**Capture and Preservation of Born Digital News**

*March 5-6*

Universidad Nacional Autónoma de México, México City

**Year 4 of the East View Global Press Archive®: Lessons Learned from a Massive Newspaper Digitization Program**

1. Bryan Benilous

Director Newspaper Products, East View Information Services, Minneapolis, USA.

bryan.benilous@eastview.com

**Barbara Krupa**

Newspaper Digitization Project Manager, Stanford University, Stanford, USA.

krupa@stanford.edu

1. James Simon

Vice President of Collections and Services, Center for Research Libraries, Chicago, USA.

simon@crl.edu

1. Frederick Zarndt

Consultant, Digital Divide Data, New York, USA.

frederick.zarndt@digitaldividedata.com

Copyright © 2020 by Bryan Benilous, Barbara Krupa, James Simon and Frederick Zarndt. This work is made available under the terms of the Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0>

Abstract:

*In 2016, East View Information Services initiated a partnership with Stanford Libraries and the Hoover Institution Library & Archives to digitize and make available 25 million pages of newspapers from around the world, most in Chinese, Japanese and Arabic, as well as Cyrillic languages. This partnership served as the catalyst for East View to greatly expand its own newspaper digitization program, now known as the East View Global Press Archive®. In 2019, East View extended the partnership to include the Center for Research Libraries to make widely available 4.5 million pages of newspapers via a combination of Open Access and consortia-wide collections. The Global Press Archive program will ensure the broadest accessibility to rare vernacular language newspaper content from around the world and further support research in the Digital Humanities.*

*This paper will highlight both partnerships as successful models for supporting the digitization and distribution of massive newspaper collections, and highlight key lessons learned from the program. Topics will include: logistical challenges; development of cloud-based workflows; publishers and copyright; challenges of working with languages in non-Roman scripts, such as Arabic, Chinese and Japanese; and broad accessibility through consortia and Open Access collections.*

**Keywords:** Newspapers, Newspaper digitization, digitization projects, Open Access, Consortia.

**Introduction**

In 2016, East View Information Services initiated a partnership with Stanford Libraries and the Hoover Institution Library & Archives to digitize and make available 25 million pages of newspapers from around the world, most in Chinese, Japanese and Arabic, as well as Cyrillic languages. This partnership served as the catalyst for East View to greatly expand its own newspaper digitization program, now known as the East View Global Press Archive®. In 2019, East View extended the partnership to include the Center for Research Libraries to make widely available 4.5 million pages of newspapers via a combination of Open Access and consortia-wide collections. The Global Press Archive program will ensure the broadest accessibility to rare vernacular language newspaper content from around the world and further support research in the Digital Humanities.

This paper will highlight both partnerships as successful models for supporting the digitization and distribution of massive newspaper collections, and highlight key lessons learned from the program. Topics will include: logistical challenges; development of cloud-based workflows; publishers and copyright; challenges of working with languages in non-Roman script, such as Arabic, Chinese and Japanese; and broad accessibility through consortia and Open Access collections.

**The Stanford Libraries and Hoover Institution Library & Archives Collection**

The Hoover Institution Library & Archives was founded in 1919 with the mission to document “war, revolution, and peace.” Curators were dispatched across the globe to collect documents with a heavy focus on newspapers that served as the first draft of history and a voice of the people. Hoover collected newspapers from over 125 countries and amassed a collection of more than 25 million pages. The collection includes runs of major daily newspapers but also very rare material with the only editions known to exist in Western libraries.

In 2001-2002, Stanford University Libraries (SUL) received about 60% of Hoover Institution Library materials, including the Hoover East Asia Collection. The print newspaper collection was a big part of this realignment. Approximately 2,700 newspaper titles from numerous countries and in multiple languages, mostly dailies and weeklies, were transferred from the Hoover tower to a storage facility in Newark, CA, located 17 miles east of the main campus. SUL staff page the newspapers on user requests twice a week and provide access to them at the Green Library or East Asia Library.

