through professional literature and interaction with colleagues, the DLG identified three basic methods in practice for collecting born digital newspapers files: harvesting from an aggregator/clipping service, harvesting from publisher websites, file transfers/submissions from publishers. To better understand how common each of these practices are, we asked respondents “What method do you use to collect born digital newspapers?” and provided them with those three choices along with an “Other” option with a box allowing for explanation. Participants were allowed to select any or all of the options as they saw fit.

The most common collection method among the respondents is file transfers and submission from publishers, which is practiced by seven out of nine respondents (77%). The remaining collection methods were less common, with three of nine institutions harvesting from publisher websites (33%) and two out of nine institutions (22%) harvesting from an aggregator/clipping service. Two participants responded “Other,” and cited submissions from a state press association and purchasing files from an aggregator as collection methods, respectively. Both of these “Other” responses could also reasonably be classified as collecting files from an aggregator, which would boost the harvesting from an aggregator/clipping service collection method to four out nine respondents (44%).

**What method do you use to collect born digital newspapers?**



The results of this question suggest that there is no panacea for collecting born digital newspaper files. The chosen method of collection can depend largely on the circumstances within a state. For example, many states do not have an aggregator or clipping service that collects newspaper files from publishers around the state. This would force a collecting institution to seek out a more piecemeal approach to file collection. Furthermore, four out of nine of the respondents employ more than one approach to collecting newspapers within their state. This suggests that a single approach may not suffice if an institution is interested in collecting born digital newspaper titles in a comprehensive manner.

**Newspaper Titles Collected**

The survey also asked respondents: “How many born digital newspaper titles do you currently collect?” We provided survey-takers with five spans of numbers that ranged from “1-25” to “More than 150.” The results were spread fairly evenly amongst the respondents with three institutions collecting more than 150 titles, three collecting between 101 and 150 titles, two collecting between 26 and 50 titles, and a single institution collecting between 1 and 25 titles. None of the participants collect between 51 and 100 titles.

**How many born digital newspaper titles do you currently collect?**



To put those raw collection numbers in context, we asked survey participants “Generally, the newspaper titles you collect make up what percentage of all of the newspaper titles published in your state?” and left a blank box for the response. The resulting percentages ranged from less than 1% to approximately 75%. The approximate average percentage of titles collected by the participants within their state was 43%.

The three institutions that reported collecting more than fifty percent of the newspaper titles in their state all take advantage of an aggregator service or institution. For a majority of the institutions in a state where an aggregation service is unavailable, the collection percentage rate was below 50%. In follow up communications, several of these institutions ascribed rejections from publishers unwilling to participate as the largest obstacle to increasing their collection percentage rates. In this way, the process is unique from a microfilming workflow, which does not necessarily rely on participation from publishers. Additionally, publishers are more reticent to allow their papers to be made freely available online than they have been to allow newspapers to be distributed via microfilm.

The Georgia Newspaper Project currently microfilms approximately 78 percent of newspaper titles in the state. Without an aggregator service available in Georgia, that collection percentage will almost certainly drop during a transition from microfilming to born digital file collection. It is likely an unavoidable consequence of moving to the less costly and more technologically sound approach of collecting born digital newspapers. This ramification should be a consideration for any institution interested in a similar transition without the benefit of an aggregating organization or clipping service.

We also asked participants, “Are you only harvesting/collecting newspapers that you have permission to present online?” in order to understand whether intentions to host materials online affected the number of titles collected by each institution. Eight of the nine participants responded to the question directly, and of those respondents, five were collecting publications for which they were not given online hosting permissions. Follow up questioning of several of these institutions has revealed that they are often able to secure permission after the fact, which justifies their collection approach. In regard to the three institutions that are not collecting titles without online distribution permission, all of them are collecting titles on a publisher-by-publisher basis and do not have access to a aggregating service or institution. In these situations, a publisher is unlikely to agree to participate in a preservation program if they do not want the collecting institution to make those files publicly available.

**Are you only harvesting/collecting newspapers**

**that you have permission to present online?**



For those institutions interested in collecting born digital newspaper titles, a decision will have to be made whether it makes fiscal sense to collect titles without a guarantee that they can be made publicly accessible online. As the survey results demonstrate, this decision may be affected by the collecting options available within a state.

**Project Time Length**

In an attempt to measure the progress of collecting institutions over time, we also asked participants: “How long have you been collecting born digital newspapers?” The DLG hoped this question would estimate the size of the workload that might be expected at the onset of such an undertaking and how the number of titles collected might increase (or decrease) over time. Participants’ responses ranged from two years to fifteen years of born digital newspaper file collection experience.

