Digitization in the Real World

Lessons Learned from Small and Medium-Sized Digitization Projects

Edited by
Kwong Bor Ng & Jason Kucsma

Metropolitan New York Library Council

Published in the United States of America by
Metropolitan New York Library Council
57 East 11th Street, 4th floor
New York, NY 10003-4605
p: (212) 228-2320 f: (212) 228-2598

Web site: http://www.metro.org

ISBN: 978-0-615-379998-2

Cover Design: Jason Kucsma *(illustration by Smartone Design, licensed via iStockphoto.com)*

Reviewers Committee: Mark F. Anderson, Jill Annitto, Anna Craft, Jody DeRidder, Renate Evers, Wei Fang, Maureen M. Knapp, Sue Kunda, Mandy Mastrovita, Ken Middleton, Emily Pfotenhauer, Mark Phillipson, Alice Platt, Mary Z. Rose, Stacy Schiff, Jennifer Weintraub, Andrew Weiss.

Digitization of the Yale Daily News Historical Archive

Kathleen Bauer, Ian Bogus, Karen Kupiec
(Yale University Library)

Jennifer Weintraub (UCLA Library)

Abstract

The Yale Daily News is Yale University's independent student run newspaper. Founded in 1878 it is the oldest continuously published daily newspaper at a United States university. From the initial print volumes until digital versions started in 2000, the entire run of the printed paper consists of 122 volumes and approximately 100,000 pages. In 2007, Yale University Library was asked to create a pilot project to digitize and make available an initial set of ten years of the newspaper with a $50,000 start-up budget. In this article, we will discuss how the project began, and issues that developed during the process related to copyright, interface design, workflow, quality control, and fundraising. This project helped Yale University Library, a large, strongly hierarchical institution, to develop workflows that allow its staff to develop new skills and work across traditional departmental boundaries. Library staff that have traditionally performed tasks related to our print collections or for smaller digital projects have developed new skills and methods for workflow, metadata creation and quality control for a large-scale digital project.

Keywords: Campus newspaper, Copyright, Newspaper digitization, METS, Quality control, Yale, Scanning.

A newspaper digitization project is one that every library, public or academic, can undertake. It is often not hard to get the rights to a local or small paper and an academic or public library has a built-in audience for this type of project. Local researchers will love having it online and genealogists from further afield will bless you. And yet, newspaper digitization, while having recently come into its own, has been somewhat difficult for libraries. Newspapers are crucial to research, providing detailed local and international accounts of events; these incredibly important primary source materials are made of poor quality material that will last a relatively short period of time. Newspapers are hard to digitize because they are published daily with hundreds of issues a year, comprised of various individual sections, and then individual articles, oversized, delicate, and contain thousands of words and pictures that require careful quality assurance. In addition they have unusual layouts, and articles often are split across two or more nonconsecutive pages. There can be numerous contributing authors, syndicated cartoons, advertisements, supplements, and even the occasional joke issue.

Fortunately, newspaper digitization is not new. Many organizations have taken on newspaper digitization and the major national and regional newspapers are now available for licensing by libraries. While many projects focus on digitization of newspapers from microfilm, there is also an increasing number of digitization projects that begin with the original paper. One important clearinghouse for information and best practices for newspaper digitization is the National Newspaper Digitization Program (The Library of Congress, 2009). This program, a joint effort between the Library of Congress and the National Endowment for the Humanities uses the power of grant dollars to enable proper newspaper digitization, research in newspaper digitization and access to the digitized papers through a central resource.

The Yale Daily News, Yale University's student run newspaper, is 132 years old and is the oldest continuously published daily newspaper at a United States university. In 2007, Yale University Library (YUL) was asked to create a pilot project to digitize the newspaper with an initial $50,000 start-up budget provided by the

Yale Daily News's parent foundation (the Oldest College Daily Foundation) and YUL. In this article, we will discuss how the project began, and issues that developed during the process related to copyright, interface design, workflow, quality control, and fundraising.