The collection covers East Asia, Middle East (including Turkish newspapers), Africa, Americas, Western/Eastern Europe, New Zealand, and Australia. Most of the titles come from the second half of the 20th century, but there are a few from the end of the 19th century. There are some rare titles from the 1940-1950s from China, endangered titles from Africa, and titles for which the SUL is the sole holder. Some titles, especially African, have very short runs, other have decades of holdings. The newspapers are stored in bundles or archival boxes. Most of them are in good condition.

SUL received the collection with many information gaps. Some titles were not listed in the online library catalog or if listed, provided only minimal information, such as catkey, title, and name of the collection only. Finding a correct record in OCLC and verifying the actual holdings was challenging. Some titles had the wrong location assigned, some had inaccurate date ranges, some were misspelled, some were missing (most probably discarded after microfilming and records were not updated). In 2012-2014 East Asia Library staff undertook a data clean-up project onsite at the Newark facility. Japanese and Chinese titles were sorted, repacked, and recorded. Also, verification of SUL/Hoover microfilm holdings versus print newspapers holdings was completed for all the geographic areas.

In 2016 Stanford University Libraries and the Hoover Institution Library & Archives decided to digitally preserve all newspapers from the Newark storage facility and make them fully accessible to the Stanford community. All digital surrogates of the print newspapers should be available 24/7 to the students, faculty, and the whole community. The scanning and digitization project in partnership with East View started in October to address storage, preservation, and access concerns related to the collection. Barbara Krupa was appointed Newspaper Digitization Project Manager, coordinating work between East View, multiple departments within SUL (Metadata, Data Control, Digital Library Systems and Services), subject specialists/curators, and Hoover staff.

**CRL’s Long History of Newspaper Preservation and Cooperative Collection Development**

The Center for Research Libraries (CRL), founded in 1949 as the Midwest Inter-Library Center, recognized early on the importance of international newspapers. CRL’s international newspaper collection was built initially from newspaper deposits made by member libraries as a space-saving measure on their part. In response to its members’ call for coverage of other world regions, CRL began subscribing to microform editions (and undertaking original filming) of 57 foreign newspapers beginning in 1952, and in 1956 was appointed the administrative home of the Foreign Newspaper Microfilm Project (FNMP). The program was devised to provide for the acquisition of some 100 foreign newspaper titles on microfilm which would then be available for loan or purchase from participating institutions.

CRL also serves as a platform for collective action among libraries, in particular relating to international and area studies resources (chief among them, newspapers). CRL’s long history in supporting collaborative collection development, as well as its capability in storing and serving shared collections, made it a natural home for programs that aggregate librarians’ expertise, collection and store rare material, and make this shared material accessible to participating institutions. Beginning in 1963 with the founding of the Cooperative Africana Microform Project (CAMP), CRL became the administrative home of six cooperative projects that each focused on a different world region: the South Asia Microform Project (SAMP) founded in 1967, the Southeast Asia Microform Project (SEAM) founded in 1970, the Latin American Microform Project (LAMP) founded in 1975, the Middle East Microform Project (MEMP) founded in 1987, and the Slavic & East European Microform Project (SEEMP) founded in 1995. In recent years, all six projects have changed their names to include the word “Materials” rather than “Microform.” This modification reflects the changing collections landscape, which now features more digital collections. All six “AMPs” acquire, preserve and maintain collections of rare and hard to obtain research materials from or about their respective world regions. Each collect annual fees from their institutional members for the acquisition and reformatting of important collections identified by the members. CRL houses these items and makes them available to members on generous terms through interlibrary loan or digital access.

Between these two organizational priorities (newspapers and global resources), CRL has formed an unparalleled collection of international news resources available for research and teaching. CRL’s catalog reflects over 18,000 news records with coverage from the mid-1700s to the present. 25% of the newspaper collection are U.S. general circulation newspapers, covering all states and territories, from colonial times to the present. CRL’s ca. 10,000 foreign titles span from the mid-1700s to the present, with particular strengths in Europe and the U.K., Latin America, Africa, China, Russia, South Asia, and more.

Libraries in North America have placed a premium on electronic access to news content. However, a large share of news produced in countries in the global South continues to be disseminated in print. There is comparatively little being done to capture or convert these resources into electronic format. Since 2006, CRL has worked to address this shortfall, scanning material of interest to its member institutions. As of 2019, CRL had captured more than 300,000 issues from 900 titles as a result of “demand scanning” (volumes or microfilm reels digitized in response to patron requests).

