Digital Migrations: A Case of Turn-of-the-Century Chicago-Immigrant Newspapers

William Schlaack

Preservation Services, University Library, University of Illinois at Urbana-Champaign, Urbana, USA.
E-mail address: schlaac2@illinois.edu

Anna Oates

Preservation Services, University Library, University of Illinois at Urbana-Champaign, Urbana, USA.
E-mail address: annaio2@illinois.edu

Abstract:

This paper will illustrate a case study on the digitization and preservation of non-English newspapers at a large research institution in the United States, providing an overview of challenges present throughout the newspaper digitization lifecycle while focusing on aspects unique to non-English newspaper titles. The case study looks at the importance and relevance of the content; evaluates complications found during collation, such as identifying supplemental editions or standard weekly supplements; intricacies of updating cataloging records to accurately provide descriptive access; actions toward the long-term preservation and sustainability of digital content and digital access portals; errors and solutions for non-English OCR; and limitations of findability imposed by search engines. To best demonstrate the IDNP's work towards access and sustainability, the paper will provide an overview of the U of I's digital newspaper portal, Illinois Digital Newspaper Collections, and digital preservation efforts with the in-house digital repository, Medusa. These practical workflows will showcase the innovative role IDNP plays in newspaper reformatting best practices.

Keywords: Newspaper Digitization, NDNP, Chronicling America, Digital Preservation.

1 INTRODUCTION

In 1987, Illinois State Historical Library received their first grant from the National Endowment for the Humanities (NEH) under the United States Newspaper Program (USNP) to foster the preservation and access of Illinois newspapers. The first step completed by the Illinois Newspaper Project (INP) was to survey cultural heritage institutions, capturing the breadth of newspaper collections across the state. This survey informed the cataloging and inventory work carried out from 1989–2007. The efforts to establish a comprehensive bibliographic record of Illinois newspapers provided justification for the selection of 471 Illinois newspapers that have been microfilmed under the project.

In 2005, the History, Philosophy and Newspaper Library (HPNL) at the University of Illinois at Urbana-Champaign (U of I) Library began work building a digitized newspaper collection. The Farm, Field, and Fireside; American Popular Entertainment; and Collegiate Chronicle projects are explored later in section 3.1. Based upon these experiences gained and success of Illinois's contributions to the USNP and previous microfilming and digitization efforts, the Illinois Digital Newspaper Project (IDNP), led by the U of I, received funding from the NEH to contribute to National Digital Newspaper Program (NDNP), a joint effort between the NEH and the Library of Congress (LC). In 2004, the NEH and the LC began this collaboration to digitize historically significant newspapers and make them freely available on the LC hosted digital library, Chronicling America. The NEH makes the NDNP possible by providing grants to U.S. states and territories. The grant awardee organizes a state-wide or territory-wide project to digitize newspapers from microfilm. Upon delivering the digitized newspapers to LC, which is responsible for their long-term storage and access.

In 2009, the IDNP became a state partner in the NDNP, and to date has contributed 336,864 pages to Chronicling America. During its first round of funding, the IDNP contributed 100,030 pages of daily publications significant to the urbanization, including *Chicago Eagle* (IL), *Broad Ax* (Chicago, IL), *Day Book* (Chicago, IL), and the *Cairo Evening Bulletin* (IL). The IDNP received two additional NDNP grants in 2011 and 2013 to continue the digitization of prominent newspapers, including continued digitization of the *Cairo Evening Bulletin*, in addition to new titles, *Rock Island Argus* (IL), the *Ottawa Free Trader* (IL), and the *Joliet Signal* (IL). These projects resulted 209,959 digitized newspaper pages. Toward the end of the 2013 grant, the IDNP applied for and received permission to begin work on a no-cost-extension to digitize non-English newspapers. This pilot project resulted in IDNP's first significant foray into the digitization of 9,122 pages of non-English content. Titles in this extension included Polish language newspapers *Telegraf* (Chicago, IL) and *Dziennik narodowy* (Chicago, IL), the multilingual Croatian and Slovenian *Proletarec* (Chicago, IL), and the German language *Vorbote* (Chicago, IL). Beginning in 2011, the NDNP technical guidelines provided measures to encode non-English languages with the inclusion of French, Italian, and Spanish. Since this time, NDNP has continuously expanded the support of non-English languages to include languages that are encoded under ISO 639-2.

Figure 1. Timeline of Illinois Digital Newspaper Project partnership with NDNP.
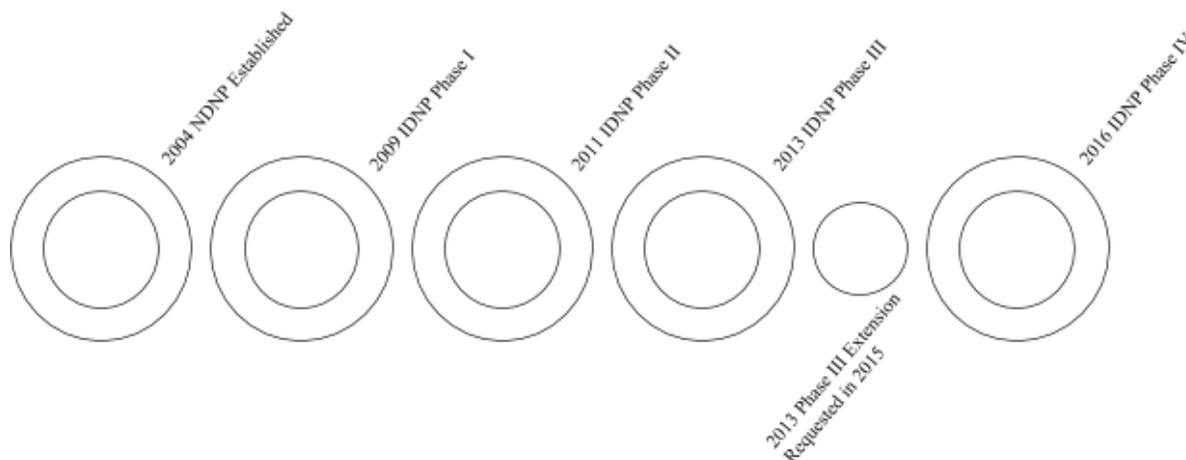
Throughout the evolution of the IDNP, contributing staff has frequently changed, resulting in the loss of institutional knowledge. Introducing new staff to the project proved particularly challenging with the digitization of non-English content as new staff were faced, not only with realizing the project objective, but also the difficulties of not possessing intimate knowledge of the languages. Understanding these challenges, the IDNP proposed to continue their contributions of unique content through NDNP to digitize 100,000 pages of non-English newspapers, focusing on essential Eastern-European immigrant newspapers published in Chicago. In 2016, the NEH funded IDNP's fourth two-year grant cycle—internally referred to as Phase IV—to digitize Czech, Lithuanian, Polish, and Slovenian newspapers. These papers were chosen through the scholarly contribution of a state-wide selection committee comprised of area studies scholars who looked at the role each paper played, not only for the history of Illinois and American multiculturalism, but also for the papers' global importance. Due to ethnic conflict in the Baltic area, Chicago's Lithuanian newspapers have a particular relevance for international Slavic researchers. One title selected, *Draugas* (Chicago, IL), is the oldest continuously-published Lithuanian newspaper in the world. While neither the Polish nor the Czech press was systematically repressed to the same level experienced by the Lithuanian press, the contributions these communities had on the fabric of Chicago-life and American-life is immense.

In July 2016, the NEH announced that it would begin accepting proposals to digitize newspapers content published between 1923 and 1963.[1] These newspapers were previously omitted because copyright law does not guarantee works published after December 31st, 1922 to be in public domain.[2] However, by including post-1922 content, awardees are required to conduct extensive research on the copyright of newspapers, as well as "third party materials"[3] printed in those

---

[1] Expanding Our Current Scope | NDNP, accessed February 28, 2018 https://www.neh.gov/news/expanding-our-current-scope-ndnp.

[2] The Copyright Act of 1976, (Title 7, App. A), accessed February 28, 2018, https://www.copyright.gov/title17/AppendixA.pdf.

For further information regarding the complexities of United States copyright law the works of Peter Hirtle are recommended. A summary can be found here: https://copyright.cornell.edu/publicdomain.

[3] About Chronicling America, accessed February 28, 2018, https://chroniclingamerica.loc.gov/about/.

newspapers. In addition to researching the copyright of post-1922 newspapers, awardees are must admit indemnification if content is found to be in copyright. As a result of the risk of possible legal action due to accepting indemnification, U of I legal administration staff chose not to allow for the U of I to accept the contractual requirements to include post-1922 content, and as such, title considerations must fall firmly within pre-1923 public domain. Given the timeframe of demographic shifts in the United States, the aforementioned limitations still allow for the inclusion of substantial amounts of content from Central-European and Eastern-European communities in Chicago.

While the LC promises to provide long-term access to and preservation of digitized newspapers, the U of I, along with 21[4] other institutions that also make their content accessible through an institutional newspaper portals. Systems used among institutions include systems implemented and maintained by a contracted vendor, such as Veridian or CONTENTdm, and systems implemented and maintained by the institution, such as in-house institutional repositories or ChronAm.[5] In addition to providing access, the U of I has developed a robust digital preservation framework to support the long-term sustainability of content.