This project helped YUL, a large, strongly hierarchical institution, to develop workflows that allow its staff to develop new skills and work across traditional departmental boundaries. Staff across the Library who have traditionally performed tasks related to our print collections or for smaller digital projects have developed new skills and methods for workflow, metadata creation and quality control for a large-scale digital project.

The Yale Daily News (YDN) is staffed and produced by student volunteers. The paper is not owned by Yale University, and the student reporters and editors are advised by the independent Oldest College Daily Foundation (OCD). OCD is comprised of former YDN staffers and Yale graduates. In 2005 the OCD came to the YUL with an idea for a project to digitize the Yale Daily News archive and provide access on the Internet. The OCD realized the complexity of the proposal especially considering they did not own a complete run of the newspaper. They asked the YUL to partner with them as OCD owned the rights to the content while YUL had the expertise and the means to make it accessible. To start the pilot project, OCD and YUL contributed $25,000 each to finance a pilot project. YUL decided that for the pilot project we would not digitize anything for which there was an existing digital edition (the YDN has been available online since 2000). Thus, we still had to choose a small amount of material from 120 years of print issues, or a fraction of the 100,000 pages in the entire run, for our initial digitization pilot.

This type of partnership, between an external group owning copyright and the campus library, can be useful for both parties. It is a good way for library staff to gain experience with a complex digitization project and digital collection building, it provides useful material for fundraising for technology projects, it enables the library to provide a useful resource to the campus community, and it enables

both the newspaper and the library to create an online product with research value freely for a product that may not have a large sales market.

Several basic principals helped guide the development of the Yale Daily News Historical Archive. Open or commonly used standards for our digital files were important in the event content needed to be migrated to new interface software in the future. We wanted to digitize each newspaper in its entirety, thereby preserving the historical context provided by editorials, cartoons and advertisements. Therefore it was important that we capture the images on each page, not only the text. Another key principle was our requirement that the Yale Daily News be freely available on the Internet. Finally, we wanted the newspaper to be fully searchable, browsable, and to include advanced search features such as byline and title searches. These principals are similar to those elucidated by NDNP and other newspaper digitization projects.

The Initial Decisions

Our first decisions concerned how we would scan the images and what the quality needed to be. Scanning from original source materials will almost always provide the truest digital image but depending on the format of the source there may be undue complications. The physical copies of the Yale Daily News held in the Library's Manuscripts and Archives department are tightly bound which hinders how well the volumes can be opened. Tightly bound materials pose a few risks for digitization projects. Not only can they be damaged during scanning but they can also inhibit producing quality digital images. Microfilm is easy and inexpensive to digitize but by nature is already a derivative format; it will always appear as a black and white photograph of an original document. Microfilm can also be of poor quality, out of focus, smeared, scratched, or otherwise unreadable. Pictures that have been microfilmed often have lost much of their detail.

Luckily, the Yale Club of New York had a complete copy of the Yale Daily News and was willing to donate it to the project. This third copy allowed us to avoid the difficult decision of inexpensively

scanning microfilm, scanning bound volumes at a high cost, or disbinding our only physical copy, which would balance the cost and the quality but would leave us with individual leaves to box and store. We were able to disbind, scan and discard these volumes without compromising library collections.

The question of how much storage space would be required for the thousands of images we would receive impacted a number of basic technical decisions. Bitonal scanning would not represent the photographs adequately so grayscale or color images would be required. We felt confident that JPG2000 as a reliable file format for these kinds of materials. Though the Yale Daily News is an essential part of research at Yale, we were not undertaking a scanning project to replace the print and microfilm versions of the YDN, but merely to provide access. JPG2000 allows large amounts of information to be stored in a compressed form, without the compression artifacts and other data losses inherent to JPEGs.

