2018 IFLA International News Media Conference

"When Risk becomes Real, Preserving News becomes critical"
18-20 April 2018 - George A. Smathers Libraries, University of Florida, Gainesville, Florida

Preserving the full historic record in holistic newspaper digital preservation: From the backend to the front

Virginia A. Dressler
Digital Projects Librarian, Kent State University, Kent, OH, USA

Abstract:
The paper will address two issues present in digital newspaper collections, namely that of long-term storage and access solutions specific to this medium, and takedown requests. The paper will outline the storage and hosting solution in place at Kent State University (KSU) of the student newspaper. The current digital archive contains ninety years of the printed newspaper (1926-2016), both reflecting both digitized and born-digital content. Decisions will be highlighted on the files retained on the dark archive (pod0), and conversely which files are stored on the hosting platform as access files.

Then, results will be presented from a recent research which investigated how practitioners from the Association of Research Libraries (ARL) member institutions address the notion of takedown requests in digital newspaper collections, with a brief conversation on how the concept of 'Right to be Forgotten' can be considered. A survey was conducted in the spring of 2017 that included a series of demographic questions in conjunction with some hypothetical and real-life scenarios revolving around takedown requests in digital newspaper collections. As the survey results show, there is a wide array of how takedown requests are currently handled and considered. The survey exposes a gap in practice and ultimately leads to some alarming outcomes in the differences of varied practice. Finally, a discussion of some examples of takedown requests at KSU will be outlined, and the subsequent internal discussions that led to a defining a process to consistently handle such requests.

Newspapers are a valuable asset to researchers, and decisions for takedown requests should be made more transparent in the age of 'open' digital collections and also lead to ramifications on full-text searching. Holistic newspaper digital preservation needs to address both the care of the raw digital files as well as the integrity of the printed word.

Keywords: Digital preservation, Takedown requests, Digital newspaper collections, Right to Be Forgotten

Introduction

This paper will focus on two aspects present in digital newspaper collections, the long-term preservation and care of the digital files and the concept of maintaining a faithful representation of the printed media. In July 2017, our institution completed the first phase of the large-scale digital initiative to provide a fully indexed and searchable digital archive of the student newspaper, *The Daily Kent Stater*[1]. Current practices will be discussed that revolve around the retention around the 'master' digital files associated with the project. Then, a discussion on the topic of maintaining the historical record in regards to takedown requests and other inquires for removal of information in sharing the results of a recent survey. Finally, an outline of subsequent conversations and decisions at University Libraries surrounding the topic will be shared to provide an insight into one example of a process and workflow to address such requests more consistently.

Storage vs. Access

At Kent State, our team has devised a method address digital preservation of the core files related to the digital student newspaper project. This entails retaining the full resolution image files to a dark archive (called the Pod0), in addition to the METS/ALTO xml. A brief overview of the process and overall production workflow of the digital initiative will be outlined here.

With the exception of the years 1926-1939 and the second half of 2016, high-resolution scans were captured from the original print version of the daily student newspaper. The newspapers had been bound into larger volumes, which were disbound for the purpose of scanning[2]. The first fourteen years of the publication (1926-1939) were scanned from microfilm since the original newsprint from these years is in very poor condition. Beginning in the summer of 2016, Kent State University Libraries worked with the Office of Student Media to obtain the original publisher digital files instead of scanning from the newsprint, effectively removing the digital capture portion (and expense) of the project. The next phase of the project will address the content published online, slated to begin in the summer of 2018 with the Office of Student Media and the University Archivist.

During the course of the digitization, encoding, and hosting of the associated digital files, KSU routinely worked with three separate vendors to complete this work. For the majority of the digitization project, Backstage Library Works acted as the first step in the process and captured the raw master image files of each page to the provided benchmarks (400 dpi, either grayscale or color depending on the original print, in uncompressed tiff images). Backstage also created some basic metadata for each issue, taken from the masthead of the newspaper (title, date, issue number) and followed a predesignated file-naming schema.

The tiff image files and original print were returned to Kent State University Libraries for quality assurance (QA) and review before sending on the encoding portion of the project. To date, this work has been completed by the second vendor, Digital Data Divide (DDD). Minimal post-processing was then run on the image files, mainly to address contrast to improve legibility,

[1] Digital Daily Kent Stater, https://dks.library.kent.edu/ Accessed January 30, 2018.