Comparing those project time lengths to the number and percentage of titles collected produces no clear correlation. The length of time spent collecting newspaper files seems to have far less impact on the number of titles collected in comparison to the method used for collecting those files. Institutions that utilize aggregators to collect born digital files collect larger numbers of titles than their counterparts regardless of the time spent building the program. Additional factors, including the willingness of publishers to participate within a state and labor allocated to the project, also appear to have a closer correlation to the number of titles collected than does the length of time spent building a project.

**Labor Allocation**

The DLG also requested information on how institutions are devoting labor to the practice of collecting born digital newspaper files. The DLG currently employs two full time equivalent (FTE) employees with clerical staff designations to microfilm Georgia newspapers, along with the assistance of a team of student assistants. The intention was to use the survey responses on this subject to compare the labor currently allocated toward newspaper microfilming to the labor devoted to the practices of those that have transitioned to digital file collection workflows.

The first question on this topic asked participants “How many full time equivalents (FTEs) do you devote to born digital newspaper collection?” A vast majority of respondents (77%) devote less than 1 FTE (less than 40 hours/week) to the process. A single respondent replied that they devote 1 FTE (40 hours/week) and another respondent devotes 2 FTE (80 hours/week). These results imply that the adoption of a digital file collection workflow would not be more labor intensive than labor devoted to the DLG’s current newspaper microfilming operation.

**“How many full time equivalents (FTEs) do you devote to**

**born digital newspaper collection?**



We also provided respondents with a list of responsibilities and tasks associated with a born digital newspaper collection workflow and asked which of the activities are handled by professional staff, clerical staff, or both. The survey activities are listed below along with their associated response rates:

* Liaising with publishers (7/9 professional staff, 2/9 both)
* Harvesting/Collection (4/8 professional staff, 1/8 clerical staff, 3/8 both)
* Quality control (3/8 professional staff, 1/8 clerical staff, 4/8 both)
* Metadata creation (5/8 professional staff, 3/8 clerical staff)
* Preservation (8/8 professional staff)

**Of the following activities, which are handled by**

**professional staff, clerical staff, or both?**



The results illustrate a significant allocation of professional staff time to the born digital newspaper file collection process, particularly in responsibilities related to publisher liaising and preservation. These two activities diverge from the responsibilities currently allocated to GNP clerical staff. For example, the department receives physical newspaper subscriptions from publishers for microfilming, so there is currently very little interaction or coordination with publishers themselves. The results suggest that a reassessment of labor allocation may be necessary to decide if professional staff will be necessary when the workflow shifts from microfilming to file collection, or if retraining clerical staff will suffice. The other listed activities, including collection, quality control, and metadata are more evenly split between professional and clerical staff and could likely be allocated to GNP staff.

**Microfilming**

The DLG was also interested in how many of the institutions that currently collect born digital newspaper files are also simultaneously maintaining a microfilming operation, so we asked participants: “Are you also microfilming newspapers?” A majority of respondents (66%) do not currently microfilm newspapers. Since many institutions are adopting born digital collection workflows to replace costly microfilming operations, the number was unsurprising.

**Are you also microfilming newspapers?**



The survey included a follow-up question that asked respondents: “If you are microfilming newspapers, what content are you microfilming (newspaper titles not available digitally, etc.)?” The three institutions that are maintaining a microfilming operation all responded to the question and their responses are as follows:

* “Titles for which we have a paying customer (library) and select titles for which we do not acquire film or digital from a vendor or publisher.”
* “We had been microfilming both printed and born digital newspapers, but now film only printed newspapers.”
* “Our film lab sells microfilm to public libraries so film is made for all standing orders. Additionally, some publishers have not allowed newspapers to be published online. When this is the case these papers are filmed.”

These responses reflect a situation in which not all newspapers can be preserved and/or distributed through born digital collection workflows. The previous survey question regarding the percentage of titles collected magnifies the reality that adopting a born digital collection approach will inevitably leave newspaper issues for a large number of titles uncollected and unpreserved. Each institution will have to make a decision whether to continue microfilming those uncollected titles, consider taking an alternate preservation approach (e.g. digitization), or leave those titles uncollected and possibly unpreserved.

**Publisher Embargos**

The survey also included a question related to publisher embargos. We asked participants, “When securing permission from publishers to present their titles online, what embargo periods do you offer them?” We gave participants the option of selecting from choices ranging from “No embargo” to “24 months” with an “Other” option for customized responses.

The answers to the question do not point to a standardized embargo practice nationwide. Most of the survey participants selected the “Other” option to explain their various approaches to offering embargo periods to publishers. Generally, collecting institutions provide multiple options for publisher embargos. The main conclusion to be derived from the respondents that offer embargos is that participating newspaper publishers are comfortable with allowing their materials to be made available online within six months of publication, if not sooner.