Other universities digitizing complex text and newspapers were using METS (The Library of Congress, 2010b) along with ALTO (The Library of Congress, 2010a). The METS files enable the program to understand the structure of a document, such as a newspaper issue with 16 pages. These files were considered "required elements" for our projects, as they would enable us to search within the issues. The ALTO files provide technical metadata for optical character recognition. The combination of the METS files which describe the issue of the newspaper with the ALTO files which describe the layout, enable full text searching and the highlighting of search words within the display of text. The user sees a representative image (a digital photograph) of a page. The searchable text created from OCR is a stored in a different file, in essence creating a layered document. The METS/ALTO links these two such that when a portion or zone of the image is highlighted it is associated with the text it represents. Users can then search for phrases and see the matching term highlighted on the page image. They can also select and copy the text directly from the image.

Yale was not in a position to write software to provide access to METS/ALTO files for their full functionality. Instead, we decided use CONTENTdm, a commercial product that would suit our needs and fully utilize the METS/ALTO structure. In order to streamline the process as much as possible, we decided to employ vendors to provide as many of these services for us as much as possible. The creation of the METS/ALTO files were included in our request for proposal (RFP) as "Highly Desirable Elements" allowing us to see potential solutions our scanning vendors could perform without totally rejecting a proposal if a vendor could not provide them.

In choosing the issues to digitize it may have seemed obvious to start with the first issue. However, since the OCD and the Library were going to need to fundraise to continue the project, both the Library and the OCD felt it would be best to do a range of interesting time periods to generate interest with potential donors. In addition, an interesting grant funded project that was already underway was digitization of World War 1 posters and pamphlets held in the YUL collection. We decided to tie into that project and digitize issues from 1913-1919. In consultation with Manuscripts and Archives it was discovered that archivists there use the Yale Daily News frequently in answering reference questions and particularly beginning when A. Bartlett Giamatti was President of Yale. The records of Yale's President, and many university offices, were closed to researchers for thirty-five years. The YDN thus provides the best available material for historical research on Yale activities between 1978 and 1981. Finally, we decided to digitize a very exciting time period that is always in demand amongst students, 1967-1970. During this time Yale and New Haven experienced student protests, a murder and resulting controversial trial of a member of the Black Panthers, and a move toward co-education.

The pilot batch consisting of 800 issues from 13 years, or approximately 8000 pages. We were fortunate to be included in a gift from the Yale Class of 1945W, which is the class of students from 1939-1945 whose initial Yale education was interrupted by World War II and completed their degree in an accelerated program upon

returning. This provided another 6 years of material bringing the total number of pages to 24,000.

Selection of Vendor and System

In July 2006 we sent the RFP to four vendors and received responses. Most of the vendors were able to fulfill our required and desired elements using similar technology. The RFP process and careful evaluation of the samples supplied by the vendors was illuminating. Through this process, we were able to learn more about the way different vendors work and we had a chance to test drive the solutions to our problem and evaluate different scanning techniques.

All of the vendors used DocWorks software to OCR the files, zone the articles, and create the METS/ALTO files. Then they prepared the package of files for loading into CONTENTdm, using a special loader developed by CCS, the company that created DocWorks. We were able to provide access to the newspaper by using CONTENTdm. This software package, used by many academic libraries, enables users to do full text searching on an issue or across all of our newspapers. The functionality not only allows viewing images of each page but also permits the user to click on an article to view the complete article by itself. The recognized text is also accessible, useful for cutting and pasting. CONTENTdm can automatically create PDFs of the issue on the fly for printing. Finally, because of the zoning and the ALTO files, when an article goes on to another page, CONTENTdm pulls all of the parts of an article together into one screen, making for easy reading

We chose to work with Digital Divide Data, a nonprofit company based in New York City. Digital Divide Data employs young people in third world countries. Their employees not only learn IT skills but also attend classes part time in an effort to break the cycle of poverty.