[2] I will note here that at least two (or more) copies of the student newspaper are held by Special Collections and Archives, so the decision to disbind one set of the student newspapers meant that there was always one set to remain in closed storage. Disbound originals returned from the vendor are now stored in flat phase boxes.

cropping and deskewing. The vendor would, in turn, create METS/ALTO xml files with article level segmentation, as well as jp2 images for each page and an issue level PDF file.

The tiff image files were then run through ABBYY FineReader software, with the vendor guaranteeing 95% (or higher) accuracy. Additionally, a METS record with MODS metadata at the article level was created. Upon agreement with the vendor, the following guidelines were provided as significant errors to be addressed and fixed: Incorrect characters (example of the letter 'p' for 'd'); Transposed characters (example of "teh" for 'the'); Missing characters (example of 'tht' for 'that') and Insertered characters (example of "c a t" for "cat"). Conversely, the following errors are considered insignificant: Capitalization differences (example of "gOLd" for "Gold"); Padded extra spacing (more than one blank space when original has only one) and extra line breaks. Finally, these would be returned for review to Kent State University Libraries staff for another round of QA and review.

Lastly, access files were transmitted to the hosting vendor, Veridian, for publication online, which are the following: jp2 images of each page, METS/ALTO xml files, and issue-level PDF files. Next, these files are uploaded and added to a test site for review, and once approved, added to the digital archive, accessible at https://dks.library.kent.edu/.

Dark Archive

After the work from the first vendor has been completed and approved, the raw tiff image files are moved to the dark archive storage system, named 'pod0'. The pod is arranged structurally by projects and subprojects, and in the example of the student newspaper, organized under the title of the newspaper as the main 'project' and then loaded into the system by academic year as the 'subproject'. Once the encoding has been completed by DDD, the METS/ALTO xml files are also ingested to these subprojects for long-term storage. The pod0 is the location for all of the master digital files and uncompressed data from projects in Special Collections and Archives, as well as all the born digital high-resolution image files from the University Photographer and selected resource files from University Communications and Marketing department. Table 1 displays the file types ingested into the dark archive, as well as the files sent to the hosting vendor.

	Image	**Text**
Long-term storage	400 dpi tiff (8-bit grayscale or 24-bit RGB)	METS/ALTO xml, METS/MODS xml
Access	Jpeg2000	METS/ALTO xml and PDF

Table 1. Breakdown of long-term vs. access files in the Digital Daily Kent Stater project

Files submitted to pod0 must adhere to the list of allowed archival file formats and also follow internal guidelines set and approved by the University Archivist. The system also incorporates record retention schedules[3] in place by the office of General Counsel, which are selected at the

[3] More information on the record retention schedule in place at Kent State University can be found here: https://www.kent.edu/generalcounsel/record-retention-schedule. Changes to record retention schedule in

project level upon creation. The retention schedule can be updated at any time as well if needed. The guidelines and protocols in place at the system work to assure adherence to the recommended file formats and overarching digital preservation standards set in place by the structure and set-up of pod0.

A digital preservation protocol was developed in 2016 to outline the practices, systems and required actions are in place to ensure proper digital preservation mandates[4]. This document covers topics such as redundancy, backup procedures, disaster recovery, file fixity, recommended file formats for long-term preservation and audit processes. Currently, the dark archive has two physical servers at different locations on the university campus with identical, redundant content on SATA hard drives, as well as one copy on a cloud location. An update to the protocol was made in January 2018 to reflect the update of practices with the removal of a third physical server and implementation of Amazon Web Services (AWS) Glacier. The associated software to support the storage includes a custom PHP/MySQL application, CentOS 6, Apache (httpd) web service, and a MySQL database, all designed and supported in-house by library developers.

The KSU library staff and faculty working the closest with digital media felt that the development (and implementation) of the digital preservation protocol was crucial to have in place to ensure the long-term access and use of digital files, and are also aware that this document needs to be adaptable to emerging ideas and new technologies that may impact future migration and framework to such a dark archive. The document is reviewed and updated by the University Archivist, the Head of Systems and the Digital Projects Librarian.

Two of the file types used by Veridian (jp2 and PDF) for access are not retained on the dark archive with the thought that these could be generated from the raw tiff image file if either of these need to be recreated for any reason, and the decision also reflects the practical aspect that retaining all variations of the content would result in a much higher storage footprint in the dark archive. Also to note, we have implemented a crowdsourced optical character recognition (OCR) tool, provided by the hosting vendor[5]. All versions of the METS/ALTO xml files are retained in the event that the content needs to be re-set to the original METS/ALTO file.