CRL has also pursued strategic opportunities for large-scale digitization of news. In 2008 CRL and Readex (a division of NewsBank) established the World Newspaper Archive project, an ongoing, multi-stage endeavor to systematically digitize and archive world newspapers and news-related materials from the major North American research libraries. This partnership has resulted (to date) in the digitization of more than 3.5 million pages of content, encompassing nearly 500,000 issues from 400 titles from countries in Latin America, Africa, and South Asia. As mentioned above, the Area Materials Projects also have targeted news resources for digitization, supporting centralized and distributed efforts to capture key 19th and 20th century newspapers.

**CRL’s Objectives in Partnering with East View**

Despite these gains, too much of the cultural and historical record of news is still “at risk.” While digitization of late-19th/early-20th century news from Europe, the U.S., and other developed areas continues to grow, the vast majority of digitized titles available are published in the global North, with only a representational amount of content from the global South (and fractional amounts in non-Western scripts) available online.

Moreover, due to copyright concerns, much of the significant content produced in the latter half of the 20th century up through the present is still inaccessible online. In CRL’s news digitization efforts member institutions continually expressed thier interest in access to more recent news resources (filling the “gap” between early 20th century copyright cutoffs and the advent of text-based articles in aggregated databases). Through a century of investment CRL member institutions and other major libraries have aggregated unique collections of news representing a range of viewpoints and diversity of languages – the unlocking of which would open new pathways for researchers, especially in areas of language processing and text mining/digital Humanities.

CRL in recent years has pivoted in its role of contributing resources for the public good. In 2016, responding to a strong endorsement from the community, CRL announced a policy in which all digital materials hosted on the web by CRL that derive from source materials in the public domain or for which CRL has secured the requisite rights and permissions will be available without restriction. The desire by CRL member institutions to create unique digital collections for widespread use became a highly motivating factor in the development of new partnership models. In early discussions with East View on its Global Press Archive program (further details below), CRL members expressed enthusiasm for seeking arrangements that would allow for widest possible accessibility of the materials.

**Digital Divide Data’s Role in Newspaper Digitization and Charitable Work**

Digital Divide Data (DDD) launched in 2001 with a team of ten people in Phnom Penh, Cambodia. Today the company has a staff of over 3000 across Asia, Africa, the Middle East, and North America. While the founder of the company was travelling in Cambodia, he was struck by the mix of poverty and progress he saw in this country which was still recovering from the devastation of the Khmer Rouge regime. While there were computer schools offering training to young people, there were no jobs for the students once they graduated. Recognizing the opportunity to make a difference, and seeing an opportunity to bring India's Business Process Outsourcing model to Southeast Asia, DDD began in a small office in Phnom Penh.

Today DDD delivers content related work to hundreds of international and local clients from operations delivery centers located in Phnom Penh, Vientiane, Nairobi, Gaza, Manila, and Virginia. DDD is the largest technology-related employer in Cambodia and Laos. DDD’s unique [Impact Sourcing](http://www.digitaldividedata.com/impact/impact-sourcing?_ga=1.229508767.262060629.1427736148) (<http://www.digitaldividedata.com/impact/impact-sourcing>) model has developed a growing network of young professionals and propelled hundreds of families out of poverty. Impact Sourcing is a segment of the BPO industry that was pioneered by DDD and has evolved into an economically sustainable approach to alleviating poverty. By employing people in developing economies in the BPO industry, this model provides high-quality digital content services to local and international businesses while creating jobs and developing a workforce capable of competing in the global economy.