Digital Divide Data gave us several options for scanning the newspapers. Because they are old, the newspapers had a yellowish tinge. We chose to scan the newspapers in grayscale, which provides a smaller file size and quicker loading of CONTENTdm. We also chose to process the text so that the background and text have a high contrast, while leaving the images in grayscale. The result is easy to

read, easy and quick to print, and true to the original intent of newsprint

Workflow

For strategic reasons, YUL does not scan the newspaper issues chronologically, making tracking the project complicated but a high priority so that the various participants could at any time see the status of the project and if any part was held up. A database was created for each volume of the newspaper. Fields were added including a pull down list for the status of each volume. Because various personnel throughout the library were working on the project this database helps everyone know what the status is of every volume.

When volumes have funding designated the status is changed so staff knows to prepare those volumes next. The volumes are collated and a manifest is created that includes basic information such as the volume, number, date, and page count in each issue. Missing volumes, printing errors in the enumeration, and possible inhibitions to scanning are also recorded. The volumes are then disbound with the manifest in hand. Pages missing a significant amount of text are removed with the rest of the issue it belongs to. These issues are recorded in the database under "Issues Needing Replacement." Microfilm will be used to scan missing issues and in order to keep a visual consistency entire issues are scanned from the same sources. For efficiency sake, missing issues are gathered and microfilm is sent and scanned in batches.

Once a volume has been disbound the manifest is printed and tied with the volume. Preparing volumes is frequently performed in batches between shipments. This allows us to have a number of batches ready and waiting until the shipment is due making deadlines easier and less harried.

Regular shipments are usually sent to Brechin Group, a digitization vendor subcontracted by Digital Divide Data, at the beginning of each month. After Brechin scans the volumes, the digital files are then stored to hard drives and sent to Digital Divide Data's Cambodia office for processing where the images are zoned, metadata

is created, and the CONTENTdm files are created. Once processing is complete the hard drives are then sent back to Yale where the files are upload directly into our CONTENTdm test server for quality control.

Quality Control

Given the quantity of items in every shipment, it is not possible to check every page. Instead sample issues are checked from every batch returned by the vendor using the ANSI/ASQ Z1.4-2003 (American Society for Quality, 2003) standard quality control procedures. This standard clearly defines how many items need to be sampled, the acceptable error rate, and when to accept or reject a batch based on the number of errors discovered. This standard was developed based on, and is almost identical to, an old military standard (i.e., MIL-STD-105E, Department of Defense, 1989) for inspecting shipments.

This method is based on the idea that there is a level of error one is willing to tolerate but if the error rate is too high the entire batch will be rejected and reprocessed. The Yale Daily News Project identifies four separate areas that need to meet specified quality rates: Image Quality, Zoning, Headlines & Authors, and OCR. If a particular area fails the vendor only has to redo that particular area, not reprocess the entire batch from the beginning. This helps zero in on particular problems for our vendors, though it complicates how we select the samples. Ideally, the sample would be totally random, but this is not feasible considering the samples are based on the particular units being checked; so the image quality is based on each page as a unit, but the zoning error rate is based on each article as a unit. It isn't reasonable to line up all the pages, let alone the articles, and randomly sample them considering complete issues are loaded into CONTENTdm. To get around this, a system was created to convert article and pages into whole issues to be checked.

A quality control tool was created that helps staff through the process that is based off the ANSI standard. The manifests for each batch are added and the inspection level is chosen. The tool then determines how big the sample sizes are as well as the rejection thresholds. A random number generator selects the issues that will

comprise the sample. The sample issues are searched in the CONTENTdm test server where the batch was loaded. Errors are recorded in the tool as well as notes such that it can be easily found again if needed. Once complete the tool calculates the results and tells staff appropriate actions. Once a batch is approved it is moved from the test system to the live system.