The authors of this paper, cognizant of the challenges to digitize non-English newspapers, developed a comprehensive and well-documented workflow, established best practices for digital preservation within U of I's institutional repository, and increased engagement efforts with communities to which the newspapers are of importance. These efforts have enabled more efficient and engaging access of digitized newspapers, and encouraged the IDNP team to reconsider the project lifecycle, evaluating digital preservation best practices at an institutional level. The primary purpose of this paper discusses methods of resolving errors encountered throughout the newspaper digitization lifecycle, focusing specifically on those obstacles unique to non-English newspapers. The authors provide an overview of non-English newspaper digitization projects and situate this in IDNP's Phase IV, exploring the importance of turn-of-the-century newspaper publication in the context of Chicago immigrant communities. This lends to an in-depth case study of the NDNP 2016–2018 project at U of I. Finally, the authors will provide practical recommendations for future newspaper digitization projects.

2 LITERATURE REVIEW

2.1 Non-English Newspaper Digitization

Digitization projects are notorious for being approached thusly—as projects. As such, project staff may fluctuate, especially within a university environment, where students are employed to complete daily tasks. Naomi Skulan recounts the successes and challenges that the University of Minnesota, Morris encountered when completing a digitization project that was largely staffed by students and student volunteers.[6] Throughout the project, permanent staff found that their

---

[4] This figure was identified by searching for participating awardees' digitized newspaper portals.

[5] chronam, GitHub repository, accessed February 28, 2018, https://github.com/LibraryOfCongress/chronam.

[6] Naomi Skulan, "Staffing with Students: Digitizing Campus Newspapers with Student Volunteers at the University of Minnesota, Morris," Digital Library Perspectives 34, no. 1 (December 7, 2017): 32–44, https://doi.org/10.1108/DLP-07-2017-0024.

most difficult challenge was "not feeling that they [could] criticize the volunteers' work."[7] Among other costs was the need to recruit, organize, supervise, and QC the volunteers' work. While Skulan reports on the extensive cost of volunteers, it is important to note that even paid non-permanent staff can be just as, if not more, costly. For example, the University Libraries of the U of I employ graduate assistants, who receive a tuition waiver in addition to a stipend for their work. Their employment is designed to be a space for career exploration that complements their study. Yet even with these high stakes, graduate assistants are non-permanent staff. As such, without meticulous documentation, institutional knowledge can be lost with changes of the non-permanent project staff. For digitization of non-English newspapers, thorough documentation is requisite to tracking errors when troubleshooting or preventing errors previously encountered.

As aforementioned, since 2011, NDNP has continuously attempted to expand non-English content available in Chronicling America.[8] Their 2015–2017 project description stipulated that "Applicants may select titles published in Danish, English, French, German, Hungarian, Italian, Norwegian, Portuguese, Spanish, and Swedish," and noted that "More languages [would] be added in future years."[9] By the following year, the NEH expanded their guidelines to permit "titles published in any language with a valid ISO 639-2 language code."[10] Since NDNP began an expansion to include non-English newspaper titles, several institutions have begun expanding their digitized collection contribution to Chronicling America. Among the most represented languages are German, Italian, and Spanish. However, the continued expansion of less commonly selected languages is largely credited to Minnesota Historical Society, which has contributed to Chronicling America newspapers in Czech, Dakota, Danish, Finnish, French, German, Icelandic, Norwegian, Polish, Slovenian, and Swedish—greater language representation than any other awardee. See figure 2 for a detailed illustration of participating states and their non-English contributions to Chronicling America.

---

[7] Ibid., 38.

[8] The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants, (August 19, 2011), http://www.loc.gov/ndnp/guidelines/archive/NDNP_201113TechNotes.pdf.

[9] Division of Preservation and Access, National Endowment for the Humanities, National Digital Newspaper Program for projects beginning September 2015, accessed February 28, 2018, https://web.archive.org/web/20150405031057/http://www.neh.gov/files/grants/ndnp-jan-15-2015.pdf.

[10] Division of Preservation and Access, National Endowment for the Humanities, National Digital Newspaper Program for projects beginning September 2016, accessed February 28, 2018, https://web.archive.org/web/20160403181741/http://www.neh.gov/files/grants/ndnp-jan-14-2016.pdf.

Figure 2. Distribution of non-English language newspaper digitization (by number of awardees).
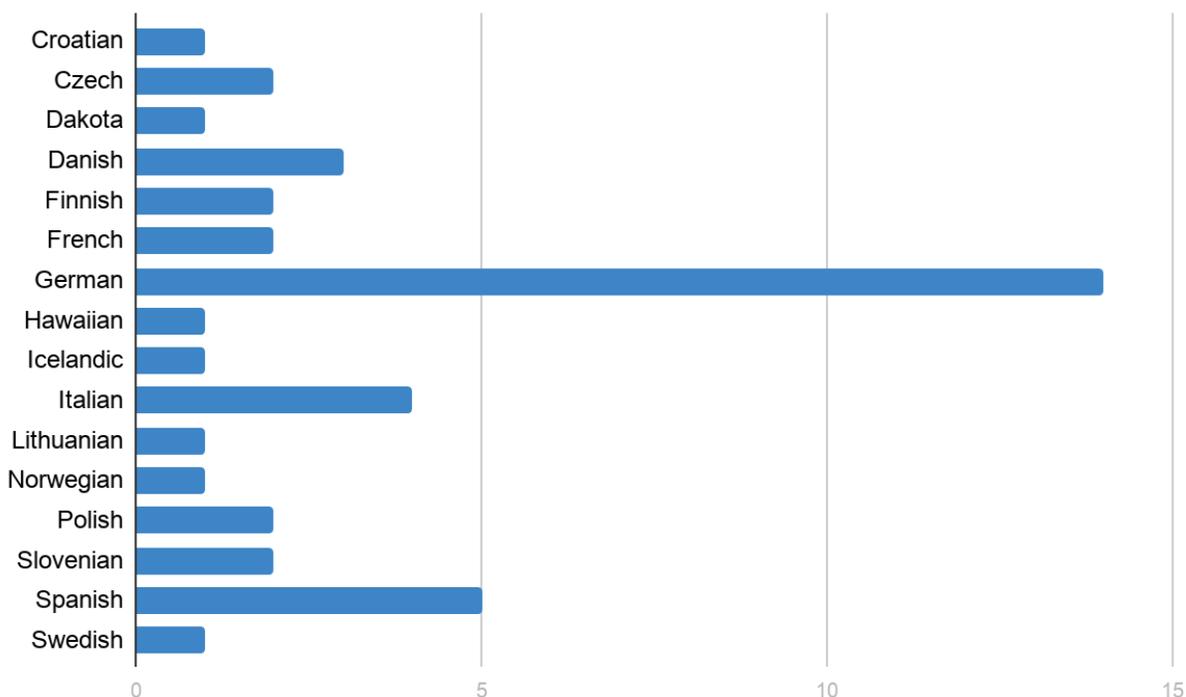
In addition to the encouragement from the NEH and LC to make Chronicling America a more comprehensive and representative digital library of United States newspapers, memory institutions across the globe have digitized their newspaper collections, some of which consist newspapers in published in the United States. Among these newspapers digitized outside of United States are *Draugas* (Chicago, IL)[11] and *Proletarec* (Chicago, IL).[12]

2.2 Chicago Immigrant Communities

The current cycle of NDNP grant funding builds on digitization of central and eastern European immigrant community newspapers from the Chicago metropolitan area conducted in, and all newspapers that serve to inform this case study are exclusively from these communities. Immigrant groups represented in IDNP Phase IV include the following ethnicities: Polish, Lithuanian, Slovenian, and Czech. As a result of political upheaval, economic difficulties, and religious persecution during 1870–1920 these ethnic groups experienced substantial demographic shifts, with some choosing to immigrate to the United States of America, and from there many subsequently settling in Chicago. Throughout the rich and extensive history of Chicago each of these ethnic groups has provided a unique contribution to the multicultural tapestry of the city. Many of these communities continue to play a unique and visible role in Chicago daily life through cultural societies, schools, museums, and churches.

---

[11] *Draugas*, Martynas Mažvydas National Library of Lithuania, accessed March 14, 2018, http://www.epaveldas.lt/object/recordDescription/LNB/C1B0001661471.

[12] *Proletarec*, Digitalna Knijižnica Slovenije, accessed March 14, 2018, https://www.dlib.si/details/URN:NBN:SI:spr-3YQG3P5E.

The Polish community comprises the largest of the ethnic groups representation in the current NDNP grant cycle in terms of both total population in Chicago and number of pages digitized: currently almost one in ten persons in Chicago claims Polish ancestry, comprising of some 821,000 people in the Chicago area. Phase IV includes 88,375 pages of Polish content across 15,074 number of issues in five titles.[13] As was common with central and eastern European immigrants from overly agrarian societies Poles flocked to jobs that were to become notorious for poor working conditions and wages, such as steel mills, slaughterhouses, and other forms of grueling industrial labor. William J. Galush, historian of Polish-American history, states that in terms of population and cultural organizations "...gave Chicago a singular statue in American Polonia."[14] Along with cultural organizations newspapers formed an important source of information both from the homeland and similar ethnic communities both on a local and national level. Newspapers often acted as a fertile sources of education in a new land, loci for organization, and catalysts for change. For most of these immigrant groups, the rights afforded by the First Amendment such as freedom of the press were hitherto unheard of, and this new found freedom let to a cornucopia of ideas and beliefs to come to the forefront of public life for immigrant communities. Early Chicago immigrant communities often congregated around church parishes and cultural community organizations, savings and loan societies, gymnastics groups, and fraternities. This was especially true of the Polish community, with titles such as *Dziennik Chicagoski* (Chicago, IL), published by the Resurrectionist Catholic movement; *Zgoda* (Chicago, IL) run by the Polish National Alliance; and *Głos Polek* (Chicago, IL), which acted as the organ of the Polish Women's Alliance of America.[15,16] Intriguingly, despite the working class status of this community, radical and partisan politics were not often at the forefront of these newspapers, and while the community at large was often loyal to the Democratic Party and far from apolitical Chicago historian Dominic Pacyga notes that Polish newspapers were "...constantly calling for support of Polish candidates despite factional or party affiliation."[17] As a result of strong community ties and industriousness Polish-American communities continue to play a vibrant role in Chicago's ethnic landscape.