In Cambodia, Laos, Kenya, and Gaza, the biggest challenge to undereducated young people is the lack of access to quality jobs in the formal employment sector. Because of low educational attainment and inadequate professional skills, the underserved youth in these countries often find themselves unemployed or employed in the informal sector or gig economy where they typically earn low wages, and have no benefits, job security, or social protection. Moreover, they remain in an intergenerational poverty loop that keeps them from escaping subsistence living. In response, through DDD's social impact model, youth enter formal employment at DDD where they are trained to deliver commercial BPO services, including digital content creation, data preparation, machine learning, and cloud computing services to global clients. At the same time, youth are provided with professional training and skills development and opportunities (loans, scholarships) to pursue higher education. DDD’s approach to Impact Sourcing is based on a comprehensive program of training, employment and higher education that enables young people to identify and attain their goals. The structured program that DDD offers enables youth to develop their skills, grow personally and improve the socio-economic well-being of their families. Over the past 18 years, our model has trained and supported over 6,000 underserved youth and increased their lifetime earnings by an estimated total of more than USD $350 million.

DDD has been recognized for industry leadership and commitment towards human capital development by being awarded the Skoll Award for Social Entrepreneurship and The International Association of Outsourcing Professionals (IAOP) includes DDD as a Rising Star in its Global Outsourcing 100 list in 2015 and 2016.

DDD is regarded as a leading BPO provider in delivering digitization and preservation services for libraries, private and public archives with rare collections, research universities and museums around the world. These services unlock and extend the value of data, documents, publications and archives, making them searchable and accessible online, on mobile devices, or in any information system. This is a core competency of DDD and we currently process (digitization and metadata creation) more than 500,000 pages per month.

DDD is incorporated in the USA as non-profit 501(c)(3) corporation to facilitate the creation of sustainable jobs and educational opportunities in emerging market countries. The USA nonprofit corporation owns for-profit companies in Cambodia, Laos, Kenya and the USA and collaborations closely with similarly organized social impact companies in Gaza and Manila. A 12-member board of directors provides oversight. DDD had revenues of more than $18,000,000 in fiscal year 2018.

The senior operations team at DDD possess well over a decade of expertise in scanning, metadata tagging, and digitization services and prior to working at DDD, have worked for firms in the USA, India, Singapore and the Philippines.

DDD’s largest office is located in Phnom Penh and is located at No. 559 Street 271. The current capacity of our 7-year old, 6-story modern building is approximately 500 seats, with 4 of the 6 floors utilized for production today. This can be expanded to 725 seats in the same building. Newspaper digitization is one of DDD’s core services; DDD has been involved in this type of work since its beginning.

Also, DDD partners with libraries and museums in Africa and Asia to preserve and safeguard their collections by providing affordable solutions to digitize, preserve, and access valuable and ‘at-risk’ materials. By working on such projects, local youth at DDD not only develop relevant skills and enhance their employability, but also engage with their culture and preserve their country's heritage for generations to come. Landmark projects undertaken by DDD include digitizing ‘at risk’ collections at the National Museums of Kenya and the Tuol Sleng Genocide Museum in Cambodia with its large collection of written and photographic records from the Khmer Rouge.

DDD has been recognized worldwide for quality service, innovation, and social impact. Among many awards and recognitions, the Global Sourcing Council presented DDD its 3S Award for Sustainable and Socially Responsible Sourcing. DDD has been consistently on the list the Top 100 NGOs worldwide in The Global Journal and is a recipient of the Google Award for Innovation in Business Process Outsourcing (BPO).

DDD is a signatory of the Lyon Declaration (<http://www.lyondeclaration.org/>) and supports the UN Global Compact Initiative (<https://www.unglobalcompact.org/>). DDD supports the American Library Association (ALA), the International Federation of Library Associations and Institutions (IFLA), and other similar organizations; DDD participates in ALA and IFLA conferences.

**Transformation of East View’s Newspaper Program into the Global Press Archive**

For over 30 years, East View Information Services has worked to bridge the gap between publishers and libraries and between East and West to bring to light “Uncommon Information from Extraordinary Places” in support of academic research. In doing so, East View has helped balance viewpoints in academic research otherwise dominated by Western sources.

Particularly with regard to digital newspaper archival products, the majority of aggregator publishers have focused on English- and Western-language content from major publishers. The first two major newspaper digitization projects were for the *Times* (of London) and *The New York Times*. Both are significant newspapers that are justified to be prioritized for digitization. Both, however, represent a Western viewpoint on global events. Additional digitization programs (both commercial and Open Access) continued to focus on Western-language content due, primarily, to the ease of digitization but even more so to demand and funding from libraries in North America and Europe. As discussed above, even today some of the largest commercial newspaper digitization programs include very little, if any, content in non-Roman scripts. Current commercial collections from East Asia focus on English-language titles from China, Korea and India—major English-language regional papers or historical missionary papers that still reflect a Western viewpoint.