Maintaining the historical record

To change gears from the more technical aspects involved in the digital preservation of the project to the more ethical, we will now address the notions of maintaining the historical record by way of retaining a faithful, accurate digital version of the printed word and the potential

place (effective July 1, 2017): https://www.kent.edu/sites/default/files/file/RetentionSchedule%20Changes.pdf

[4] The decision to call the digital preservation document a 'protocol' as opposed to the more commonly used 'policy' was under recommendation from the Office of General Counsel, as the application of the word policy would mandate a different level of review and approval at the Board of Trustee level.

[5] More information about the crowd sourced OCR correction can be found here: https://www.library.kent.edu/special-collections-and-archives/digital-daily-kent-stater-text-correction. Monthly logs are sent from Veridian for review to ensure information is not tampered with maliciously in the digital archive.

impact of the takedown request in digital newspaper collections. I will argue that this element is as important as the digital preservation component addressed in the section above, though problematic when notions surround 'The Right to Be Forgotten' are considered[6]. The next section will reference a recent study that looked at how surveyed academic libraries deal with takedown requests in digital newspaper collections, and outline some key issues for practitioners to consider when receiving such requests.

Takedown requests: A survey

In the spring of 2017, 124 ARL member institutions were surveyed on the topic of takedown requests of content present in digital newspaper collections. A series of hypothetical questions were posed, and also asked respondents to share relevant policies in place at their institution as well as any real-life scenarios of takedown requests[7]. A brief overview of survey results will be provided below, with a more comprehensive report provided in an accepted article in College and Research Libraries (Anticipated publication in January 2019, with a preprint currently available on the journal website).

All of the ARL member institutions were invited to take the survey, in a Qualtrics invitation addressed to identified digital librarians or a job title that indicated some level of involvement with digital collections (with a last resort of member of library leadership if no person could be identified from publicly availble staff directories). The survey received a 25.8% response rate over the course of one month (February to March 2017), and was provided in both English and French. The first half of the survey focused around some basic demographic information about the general nature of digital collections at the institution, how long the digital collections have been in place and other questions addressing staffing, platform, and accessibility of digital collections. Additionally, survey respondents were asked if there is a policy (or policies) currently in place to address takedown requests and were also asked optionally to share the policy if they were able directly into the Qualtrics survey or provide a URL to the content.

The second half of the survey centered around a set of hypothetical questions surrounding variations of takedown requests around a digital newspaper collection (Figure 1). Additionally, a question was included to ask what person or person(s) at the institution would be charged with decision making around takedown requests. And lastly, survey respondents were asked if they

[6] In a *very brief* nutshell, the Right to be Forgotten revolves around a request from a Spanish citizen in 2010 regarding a newspaper article from 1998 (*La Vanguardia*), with information that the individual felt was no longer relevant and requested removal from first the original publisher, but was later granted to remove the information from search engine results. This decision has had a huge impact in the European Union, and, since 2014, there is a mechanism in place for individuals to request removal of private information directly to Google. More information can be found at Transparency Report, https://transparencyreport.google.com. To date, the United States has not had this kind of mandate or data privacy laws, with the exception of a similar court case in 2008 regarding a Cornell alum and the student newspaper (Vanginderen v. Cornell, Case No. 08cv736 BTM(JMA). United State District Court, S.D. California. Jan. 6, 2009).

[7] Virginia Dressler and Cindy Kristof. "The Right to be Forgotten and Implications on Digital Collections: A Survey of ARL Member Institutions on Practice and Policy". College and Research Libraries Journal (expected publication, January 2019). Preprint available: https://crl.acrl.org/index.php/crl/article/view/16890. KSU Institutional Review Board (IRB) #17-059.

were willing to share a real-life scenario along the lines of the hypothetical scenarios provided in the survey.

> You receive a request for a name to be removed from a particular item in your digital library, directly from the individual in question. The requestor claims that the inclusion of their name in an openly accessible digital library violates their privacy. The name appears in print in your digital regional newspaper collection, within the student newspaper that was published in print at your institution and later digitized for the digital collections. This content has been run through optical character recognition (OCR) software, and has been fully indexed by search engines such as Google. How would you respond?

Figure 1. Hypothetical question #1 from the survey

Some selected responses to the hypothetical question include: “I would check with [General Counsel], but would assume that no change would be required- we are merely providing access to an already existing item and would not want to modify the historical record.”, and three other respondents indicated a dialogue with General Counsel would take place initially as a result of the request. Other answers provided thoughtful responses regarding the logistics and long-term complications of takedown requests, as in this answer: “Administratively, keeping track of what’s been redacted has been troublesome.”