Copy Right

When we started the Yale Daily News Historical Archive, the copyright situation seemed straightforward. Copyright of the YDN belongs to the Yale Daily News Publishing Co., Inc. This company is run by the student officers of the YDN, under advisement of a professional manager. The project was conducted in close collaboration with the officers of YDN and the Oldest College Daily Foundation, and we had their permission to digitize and make material freely available. Early in the project we concentrated on the earliest material from the 1880. This material was mainly text written by students, with some advertisements. There were no photographs. Subsequently material from World War I was digitized, but material was pre-1923 and again consisted mainly of text written by students, although a few photographs began to appear.

It was not until we started work on material from the 1960's that we grew concerned that we might have a problem. In the 1960's the YDN began to run comic strips, some of which were produced by Yale students. Most notably, a strip called Bull Tales first appeared in September 1968 written by an undergraduate named Garry Trudeau. After Mr. Trudeau graduated he changed the name of the strip to Doonesbury which was then syndicated nationally, including in the YDN. Peanuts also regularly ran in the YDN throughout the 1960's, 1970's and later years. The inclusion of these popular and copyright protected comic strips raised red flags for us: were we allowed to include this material, and did we need to seek permission to do so? We worried that we would need to excise this material from the digitized copies of the papers.

The comic strips were not unique material that could not be found elsewhere, but it did seem that at a time of social and political upheaval at Yale and the entire nation, the YDN staff included the strips for a purpose, and they did play a part in the tone of the YDN. Doonesbury was often overt political comment on current events: Peanuts less so, but still was part of a social commentary. The digitized YDN would not be complete without the inclusion of the strips as they first ran in the paper. We decided that a safe course of action would be to contact the rights holders, Trudeau and the estate of Charles Shultz. In both cases, permission was given to use this material.

We were lucky that this issue was a problem for other projects as well. In June 2008, at the same time we were granted permission by the authors to run both strips, the 11th circuit court of Appeals in Atlanta rendered a decision in the case of Greenberg v. National Geographic Society siding with the NGS. Greenberg had sued NGS for including in a digitized version of the magazine material written by Greenberg and originally published in the printed magazine. Greenberg claimed that NGS only had permission to use his material in the original publication, and in reusing his work in the digital version had violated his rights under Section 201c of the Copyright Act. In siding with NGS, the court found that the use of Greenberg's and other's work was permissible because the magazine was faithfully reproduced and presented as the original, with material presented in its original context. This was in contrast to the 2001 finding in Tasini v the New York Times Co., where the Supreme Court found that the rights of freelancers were violated because in the digital product in question individual articles could by viewed individually, without the original context of surrounding material. The difference between these decisions ultimately lays in how the digitized content is displayed. The publisher has the rights to the issue as a whole and can repackage it as a whole, but they cannot split out the parts and make them into a different product. In our presentation of the YDN, each issue is presented in its entirety as a faithful and full reproduction of the original. While searches will indicate individual articles (and occasionally comic strips) the user always finds that material in the

context of its original issue. This decision meant that although we did get permission from Garry Trudeau and Charles Schultz to use their strips it was not required that permission was specifically granted to include them.

Challenges involving Workflow and Data Correction

Anytime a new project starts, especially with a new vendor, there is a period of adjustment and problems are expected. Vendors work with numerous customers with various expectations. It is extremely important for customers to be as clear about their expected outcomes and requirements as possible. Good vendors will make every attempt to satisfy those requirements.

In earlier volumes of the Yale Daily News it was not common for authors to be named in the articles. During a batch of later years, when it became common to name authors, it was discovered that the authors' names did not display with other metadata in the header bar in the pop up window. This was recorded as an error as it was our understanding that fields tagged as "authors" would show up with the title of the article. In discussing this with the vendor they were in fact tagging them correctly and gave us a short list of issues to check in our CONTENTdm installation for why author names were not being displayed.

Other problems were found before the material left the library. As we started disbinding older volumes there was some damage to the pages. Examples of types of damage were photographed and sent to the project staff so decisions could be made on where the cut-off of acceptable versus unacceptable damage would be made. It was ultimately decided that some minor loss of text was acceptable as long as the user could still surmise the lost text and it was not in an area that may be in a targeted search, such as the title of the article.