East View recognized early on the gap in available research content from the East—particularly from Eastern Europe, Russia and the former Soviet Union, East Asia and the Middle East. East View’s first digital newspaper archive was *Pravda*, the official paper of the Soviet Union and the most significant paper in Russia today. East View continued to leverage its expertise in digitizing Cyrillic script with other major papers from Russia and Eastern Europe, such as *Izvestiia* and *Literaturnaia gazeta*. East View also distributed digitized newspaper content from partners in East Asia including such prominent titles as *Jiefangjun Bao (PLA Daily).* In doing so, East View provided critical Eastern viewpoints to research of the Cold War, Post-Cold War and other critical topics both international and domestic.

In 2016, East View partnered with Stanford Libraries and the Hoover Institution Library & Archives to provide a digital solution for the entire off-site collection of newspapers collected by Hoover throughout the 20th century from over 125 countries and totaling over 25 million pages. The partnership, intended to take 10 years to complete, includes identifying available digital surrogates, then digitizing and commercializing the remainder. The astonishing cost of digitizing, wholesale, the entire collection and the fact that the vast majority of the material remains in-copyright, requires a commercial aspect to fund the program and get appropriate buy-in from publishers to secure rights.

This corpus of materials allowed East View to reimagine its newspaper program and invigorate it with a mission to become the largest international digital newspaper program in the world. This program was branded the East View Global Press Archive®. Building on more than 20 titles already digitized by East View prior to the partnership with Stanford and Hoover, East View accelerated its commercial newspaper digitization program with compelling titles from the Stanford-Hoover collection including *Moscow News, The Rafu Shimpo, Gudok* and *Novoe russkoe slovo*. At the same time, East View recognized that the breadth of content required additional partnerships to ensure a robust digitization program that supported the cost of digitization and, more importantly, supported research through broad accessibility of content.

Almost immediately after announcing the partnership with Stanford and Hoover, East View began close collaboration with the Center for Research Libraries (CRL) on further developing the Global Press Archive program. Partnering with CRL addressed three key challenges with the program: 1) access to more cost-effective surrogate and complementary material through decades of microfilming initiatives by CRL and its large member community, 2) broad distribution of the completed material to the academic community through the sponsorship of both Open Access and consortia-wide collections, and 3) funding for digitization of rare material through cost-sharing across CRL’s membership community to support digitization of material that otherwise did not have sufficient market demand. Both Stanford and CRL also receive digital archival copies of the material in multiple formats including METS/ALTO XML, JPEG2000, searchable PDF and TIFF files. This ensures long-term digital preservation of all the material at multiple locations across the U.S., including Stanford, CA, Chicago, IL, and Minneapolis, MN.

While the broader agreement with Stanford is focused on the full collection of 25 million pages over 10 years, the agreement with CRL is narrower in focus and based on funding abilities of the member institutions with a commitment to digitize 4.5 million pages over three years. This CRL “Charter Alliance” is viewed as a pilot stage with the hope from all sides that the Alliance will extend beyond this initial phase and not only support the complete digitization of Stanford’s archive but also support the digitization of the extensive microfilm and print holdings of CRL and its members. The program is designed to support what the community can bear at any given time but has the ability to scale in annual output and extension for further years.

CRL and East View listened closely to the desires of CRL members and the broader academic research community when designing the CRL Alliance program. There was much interest in 1) making as much material as possible available as Open Access collections free to the world, 2) not forcing purchasing decisions on member institutions and ensuring equitable pricing based on institution size, 3) ensuring CRL and its members are deeply involved in the scope and selection of content to be digitized, and 4) focus on long-term preservation and support for Digital Humanities research.