While Lithuanian, Slovenian, and Czech immigrant groups (of which Phase IV will digitize 12,705, 3,350 and 1,234 pages of newspaper, respectively) did not flock to Chicago in comparable numbers to Polish immigrants they nonetheless set up their own community ethnic organizations, churches, and of course, newspapers. Many of these groups congregated around the Pilsen neighborhood, as well as the infamous Back of the Yards area, made famous in part for its role in Upton Sinclair's 1906 novel *The Jungle*, which prominently featured the Lithuanian immigrant protagonist Jurgis Rudkus's fight for economic justice in abysmal working conditions. By 1914 due to political oppression at home Chicago's Lithuanian population swelled to upwards of 50,000, making it the largest Lithuanian city in the world. From this community

---

[13] Titles include *Dziennik Chicagoski*, *Głos polek*, *Zgoda = Unity*, *Zgoda : Wydanie dla mężczyzn (Men's edition)*, and *Zgoda : Wydania dla niewiast (Women's edition).*

[14] William John Galush, *For more than bread: community and identity in American Polonia, 1880–1940* (Boulder, Colo.: Columbia University Press, 2006), 26.

[15] Victoria Granacki, *Chicago's Polish Downtown* (Charleston, SC: Arcadia, 2004), 69-70.

[16] Andrzej Brożek, *Polish Americans, 1854–1939* (Warsaw, Poland: Interpress, 1985), 224.

[17] Dominic A. Pacyga, *Polish Immigrants And Industrial Chicago: Workers On The South Side, 1880–1922.* (Chicago, University of Chicago Press, 2003), 198.

came both *Draugas* (Chicago, IL) and *Lietuva* (Chicago, IL)*,* the former of which has the distinction of being the longest running Lithuanian newspaper in the world, forming an important voice for Lithuanian Catholics, while the later espoused a moderate nationalist worldview.[18] Slovenian immigrants, connected by periodicals such as *Glas Svobode* (Chicago, IL), formed a nuanced population comprised of socialist freethinkers and more traditional Catholics.[19] Czech populations were considerably more secular, with most Bohemians reading *Tydenni Hlastel* (Chicago, IL) and *Hlastel* (Chicago, IL) in the unique and especially popular Bohemian saloons.[20] Regardless of their political or religious affiliation, many of these newspapers and organizations lasted through periods of assimilation, economic downturn, immigration restrictions, and demographic shifts; and continue to serve their respective communities to this very day.

3 CASE STUDY OF NON-ENGLISH NEWSPAPER DIGITIZATION AT U OF I

3.1 Previously Digitized Content

With both preservation and access goals in mind, titles were selected and given to outside vendors to produce the digital surrogates and metadata required to build an online repository. Concurrent with work for NDNP, the U of I digitized three major and distinct thematic collections: the Collegiate Chronicle; American Popular Entertainment; and Farm, Field, and Fireside.[21] Motivated due to patron use and, in part due to an interest in reducing collection footprint space, opening access to previously unavailable brittle newspaper content, and refilming material filmed prior to the introduction of contemporary best practices for preservation microfilming, the U of I successfully digitized 22 titles, comprising of 22,737 issues, under the Farm, Field, and Fireside project. In addition to the rural newspapers digitization project, the U of I conducted a second major reformatting project, this time focusing on entertainment industry related newspapers and trade journals published in the US between 1853 and 1929. Dubbed the American Popular Entertainment collection, these works, made up of three titles comprised of 3,993 issues, focus on popular entertainment of the period, including extensive information on vaudeville, burlesque, and musical performances.[22] These projects were the first major foray into large-scale newspaper digitization at the U of I and supported initiatives of sustainable and systematic planning to accomplish the scale and programmatic requirements set out by NDNP. Initial digitization efforts

---

[18] Danguole Kviklys, "'Draugas' The Lithuanian World-Wide Daily Newspaper, 1909–2009," *Lituanus* 55, no. 4 (2009), http://www.lituanus.org/2009/09_4_01%20Kviklys.html. This article provides an impressively detailed overview of *Draugas* and the Chicago-based Lithuanian press as a whole.

[19] Emily Brunner, "Slovenes," The Electronic Encyclopedia of Chicago, 2006, http://www.encyclopedia.chicagohistory.org/pages/1153.html.

[20] Melvin G. Holli and Peter d'Alroy Jones. *Ethnic Chicago: a Multicultural Portrait*. (Grand Rapids, Mich. W.B. Eerdmans Pub.Co, 1995), 509.

[21] Monetary support came from a variety of sources including a Library Services and Technology Act grant from the Institute of Museum and Library Services via the Illinois State Library, the Clifford Family Endowment, Lancaster Farming, Inc., the Minnesota Historical Society, the Pennsylvania State University, the Wisconsin Historical Society, the Norman Jean Johnston Estate, and collection development funds from the University of Illinois Library. A guide with more information on the Farm, Field, and Fireside collection can be found here: http://uiuc.libguides.com/fff/.

[22] The American Popular Entertainment Collection, digitized thanks to the generosity of donor Robert O. Endres, is accessible here: http://go.library.illinois.edu/APEcollection.

focused on local Illinois newspapers, these local collections worked in tandem with the Collegiate Chronicle collection to present college newspapers from different regions and time periods across the United States. In addition, the U of I also developed the American Popular Entertainment collection. This collection features three English language newspapers and trade journals published for the entertainment industry in the US between 1853 and 1929. Furthermore an additional collection: The Farm, Field, and Fireside Collection, supported by a wide range of federal, state, and private funding ranging from the Institute of Museum and Library Services (IMLS) to the U.S. Department of Agriculture, digitized 22 unique rural American titles with a focus on agriculture, politics, family life, and farming technology. This project which utilized one of the largest collections of original farm newspapers in the world showcased the transformation of American agriculture during the turn of the 20th century. Most recently, from 2015–2016, the University Library conducted a grant-based partnership with ISL which administered Library Services & Technology Act Funds to digitize 210,895 pages of microfilmed Illinois-based English language newspapers.

3.2 Workflow

Upon receiving grant funding, the IDNP finalizes a possible list of titles to digitize and consults the Advisory Board to determine which title candidates would have the most research impact. IDNP then outsources duplication of microfilm to create a third-generation positive microfilm to be used for collation and a second-generation silver negative microfilm to be used for digitization and which will be a final deliverable to the LC as the master preservation copy. An IDNP staff member then collates the microfilm, documenting issue level metadata and page counts, reel level metadata, and elements of bibliographic significance in an authority file, which is sent with the silver negative microfilm to the digitization vendor. The digitization vendor is responsible for creating a digital facsimile (TIFF) and derivative access files (JPEG2000, PDF, and ALTO XML).

Each NDNP awardee follows a similar process from title selection and title cataloging to identifying microfilm holdings and duplicating microfilms. While non-English newspapers require additional expertise for subject analysis during these primary steps, the greatest challenge IDNP encountered was the creation of digital content and QA of content. Even with language support for special cases, the lack of permanent language support staff completing collation of microfilm and review of digital derivatives imposed challenges to delivering quality content. Challenges experienced during the pilot project have informed a mainstreamed, more in-depth workflow. The project staff implemented these additional steps throughout both the collation and QA processes to avoid potential errors in digitized content.

During the collation process, project staff rely upon understanding of the general format of a particular newspaper. From this, staff will know where to identify section or edition information. The project staff then determines whether their relevance is substantial enough to be encoded in the METS file as a section label or as an edition. For example, from September 1914 through January 1916, *Dziennik Chicagoski* (Chicago, IL) published a section titled "Ostatnie Wiadomosci Z Ojczyzny" (Recent News from the Motherland), featuring news from Poland on the progression of World War I. This was considered bibliographically and historically significant and was thus encoded in the METS XML file.

The silver negative microfilm reels are shipped to a vendor for digitization and file processing. The vendor performs preliminary QA before sharing the files with the IDNP staff for in-depth QA. The QA process requires further steps to accommodate the inclusion of several languages. Among

verifying bibliographic correctness, content quality review common to all newspapers includes checking for:
- column zoning,
- order of OCRd text in both ALTO and PDF files,
- search highlights of ALTO content blocks,
- and the ISO 639-2 language encoding of ALTO TextBlocks[23]

In addition to these checks, non-English content requires more meticulous review of the ISO 639-2 codes, specifically for multilingual issues, and review of diacritics. In newspaper issues containing more than one language, ISO 639-2 codes may need article-level review.[24] At U of I, diacritic errors have occurred because of incorrect input information (e.g., OCR software not set to correct language dictionary;[25] not set to correct font capture)[26] and failure of the operating system.