Other challenges have come up during the course of the project but because of our good relationship with the vendor we can work out solutions easily.

Funding and Sustainability

The project was originally funded with contributions from the YUL and the Oldest College Daily Foundation. In addition the YUL matched its original funding amount for one additional year. At the same time the Library Development Office added the YDN digitization project to its priority list and began actively seeking contributions. The Yale Class of 1945W (the World War II years) signed on quickly to fund the eight academic years covering 1940-1948.

As we moved the project from the pilot phase into production we faced challenges brought about by the depressed economic climate of early 2009. This became the most important factor in determining how the project would continue. Decreases in the Library budget resulted in the elimination of additional library funds. Therefore, digitization of additional content is currently being funded entirely by donor contributions.

The Library's Development Office continues to prioritize this project on its development list and actively works with potential donors to identify time periods that may be of individual interest and are available for funding. Contributions are publicly acknowledged on the Library's website (http://images.library.yale.edu/digitalcollections/ydnAcknowledgments.aspx).

As we move forward issues for digitization will be selected based on a variety of criteria. Donor funding may be given for specific years. These years will be given the highest priority in the digitization queue. Next, if donors do not select specific years for digitization, priority will be given to years that are in demand by researchers at Yale. Specific years or eras are requested repeatedly – including the 1960's. Identification, digitization and availability via CONTENTdm of these specific years can improve services to researchers by providing them immediate access to the information they require while at the same time increasing staff productivity by avoiding repeated copying of the same articles. Once we have digitized all content identified as highly desirable by researchers we will fill in remaining year gaps beginning with the oldest content.

Conclusion

The Yale Daily News digitalization has been a challenging and rewarding project. It has been a partnership with organizations outside of the library, such as the OCD Foundation, and also a great opportunity within the library for various departments to work together. The core team within the library comes from five separate departments: Cataloging and Metadata Services, Electronic Collections, Library Access Integrated Services, Preservation, and Usability & Assessment. This group meets regularly and work out issues that each department have interests in finding collaborative solutions, many of which were able to move into other fledgling projects. In addition, Preservation department staff performs the tasks of preparing the newspaper for digitization and for checking the quality of the digital files as they return. This process utilized skills the department already had, expanding them for newspaper digitization.

On the fundraising side it has proven to be a great springboard. Yale graduates tend to be quite loyal and they have frequently been very interested in looking back into the digitized YDN content that has been created. Not only are they giving back to help complete the project but it has also been an eye catching project where other potential projects can be discussed and funded.

Finally, digitizing the Yale Daily News has enabled the library to produce a free, highly useful, and unique digitized resource for both Yale University and other researchers. The expertise gained in this project has enabled YUL staff to build on this success with other, more complex, digital projects. YUL staff can now successfully digitize varied material such as maps, annotated manuscripts and books. Mass digitization of books may free libraries to concentrate on high quality digitization of unique material that is present in nearly all public and academic libraries. A newspaper digitization project can be an excellent springboard for digitization of this diverse set of material.

References

American Society for Quality. (2003). *ANSI/ASQ Z1.4-2003: Sampling procedures and tables for inspection by attributes*. Milwaukee: American Society for Quality.

Department of Defense. (1989). *Military standard. Sampling procedures and tables for inspection by attributes*. Washington, D.C.: Department of Defense.

The Library of Congress. (2009). *National digital newspaper program*. Retrieve May 20, 2010 from http://www.loc.gov/ndnp/

The Library of Congress. (2010a). *ALTO: Technical metadata for optical character recognition*. Retrieve April 10, 2010 from http://www.loc.gov/standards/alto/.

The Library of Congress. (2010b). METS *Metadata encoding & transmission standard*. Retrieve April 10, 2010 from http://www.loc.gov/standards/mets/