The DLG followed up with survey participants to request samples of their publisher memorandum of understanding/agreement forms in an attempt to determine the considerations involved in securing permission to distribute born digital files online. Most of the permission forms are simple one-page agreements that share the following attributes:

* Non-exclusive permission to distribute and preserve publisher files
* Restrictions on commercial distribution of the materials
* Unrestricted use for research and educational purposes
* Perpetual distribution rights

Surprisingly, a preponderance of the submitted forms do include embargo options for the copyright holder to select from. Those embargo agreements, if they are made, are done so outside of the memorandum of understanding.

**Online Distribution**

The DLG was also interested in understanding the systems utilized to host born digital newspaper content. We asked respondents “What system do you utilize to host your born digital newspaper content?” and they were provided with eight options of commonly used newspaper distribution platforms along with a blank box “Other” option. Interestingly, no two participants listed the same system for distribution. The responses received by the DLG included:

* Internet Archive (Blacklight front end)
* Solphal - a homegrown system
* Just a simple on site file share
* Open ONI
* Islandora
* Portal to Texas History
* Sobek
* Veridian

Although there is no standard approach to distributing born digital newspaper images online, the results were promising in that there are a wide variety of platforms that are compatible with files created by newspaper publishers. This is an encouraging result for institutions that have a pre-established system for distributing digitized newspapers and hope to seamlessly fit born digital newspaper files into the same platform.

In addition to the platform used to distribute these materials, the DLG was interested in the circumstances under which they are made available. To that point, we asked participants, “Are the born digital newspapers you present online freely available, available on-site, or behind a paywall?” A majority of the respondents make their materials freely available online (66%), while 33% of respondents make materials available onsite, and a single respondent (11%) makes materials available behind a paywall. Two of the organizations make their materials available in multiple ways due to the nature of their distribution agreements.

**Are the born digital newspapers you present online**

**freely available, available on-site, or behind a paywall?**



**Conclusion**

Statewide collection and preservation of born digital newspaper files in the United States is still a burgeoning practice. There are far more organizations interested in adopting the practice than there are organizations that have already done so. That being said, early adopters of born digital newspaper workflows have laid a valuable foundation of practices and procedures for others to follow. Moreover, those practicing institutions are exceedingly generous in sharing their expertise on the subject.

The survey itself reveals that due to differences in circumstances, there is no standardized approach to born digital newspaper file collection in the United States. Practices and outcomes vary due to a variety of factors, including funding availability, publisher receptiveness, and most significantly, the presence of an aggregating organization or service. These situations complicate the landscape for organizations interested in adopting born digital newspaper collection workflows. The survey demonstrates, however, that despite the obstacles, many institutions have successfully transitioned. For those interested in the following their path, the promise of an affordable and technologically sound alternative to microfilming makes the complex landscape worth navigating in order to preserve and make available ever-popular newspaper materials.

1. Widmer, Lois J., Laurie N. Taylor, and Mark V. Sullivan. “Born Digital Newspaper Preservation Workflows for the Florida Digital Newspaper Library and the Caribbean Newspaper Digital Library,” [n.d.], https://ufdcimages.uflib.ufl.edu/AA/00/01/71/21/00001/LJWidmerLNTaylorMVSullivan\_FDNL.pdf. [↑](#footnote-ref-1)
2. Krahmer, Ana, Mark Phillips and Ron Larson. “Two states provide models for newspaper preservation, Newspaper Research Journal, 2015, Vol. 36(3) 328–336. [↑](#footnote-ref-2)
3. Terry, Kopana and Eric Weig. “Preserving Kentucky’s Newspapers: Analogue Beginnings to Digital Frontier.” Paper presented at IFLA WLIC 2015. https://www.ifla.org/files/assets/newspapers/Pretoria/pdf-000-terry-en.pdf. [↑](#footnote-ref-3)
4. Rabun, Sheila. “Microfilm to Born-Digital for Current Newspapers: A Case Study from the University of Oregon Libraries.” Paper presented at IFLA WLIC 2016. http://library.ifla.org/2064/1/S21-2016-rabun-en.pdf. [↑](#footnote-ref-4)
5. Zarndt, Frederick, Dorothy Carner, and Edward McCain. “Born Digital Legal Deposit Policies and Practices.” Paper presented at IFLA WLIC 2017. http://library.ifla.org/1905/1/S18-2017-zarndt-en.pdf [↑](#footnote-ref-5)
6. Salamon, Jennifer. “Born-Digital News in Ohio: What Is (and Isn’t) Happening.” Paper presented at IFLA WLIC 2016. http://library.ifla.org/1520/1/090-salamon-en.pdf. [↑](#footnote-ref-6)
7. Skinner, Katherine and Matt Schultz, “Guidelines for Digital Newspaper Preservation Readiness,”

Educopia Institute, 2014. https://educopia.org/guidelines-for-digital-newspaper-preservation-readiness. [↑](#footnote-ref-7)