Three respondents indicated that the item in question would be temporarily removed from public access while the issue is being addressed and resolved. Another handful of respondents stated that they would most likely remove the name upon request, often not needing more reason than the initial request from the individual in question. This removal included the full redaction of the name from the image and OCR, as well as the more subtle; “We would maintain the digital representation of the newspaper while removing the name from the OCR text file to prevent crawlers from indexing the name and making it easily discovered.” And lastly, one respondent indicated that they would use the request as a chance to educate the requestor, “We would discuss their reasons and explain it’s a news source and we can’t change it. It would be unethical to alter the news from the past. If they claim the article is defamatory, we would refer them to University Council.” The answers provided to the hypothetical questions in the survey displayed a huge array of how practitioners consider these requests and ultimately act on the takedown request.

The other interesting aspect that can be taken from the results was the personnel involved with decision-making around takedown requests. Some respondents noted that this would likely be addressed by a single staff member (normally the person in charge of managing the digital collection most directly). Others indicated that the more complex issues be taken back to a working group of many individuals at the institution and could include a department head, copyright librarian, scholarly communications librarian or other staff who may be able to help form a consensus on a decision for the request. And again, some institutions cited using someone from a General Counsel office or equivalent for consultation and the ultimate decision.

And finally, the real-life scenarios were intriguing to get some direct insights into how decisions are made around the more problematic requests. One respondent cited an oral history collection, where the interviewee had talked about another individual in the interview, whose family ultimately felt the comments were slanderous. Since the institution was currently courting the latter's family for a donation, the request to pull the item offline was honored in hope of a potential collection. Another cited some cultural sensitivities in a Native American collection, and did weigh issues of privacy heavily in this scenario more so than other projects. And finally, one institution indicated they had redacted social security numbers were had been published within a digital collection. There can be some interesting considerations for practicioners to keep in mind, and in some instances, may find scenarios where the takedown is a valid decision that is made as an institution. Overall, there is a need for more discussion and documentation of these issues as a profession.

Role of Policy

Another aspect also addressed in the survey was the role of policy at the institution regarding takedown requests. Seven respondents indicated a policy in place, though after further analysis only three of these really addressed the issue and processes involved in a takedown request of information contained in a digital collection. The remaining policies centered around copyright isssues, and did not specifically request the notion of takedown besides copyright violations. The more helpful policies addressed the process of the request, the required information (Title, URL, issue) for the requestor to submit, and the expected turnaround time for a decision. A well-written policy can be a place to provide a useful framework for both the institution and the requestor to set guidelines and expectations for the practice of handling such requests.

Impact on local practice

Since the survey, Kent State University Libraries has addressed the current internal processes involved in takedown requests and have also created a working group to address such requests, in part to improve awareness and communication of further requests. We have also formulated the following response to takedown requests of information contained in the digital student newspaper collection specifically.

The individual requesting removal of any information from the digital collection will be directed to:

1. Make a public records request
2. Submit documentation as proof of inaccuracy to Office of Student Media to make a decision on whether or not to redact
3. University Libraries can then correct or remove information from digital archive, if required

The KSU working group feels that putting in some standard language on the website (not currently in place), and setting out a list of steps to relay to the requestor will move towards a more systematic process and internal workflow surrounding takedown requests.

Conclusion

In conclusion, a holistic approach to preserving the printed word necessitates consideration of both the digital preservation underpinnings for long-term storage and access as well as addressing internal practices regarding takedown requests. I believe well-formed policies addressing both digital preservation and takedown requests, when combined with consistent practice will aid in a more transparent management of digital collections, can assist current (and future) practitioners complete these endeavors. From the survey results, one can note the inconsistent practices when takedown requests are made to digital collections, and these decisions can have long-term impacts on the notion of truly open and accessible digital collections.

References

Backstage Library Works, accessed February 8, 2018. http://www.bslw.com/.

Digital Daily Kent Stater, accessed January 30, 2018. http://dks.library.kent.edu.

Digital Divide Data, accessed February 8, 2018. https://www.digitaldividedata.com/.

Dressler, V. and C. Kristof. “The Right to be Forgotten and Implications on Digital Collections: A Survey of ARL Member Institutions on Practice and Policy”. *College and Research Libraries Journal* (expected publication, January 2019). Preprint available: https://crl.acrl.org/index.php/crl/article/view/16890.

Kent State University Libraries, Digital Preservation Protocol (version 2). 2018.

Veridian, accessed February 3, 2018. https://www.veridiansoftware.com/.