Upon acceptance receipt from the LC, newspapers batches are restructured[27] and ingested into Medusa. After the ingestion of files completes, the TIFFs are removed from the batch and the remaining access files are ingested into IDNC.

3.3 Access and Sustainability

3.3.1 Access

Comparison of Chronicling America and IDNC
Chronicling America is a website that provides access to newspapers digitized as part of the NDNP. The website is built upon chronam, an open source application that uses "the Python programming language, Django Web framework, RDFLib, Apache Solr search server, Apache Web server, and MySQL database engine."[28] Like Chronicling America, IDNC is a website that provides access to digitized newspapers. IDNC is maintained by Veridian Software of DL Consulting. Veridan Software runs on Apache Tomcat and Apache Solr, and enables patron engagement with their User Text Correction (UTC) OCR editing feature.[29,30]

Both Chronicling America and IDNC support full-text search, browsing by calendar, and download of newspaper files. The two differ most substantially in their patron engagement features and their

---

[23] NDNP Awardee Wiki, "OCR Accuracy and Quality Review," accessed March 21, 2018, http://www.loc.gov/extranet/wiki/library_services/ndnp/index.php/OCR_Accuracy_and_Quality_Review (NDNP awardee login required).

[24] For example, English articles were printed amongst Polish articles in *Dziennik Chicagoski*, and Croatian articles were printed on the same page as Slovenian articles in *Proletarec*.

[25] For example, Croatian language content was set to a Slovenian OCR dictionary for file processing of *Proletarec*.

[26] When digitizing *Vorbote*, a German-American newspaper, the OCR was not set to capture fraktur typeface, which caused myriad errors in the capture of diacritics.

[27] The file structure of newspaper content in Medusa is further illustrated and explained in Section 3.3.2.

[28] About Chronicling America.

[29] Veridian discovery and delivery platform, accessed March 21, 2018, https://www.veridiansoftware.com/knowledge-base/veridian/.

[30] Stefan Boddie, Crowdsourced User Text Correction (UTC), June 24, 2014, https://www.veridiansoftware.com/knowledge-base/crowdsourced-user-text-correction/. Text corrections are not made to the original ALTO file: https://www.veridiansoftware.com/knowledge-base/user-text-corruption/.

image view. As aforementioned, Veridian Software enables text editing of the ALTO OCR files. Due to system priorities, Chronicling America does not offer this feature.[31] LC prioritizes providing sustainable and scalable access to a continuously expanding digitized collection numbering in the tens of millions of pages. As a result, this priority takes precedence over providing an enhancement tool such as crowdsourced text correction. Nonetheless, LC continues to offer the Chronicling America software, chronam, as an open-source software. NDNP awardee institutions that chose to implement chronam software for their digitized newspaper collections collaborated to establish openoni, which continues to be maintained by the NDNP awardee community.[32] Chronicling America offers an issue-level image view and a page-level image view. As seen in figures 3 and 4, in the issue view, all pages of the issue are visible. In the page view, the page image can be zoomed in and out, and page derivatives can be viewed or downloaded (including PDF, JPEG2000, ALTO OCR, and plain text OCR).

Figure 3. All page issue view in Chronicling America.

---
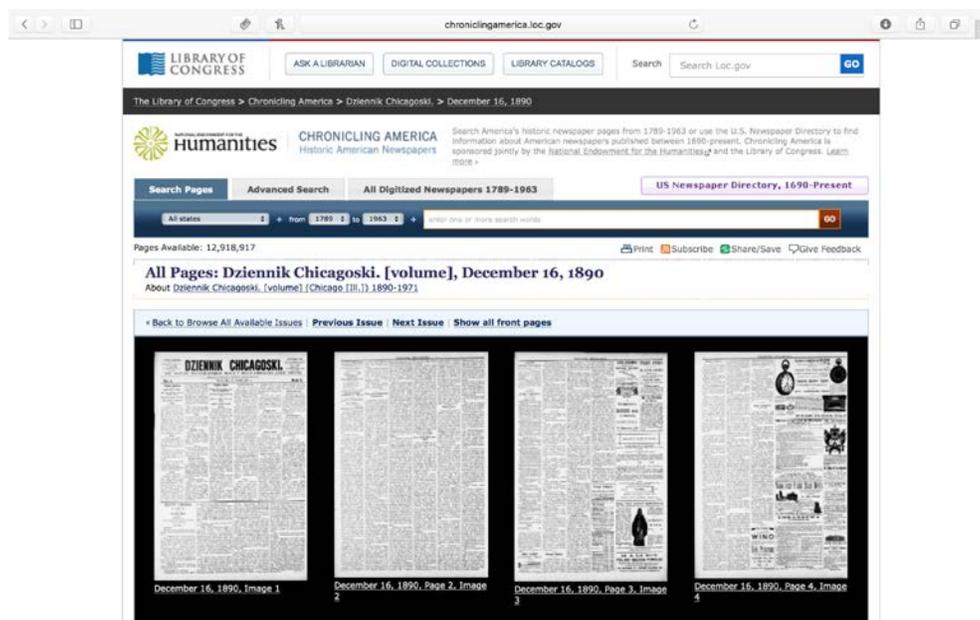
[31] In 2017, the LC released "Beyond Words," a project to identify and annotate WWI-era newspaper photographs and illustrations as part Labs: https://labs.loc.gov.

[32] openoni, GitHub repository, accessed March 21, 2018, https://github.com/open-oni/open-oni.

Figure 4. Issue view in IDNC.
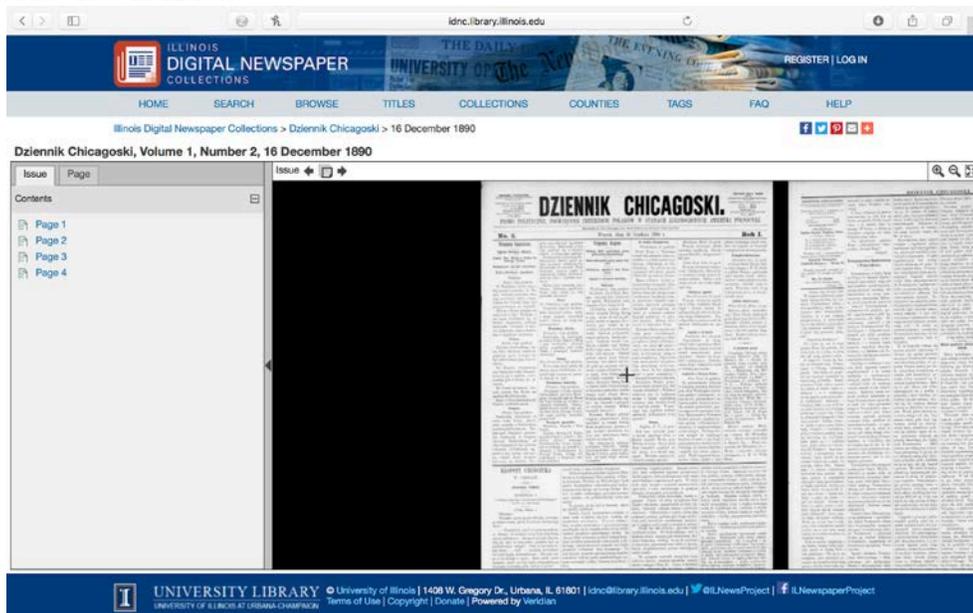
The ALTO XML files contain text content data and text location data. The text location data enables a newspaper portal to return pages where the search term is highlighted on pages where the term appears. A search of "zarzadzie" was input into both Chronicling America and IDNC. The output of each as it appears in the issue view is illustrated in figures 5 and 6.

Figure 5. Chronicling America search highlights of "zaradzie."
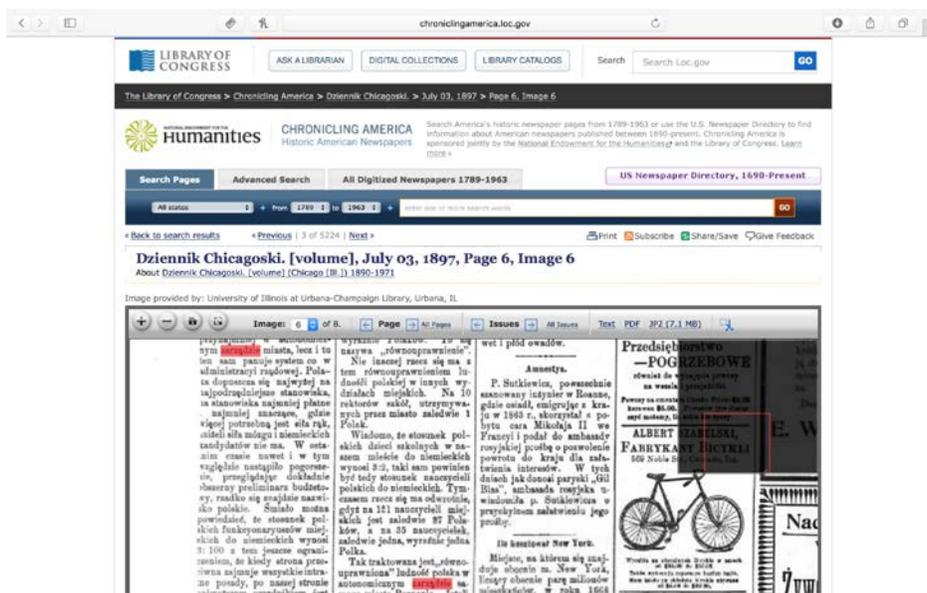
Figure 6. IDNC search highlights of "zaradzie."
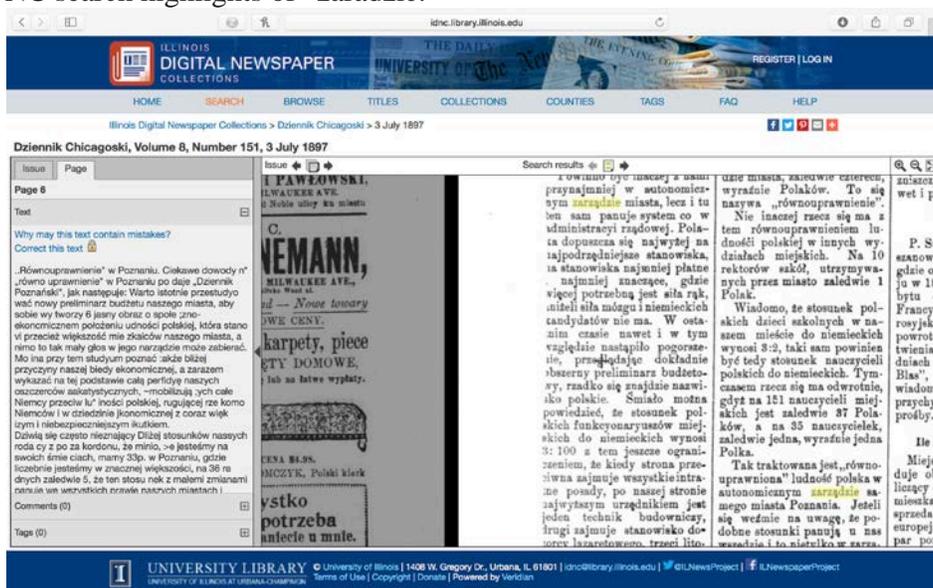
Interoperability and Universality

As digital libraries continue to function as a primary access point for information, standardization and interoperability remain vital to the long-term sustainability of content access. While the NDNP technical specifications have become the ultima Thule of newspaper digitization with specifications based upon standards, the chronam software continues to compete as a preferred digital access newspaper system (DANS).[33] Other DANS, such as Veridian Software, provide various functionalities for searching, image viewing, and downloading. Among the most common image support is the International Image Interoperability Framework (IIIF). IIIF defines itself as "a set of shared application programming interface (API) specifications for interoperable functionality in digital image repositories" and promotes, per its namesake, interoperable of image access.[34] Despite its uptake as a standard for image APIs and image viewers, IIIF is not a mainstay for newspaper portals.[35] Fishrappr, the University of Michigan newspaper portal, is an instance of IIIF implementation. In addition to use of IIIF, Fishrappr continues to apply popular systems, including Blacklight and Samvera.[36]

The LC has taken measures to improve Chronicling America's search functionality of non-English content by offering two levels of search capacity: 1) language analysis search support and 2) exact match search. Language analysis supports "stemming, stop words, and/or variants" searches but is

---

[33] The authors have coined this term to refer to systems that provide digital access of a newspaper form, whether born digital or digitized.

[34] IIIF Frequently Asked Questions (FAQs), accessed March 21, 2018, http://iiif.io/community/faq/#what-is-iiif.

[35] The discussion of IIIF is a potential image viewer and API would require an evaluation of the use of aggregation of newspaper images, which Chronicling America does to an extent considering the scale of NDNP. However, the benefits and risks of IIIF as a future step for digitized newspaper portals is out of scope for this paper.

[36] Nabeela Jaffer, Michigan Daily - From Newspaper Prints To Digital Archives, April 18, 2017, https://www.lib.umich.edu/blogs/library-tech-talk/michigan-daily-newspaper-prints-digital-archives.

limited to the languages listed in appendix 1.[37] All other languages with an ISO 639-2 code are limited to exact match searches. Still, the nature of the digitized content does restrict access. With OCR computer-generated text, there exist errors inherent to the content being searched. Thus, even if a patron has entered a search-term correctly, the underlying data may produce unexpected search results.

In 2015, the IDNP launched their updated digitized newspaper access portal, Illinois Digital Newspaper Collections (INDC), with Veridian Software of DL Consulting. The decision to migrate from the previous platform, Olive ActivePaper, to Veridian came with thorough research of preferred features, in addition to informal surveys of IDNC users. This research began in 2012 with the formation of the Newspaper Delivery and Preservation Working Group. The preferred features for the newspaper access site included, article segmentation, downloadable content, aggregated search, access and view of images and OCR text, crowdsourced OCR correction, and article tagging. Among the back-end features were collection arrangement options, METS-ALTO encoding standard, lost-cost production and sustainable workflows, reliable download from any IP address, and efficient content management. The working group reviewed five content management systems and services and found that Veridian met all preferred criteria for user access and administrative functionality.

Perhaps the greatest challenge institutions encounter is deciding what kind of software best matches their institutional needs and priorities. While open source software is attractive for as a "free" option, for some institutions, the long-term cost can be more extensive than contracting a vendor. However, by contracting a vendor, an institution removes their ability to fully moderate the longevity and sustainability of a platform, due to total dependence upon the vendor support and continuation.

3.3.2 Sustainability

Preservation at the Library of Congress
The LC asserts that preservation is fulfilled through each awardees' deposit of the silver negative microfilm used for digitization, which the LC treats as the preservation master. However, they also take measures to preserve the digital content in their digital repository "based on the Open Archive Information System (OAIS) Reference Model."[38] Upon ingest into the repository, the digital content is written to Linear Tape-Open (LTO) tape. Every few years, the content is rewritten to new LTO tape. The file formats specified in the NDNP technical guidelines have been thoroughly reviewed by the LC digital preservation practitioners for the formats' sustainability factors. These include disclosure, adoption, transparency, self-documentation, external dependencies, impact of patents, and technical protection mechanisms.[39] Format profiles for each file type are available on the

---

[37] NDNP Awardee Wiki, "Encoding Non-English Language Text," accessed March 21, 2018, http://www.loc.gov/extranet/wiki/library_services/ndnp/index.php/Encoding_Non-English_Language_Text (NDNP awardee login required).

[38] About Chronicling America.

[39] Sustainability of Digital Formats: Planning for Library of Congress Collections, accessed March 21, 2018, https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml.

"Sustainability of Digital Formats: Planning for Library of Congress Collections" website: TIFF 6.0,[40] JPEG2000,[41] PDF,[42] and XML.[43]

Preservation at the University of Illinois at Urbana-Champaign
Critical to the long-term digital preservation of digital content at the U of I is Medusa.[44] Medusa functions as a digital preservation repository and collection registry to provide an enduring storage and management environment for digital collections across the U of I. The collection registry is accessible only to digital content producers at the U of I. Storage is managed in partnership with the National Center for Supercomputing Applications (NCSA) acting as an additional on-campus storage node in addition to off-campus storage currently utilizing Amazon Glacier services, upholding Level 3 compliance with the National Digital Stewardship Alliance (NDSA) levels of preservation.[45] Through Medusa, the U of I Library also provides a Digital Collections portal.[46] This portal brings together critical library collections such as digitized and born-digital books, photographs, audiovisual materials, and newspapers. While primarily built to showcase collections housed in the Medusa service, the Digital Collections portal also links out to other services and platforms, such as the Illinois Digital Newspaper Collections, Veridian-based access points. In addition to hosting, managing and preserving internal library collections, Medusa also accommodates students and scholars across the U of I system by automatically syncing data ingested into the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS) and Illinois Data Bank platforms.[47,48]

Medusa enables collection moderators, such as newspaper reformatting staff, the ability to edit descriptive, administrative, and rights metadata associated with collections. Preservation staff are easily able to track the provenance and statistics of files, in addition they are able to verify the MD5 checksum to ensure that files have not changed since ingest. Ongoing checksum verification is run across Medusa, automatically checking files every 90 days. Medusa also provides the ability for preservation staff to extract technical metadata utilizing the File Information Tool Set (FITS) on-demand. All of these tools ensure that robust and thorough digital preservation practices are followed in a uniform fashion across the digital collections.

---

[40] "TIFF, Revision 6.0," Sustainability of Digital Formats: Planning for Library of Congress Collections, updated September 28, 2009, https://www.loc.gov/preservation/digital/formats/fdd/fdd000022.shtml.

[41] "NDNP Profile for JPEG 2000, Part 1, Core Coding," Sustainability of Digital Formats: Planning for Library of Congress Collections, updated August 16, 2006, https://www.loc.gov/preservation/digital/formats/fdd/fdd000192.shtml.

[42] "PDF/A-1, PDF for Long-term Preservation, Use of PDF 1.4," Sustainability of Digital Formats: Planning for Library of Congress Collections, updated October 6, 2010, https://www.loc.gov/preservation/digital/formats/fdd/fdd000125.shtml.

[43] "XML (Extensible Markup Language) 1.0," Sustainability of Digital Formats: Planning for Library of Congress Collections, updated March 9, 2008, https://www.loc.gov/preservation/digital/formats/fdd/fdd000263.shtml.

[44] Medusa: https://medusa.library.illinois.edu/.

[45] Megan Phillips, Jefferson Bailey, Andrea Goethals, and Trevor Owens, The NDSA Levels of Digital Preservation, 2013, http://www.digitalpreservation.gov/documents/NDSA_Levels_Archiving_2013.pdf.

[46] Digital Collections: https://digital.library.illinois.edu/.

[47] IDEALS: https://www.ideals.illinois.edu.

[48] Illinois Data Bank: https://databank.illinois.edu.

Medusa provides the ability for collection managers with proper permissions to download files or batches of files. Due to a comprehensive restructuring and revitalized ingest workflow project, digitized newspaper collections at the U of I are now structured in such a way to unify best digital preservation practices by elaborating file hierarchy policies as well as ease of collection manager access. The evolution of these changes can be seen in appendices 3-5. In many cases, library staff require a higher-quality image for a patron request, and with the re-envisioned, description-based collection hierarchy, they are more easily able to access and download in Medusa.

3.4 Outreach

3.4.1 Online efforts

One critical, albeit often challenging, aspect of U of I's involvement as a producer of digitized historical newspapers is properly showcasing and connecting relevant patrons to the breadth and depth of dozens of titles and hundreds of thousands of pages digitized under the auspices of NDNP. Digitization done in relative isolation, even with adherence to national file format standards and robust digital preservation will often neglect patron engagement through appropriate, targeted outreach efforts. Outreach efforts and subsequent patron usage provide an answer to the question of practical and demonstrative use that often comes with mass digitization projects. Online outreach efforts have taken many relatively conventional forms, with LibGuides and social media efforts forming the primary means of project promotion.[49] The authors are investigating additional avenues of connecting content to both relevant users and those previously unaware of the IDNC platform. Ongoing projects include the development of a Libguide for use with the U of I's University Laboratory High School 8-12 grade history curriculum. In addition, efforts are underway to promote increased crowd-sourced OCR correction participation through the means of contests and promotional flyers. These efforts can often be accomplished at minimal cost and relatively low staff time commitment. Nonetheless, for a truly successful effort, the authors recommend creating a social media framework and action plan, in addition to identifying possible collaborators both within the home institution and project partners (e.g., institutions with previously successful online patron engagement efforts).[50]

3.4.2 In-person efforts

In addition to online efforts, the U of I NDNP program is planning and expanding in-person outreach efforts. Of particular relevance for working with non-English content is connecting with relevant cultural centers that have a historical connection to the content. Thus far, the authors have promoted Phase IV NDNP efforts via historical regional conference presentations,

---

[49] IDNC Facebook: https://www.facebook.com/ILNewspaperProject/,

IDNC Twitter: https://twitter.com/ILNewsProject,

IDNC Pinterest: https://www.pinterest.com/idncuiuc/.

[50] Erenst R. Anip, of the University of Vermont's Vermont Digital Newspaper Project presents an excellent overview in his publication "START SPREADING THE NEWS. A journey in social media usage for outreach in a cooperative digitization project" of strategies employed by Vermont to increase collection visibility through outreach efforts. One recommendation is that NDNP awardee institutions make greater use of communication systems in place (e.g. the NDNP Awardee Wiki and listserv) to share ideas concerning outreach and better coordinate efforts across the program. URL: https://scholarworks.uvm.edu/cgi/viewcontent.cgi?article=1031&context=libfacpub.

networking with relevant institutional staff such as those within the aforementioned University Laboratory High School as well as the Russian, East European, and Eurasian Center and International Area Studies Slavic reference librarians. Along similar lines, the U of I's NDNP program employs student assistants with native-level knowledge of the newspapers. This employment allows program staff to develop title essays utilizing both primary and secondary sources, in addition to connecting members of the community to historical newspapers relevant to their personal history. Though these efforts, the authors have gained a deeper understanding of the content digitized and a deeper appreciation for the historic importance these papers have on the local communities they served. One area of collaboration currently being piloted is possible collaboration with relevant ethnic community centers (e.g., the Lithuanian Research and Studies Center in Chicago). Collaboration with cultural centers allows for greater promotion of non-English materials to the communities with a particular connection to the content, both domestically and internationally. Similar to staff in other cultural heritage institutions, library staff should take into account from the beginning of a project the unique attachment that groups may attach to culturally relevant materials such as in the authors' case with ethnic newspapers. Working with these groups and articulating and clarifying the digital newspaper lifecycle provides, not only greater participation and insight, but also the potential for increased project visibility and the need for robust digital preservation efforts.

Digitization projects dealing with non-English content can present a unique problem for outreach efforts, both online and in person. Many institutions have circumvented this difficulty through online resources that aim to connect both native speakers and learners of languages found within NDNP digitization projects. In particular, the authors find the efforts of the Minnesota Historical Society and their Swedish American Newspapers website particularly noteworthy.[51] Additionally the University of Arizona also promotes non-English content via their unique Historic Mexican & Mexican American Press website.[52] Both of these platforms showcase robust outreach efforts that showcase content in historical and demographic content. Well planned and conceived outreach efforts serve to better amplify use of materials by bringing them to the attention of often overlooked patron groups.

4 CONCLUSION AND RECOMMENDATIONS

Throughout their work with U of I's Phase IV NDNP efforts the authors stress two essential, practical elements that would assist other institutions undertaking or planning to undertake large-scale non-English newspaper digitization projects: (1) the development of comprehensive documentation for common challenges encountered throughout the newspaper digitization lifecycle and (2) planned outreach efforts both on and offline geared to connecting patrons with the content, bringing the project to the attention of relevant user groups, and engaging in both intra and inter-institutional efforts to enable greater publicity and creative use of digitized content.

Robust, detailed, and up-to-date documentation is critical to the success of programmatic digitization efforts. This is especially the case for institutions that experience staff turnover, as if often the case for academic institutions that often rely on a student workforce. With proper

---

[51] Swedish American Newspapers: http://www.mnhs.org/newspapers/swedishamerican.

[52] Historic Mexican & Mexican American Press: http://www.library.arizona.edu/contentdm/mmap/index.html.

documentation in place, time can be more effectively utilized so as to allow more extensive QA, more outreach planning, and more engagement with patrons. Documentation allows for a reduction in staff training time, an increase in staff accountability, workflow transparency for project managers, and the ability to pinpoint possible inefficiencies or redundancies in order to better streamline workflows. All too often, staff neglect to share workflow updates or procedures that allow for greater productivity. Compiling documentation on a shared, easily editable platform such as a wiki, a Google Doc, or other collaborative online workspace encourages staff collaboration and up to date information.

The authors have found it especially useful to develop a common shared online space to house documentation relating to each step of the digitization process in addition to relevant resources such as language collation documentation, meeting minutes, and updated stakeholder contact information. Maintaining these documents may initially seem daunting, but once implemented into daily activities, allow for better project management and reduce the possibility of institutional memory loss. Upon development of documentation, the authors noticed a number of oversights within U of I's newspaper digitization process. Many of these oversights related to a lack of comprehensive standardized collation and QA review processes as well as continued, systematic digital preservation efforts for returned files. Thanks to the implementation of comprehensive documentation efforts, the U of I was thus able to apply best preservation practices and ensure that future efforts will be much more productive and standardized.

An area of continued exploration is the development of outreach efforts which strike a balance between sustainable, engaging, and unique. It is the opinion of the authors that outreach efforts should be part of the overarching newspaper digitization lifecycle, insofar as that project staff engage with the content from the time it is selected to the time it is ingested and accessible to patrons. Greater awareness and engagement with the content coupled with planning efforts will enable greater visibility and use of the collection. While outreach can prove a unique problem if library staff are unable to fully understand the language of the newspaper thus can nonetheless provide an excellent opportunity to work with students, academic faculty, or engage with the local communities the papers historically serve.

Acknowledgments

References

Arms, Caroline R., Carl Fleischhauer, and Kate Murray, eds. "Sustainability of Digital Formats: Planning for Library of Congress Collections." Library of Congress. Accessed February 28, 2018. https://www.loc.gov/preservation/digital/formats/index.html.

Anip, Erenst R., "Start Spreading the News. A journey in social media usage for outreach in a cooperative digitization project," (2014). University Libraries Faculty and Staff Publications 31 (2014). https://scholarworks.uvm.edu/libfacpub/31

Boddie, Stefan. "Crowdsourced User Text Correction (UTC)." *Veridian Knowledge Base* (blog). June 24, 2014. https://www.veridiansoftware.com/knowledge-base/crowdsourced-user-text-correction/.

Boddie, Stefan. "User Text Corruption." *Veridian Knowledge Base* (blog). November 21, 2011. https://www.veridiansoftware.com/knowledge-base/user-text-corruption/.

Brożek, Andrzej. *Polish Americans, 1854–1939*. Warsaw, Poland: Interpress, 1985.

Brunner, Emily. "Slovenes." The Electronic Encyclopedia of Chicago, accessed March 7, 2018. http://www.encyclopedia.chicagohistory.org/pages/1153.html.

chronam. GitHub repository. Accessed February 28, 2018. https://github.com/LibraryOfCongress/chronam.

Digital Library of Slovenia. "Proletarec." Accessed March 14, 2018. https://www.dlib.si/details/URN:NBN:SI:spr-3YQG3P5E.

Division of Preservation and Access, National Endowment for the Humanities. "National Digital Newspaper Program for projects beginning September 2015" Accessed February 28, 2018. https://web.archive.org/web/20150405031057/http://www.neh.gov/files/grants/ndnp-jan-15-2015.pdf.

Division of Preservation and Access, National Endowment for the Humanities. "National Digital Newspaper Program for projects beginning September 2016." Accessed February 28, 2018. https://web.archive.org/web/20160403181741/http://www.neh.gov/files/grants/ndnp-jan-14-2016.pdf.

Galush, William J. *For more than bread: community and identity in American Polonia, 1880–1940*. Boulder, Colo.: Columbia University Press, 2006.

Granacki, Victoria. *Chicago's Polish Downtown*. Charleston, SC: Arcadia, 2004.

International Image Interoperability Framework. "IIIF Frequently Asked Questions (FAQs)." Accessed March 21, 2018. http://iiif.io/community/faq/#what-is-iiif.

Jaffer, Nabeela. "Michigan Daily - From Newspaper Prints To Digital Archives." *Library Tech Talk* (blog). April 18, 2017. https://www.lib.umich.edu/blogs/library-tech-talk/michigan-daily-newspaper-prints-digital-archives

Jones, Peter d'Alroy and Melvin G. Holli, eds. *Ethnic Chicago: A Multicultural Portrait*. Grand Rapids, Mich.: W.B. Eerdmans Pub. Co., 1995.

Kviklys, Danguole. "'Draugas' The Lithuanian World-Wide Daily Newspaper, 1909–2009." *Lituanus* 55, no. 4 (2009): http://www.lituanus.org/2009/09_4_01%20Kviklys.html.

Library of Congress. "About Chronicling America." Accessed February 28, 2018. https://chroniclingamerica.loc.gov/about/.

Library of Congress "Beyond Words." Accessed February 28, 2018. http://beyondwords.labs.loc.gov/#/.

Library of Congress. "The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants." Modified August 19, 2011. http://www.loc.gov/ndnp/guidelines/archive/NDNP_201113TechNotes.pdf.

Minnesota Historical Society. "Swedish American Newspapers." Accessed February 28, 2017. http://www.mnhs.org/newspapers/swedishamerican.

National Endowment for the Humanities. "Expanding Our Current Scope | NDNP." accessed February 28, 2018. https://www.neh.gov/news/expanding-our-current-scope-ndnp.

National Library of Lithuania. "Draugas. - 1909-." Accessed March 14, 2018. http://www.epaveldas.lt/object/recordDescription/LNB/C1B0001661471.

NDNP Awardee Wiki. "Encoding Non-English Language Text." Accessed March 21, 2018. http://www.loc.gov/extranet/wiki/library_services/ndnp/index.php/Encoding_Non-English_Language_Text. (N.B. Requires NDNP Awardee Account)

NDNP Awardee Wiki. "OCR Accuracy and Quality Review." Accessed March 28, 2018. http://www.loc.gov/extranet/wiki/library_services/ndnp/index.php/OCR_Accuracy_and_Quality_Review.

openoni. GitHub repository. accessed March 21, 2018. https://github.com/open-oni/open-oni.

Pacyga, Dominic A. *Polish Immigrants And Industrial Chicago: Workers On The South Side, 1880–1922*. Chicago: University of Chicago Press, 2003.

Phillips, Megan, Jefferson Bailey, Andrea Goethals, and Trevor Owens. "The NDSA Levels of Digital Preservation: An Explanation and Uses." Proceedings at the Archiving (IS&T) Conference, Washington, DC, April 2–5, 2013. http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf

Sroka, Marek and Tracy Nectoux. "'The Dwindling Legacy That Is Food for Mice and Flames': Discovery and Preservation of Illinois Historic Newspapers through the Illinois Digital Newspaper Project, 2009–2015." *Journal of the Illinois State Historical Society (1998–)* 110, no. 1 (Spring 2017): 87–107. doi:10.5406/jillistathistsoc.110.1.0087.

Sroka, Marek. "'Forsaken and Abandoned': The Nationalization and Salvage of Deserted, Displaced, and Private Library Collections in Poland, 1945–1948." *Library & Information History* 28, no. 4 (December 2012): 272–288.

Rimkus, Kyle R., and Thomas G. Habing. "Medusa at the University of Illinois at Urbana-Champaign: A Digital Preservation Service Based on PREMIS." Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. Indianapolis, Indiana, USA, July 22–26, 2013. http://hdl.handle.net/2142/45232.

Scott, Alyce L. "The Illinois Digital Newspaper Project." *Microform & Imaging Review* 37, no. 4 (2008): 163–169. doi:10.1515/mfir.2008.018.

Skulan, Naomi. "Staffing with Students: Digitizing Campus Newspapers with Student Volunteers at the University of Minnesota, Morris." *Digital Library Perspectives* 34, no. 1 (December 7, 2017): 32–44. https://doi.org/10.1108/DLP-07-2017-0024.

University of Arizona. "Historic Mexican & Mexican American Press." Accessed February 28, 2013. http://www.library.arizona.edu/contentdm/mmap/index.html.

University of Illinois related resources and websites (all accessed April 6, 2018):
American Popular Entertainment Collection: http://go.library.illinois.edu/APEcollection.
Digital Collections: https://digital.library.illinois.edu/.
Farm, Field, and Fireside Collection Libguide: http://uiuc.libguides.com/fff/.
IDEALS: https://www.ideals.illinois.edu.
Illinois Data Bank: https://databank.illinois.edu.
IDNC Facebook: https://www.facebook.com/ILNewspaperProject/.
IDNC Pinterest: https://www.pinterest.com/idncuiuc/.
IDNC Twitter: https://twitter.com/ILNewsProject.
Medusa: https://medusa.library.illinois.edu/.

Veridian Software. "Discovery and delivery platform." Accessed March 21, 2018. https://www.veridiansoftware.com/knowledge-base/veridian/.

Appendices

Appendix 1. List languages supported by "language analysis" search.

Arabic	Galician	Latvian
Armenian	German	Norwegian
Basque	Greek	Persian / Farsi
Bulgarian	Hebrew	Polish
Burmese	Hindi	Portuguese
Catalan	Hungarian	Romanian
Chinese	Indonesian	Russian
Czech	Irish	Spanish
Danish	Italian	Swedish
Dutch	Japanese	Thai
English	Khmer	Turkish
Finnish	Korean	
French	Lao	

Appendix 2. High-level illustration of IDNP workflow.
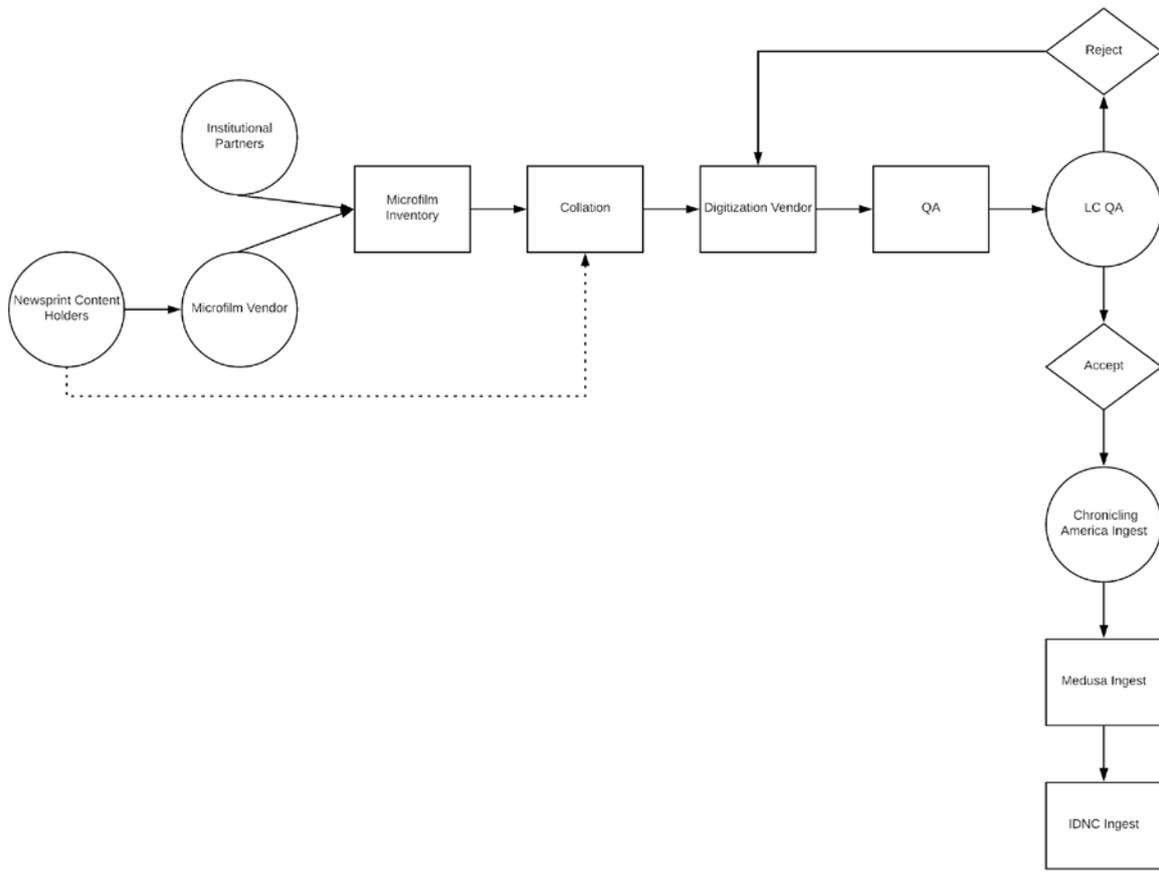
Appendix 3. Medusa file structure for files produced under the 2015–2016 ISL grant.
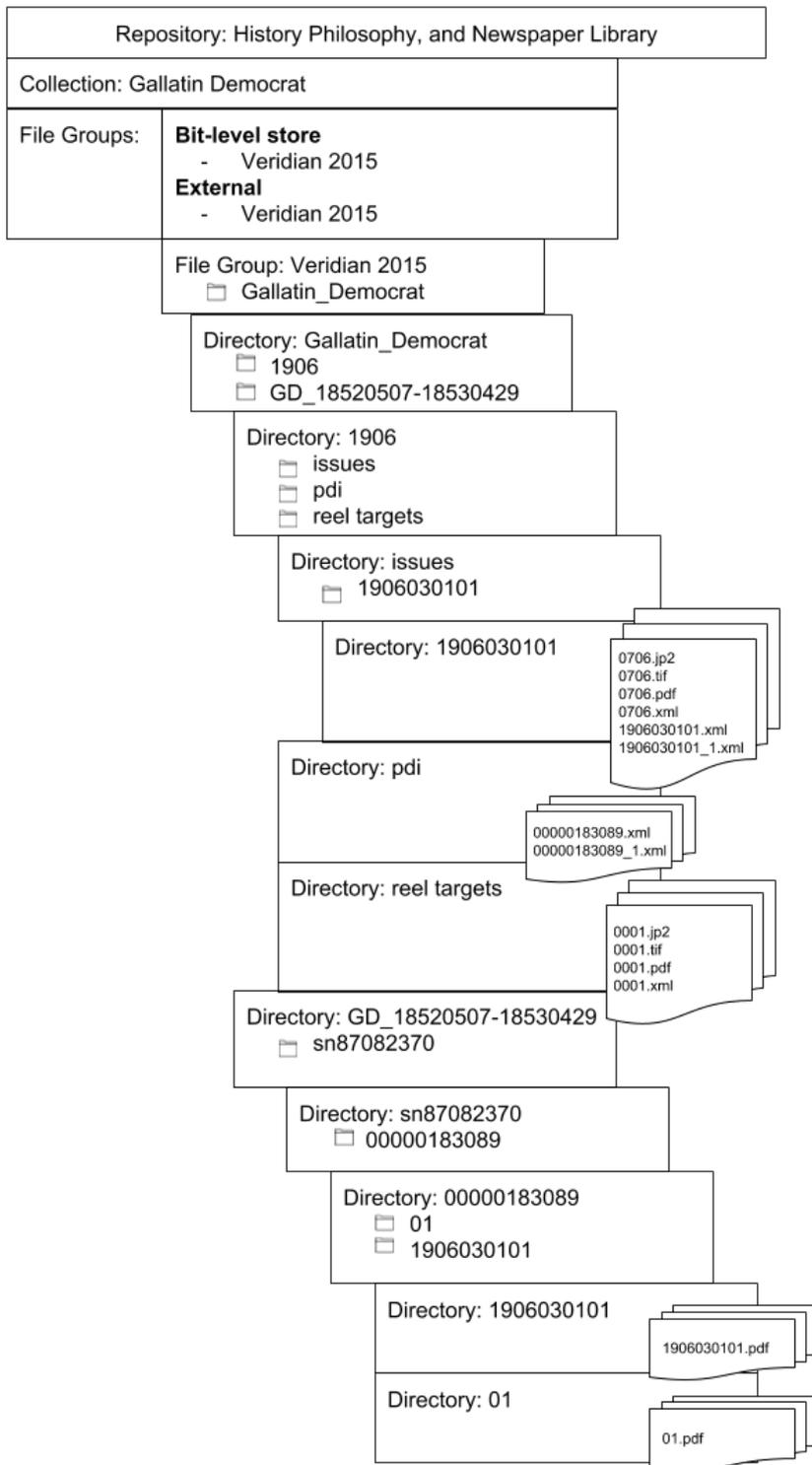
Appendix 4. Medusa file structure for files produced under NDNP (2009, 2011, 2013).
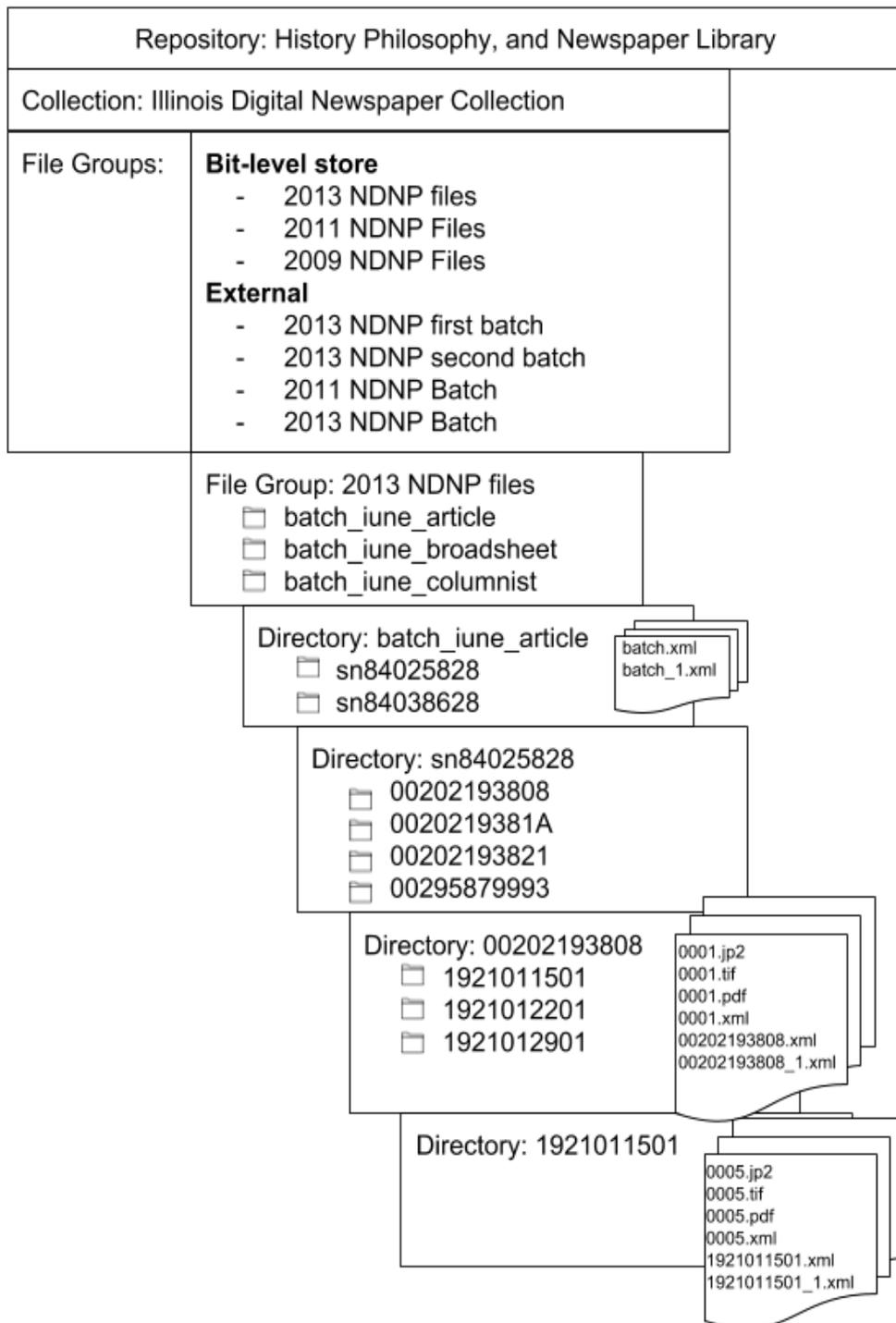
Appendix 5. Medusa file structure under new organization.

Repository: History Philosophy, and Newspaper Library	
Collection: Chicago Eagle	**Collection: Illinois Digital Newspaper Collection**

Collection: Chicago Eagle

File Groups:
- **Bit-level store**
  - content
- **External**
  - 2009 ingest
  - 2013 ingest

File Group: content
- 🗀 issues
- 🗀 reel targets

File Group: issues
- 🗀 1921011501
- 🗀 1921012201
- 🗀 1921012901

Directory: 1921011501

0005.jp2
0005.tif
0005.pdf
0005.xml
1921011501.xml
1921011501_1.xml

File Group: reel targets
- 🗀 00202193808
- 🗀 0020219381A
- 🗀 00202193821
- 🗀 00295879993

Directory: 00202193808

0001.jp2
0001.tif
0001.pdf
0001.xml
00202193808.xml
00202193808_1.xml

Collection: Illinois Digital Newspaper Collection

File Groups:
- **Bit-level store**
  - content
- **External**
  - 2009 ingest
  - 2013 ingest

File Group: newspaper collections pdi
- 🗀 batch_iune_article
- 🗀 batch_iune_broadsheet

Directory: batch_iune_article

originalSubmissionBag
_manifest.txt