As such, the program was designed to build nine collections over three years, each roughly 500,000 pages in size with CRL members funding the initiative for the benefit of the global academic community. As much content as possible (six of the nine collections, totaling 3 million pages) will be presented as Open Access. The remaining three collections will consist of in-copyright material and be made available to all CRL member institutions and commercially available to non-CRL members. The commercial collections are necessary to extend the scope of the program to include copyrighted content that by definition requires a commercial aspect to attain rights from active publishers. The program was designed with an opt-in model in which institutions volunteered to commit to the program and allowed for less than 50% of the CRL community to fund the program for all CRL members. This ensured that smaller institutions and those whose funding is strained still benefit from the program. Additionally, a CRL advisory committee was formed to advise on the program and help prioritize collection development. The advisory committee also coordinates with subject matter experts in title selection for the individual collections. As noted previously, all digitized files are delivered to CRL for long-term preservation and support of Digital Humanities research.

The Global Press Archive and the CRL Alliance not only transformed East View’s commercial newspaper digitization program, it represented a new model for collaborative digitization programs, one that ensures the broadest access to rare materials, including Open Access for out-of-copyright orphaned material, in a cost-effective model that leverages the coordinated purchasing power of large consortia and brings librarians and academics into the deliberative process of product development in support of long-term preservation and Digital Humanities research.

Such an ambitious project poses unique challenges including managing logistics, ensuring quality digitization, engaging with publishers and ensuring long-term viability of the program. As the Global Press Archive enters its fourth year in partnership with Stanford and Hoover and concludes the first year of the CRL Alliance, further insight is gleaned through implementation and is addressed below.

**Lessons Learned: Logistics**

There are many logistical challenges involved in a massive newspaper digitization project such as the Global Press Archive, including: metadata, shipping and handling, and sharing digital output.

The content from the Stanford-Hoover collection was collected and cataloged over the course of a century. In that time, cataloging rules changed particularly with regard to transliteration of vernacular language titles. Other key metadata also changed over time including the names of cities. This resulted in much duplication of title data. In some instances, key metadata was missing and there were no uniform identifiers to align duplicate title entries. As East View coordinated with other partners, metadata became even more of a challenge as other institutions used different transliteration systems and city names. Publication titles presented further challenges: just as in Western cultures where newspapers share similar names (e.g., *Daily News*, *The Tribune*, *The Times*), non-Western newspapers also have very similar names (e.g., *Akhbar* across the Middle East, *Li Bao* in China, etc.). Are these truly different titles or were they just duplicated in cataloging? Were these related titles published by the same publisher in different cities? Such issues have major implications when determining rights. In such cases, East View utilized catalog experts in those languages to validate that the titles were unique and research the publisher. Ultimately, East View spent six months researching and cleaning the data to merge duplicate titles, correct and improve the metadata and, in the process, perform due diligence on publisher rights and identify additional complementary holdings.

Despite these challenges, the clean-up of metadata is fairly painless compared to the physical requirements of shipping rare, fragile print material across the country. East View estimates that the Stanford-Hoover collection is over 25 million pages of print material, most of which is stored in over 25,000 wrapped bundles of newspapers, with more content stored in thousands of archival boxes. Each bundle weighs roughly 10-15 pounds, with the entire collection weighing easily well over 250 tons. Along with the physical toll of moving newspapers to scanning facilities, there is also much work to inventory the material and confirm nothing is lost or damaged in shipping. Shipping costs for such a large amount of content can also greatly impact the budget.

Early on, East View recognized this challenge and performed a great deal of research to identify potential surrogate material in digital or microfilm formats. Fortunately, millions of pages were digitized already as Open Access collections or in commercial databases. These titles were deprioritized for digitization. Millions more pages existed in microfilm with good enough quality to be surrogates for scanning. The CRL Alliance granted East View access to still more surrogate material from the extensive collections of CRL and its member institutions. There is still a large amount of material that requires scanning from print, whether that is to fill gaps in microfilm holdings, to deliver higher quality output for poor quality images from degraded microfilm, or, most importantly, to scan rare print material that has never been microfilmed or digitized and where Stanford-Hoover holds the only known copies in North America, if not globally.

The logistical challenges do not end with the shipping of materials. East View digitizes all the newspapers to uniform standards based largely on those developed for the U.S. Library of Congress National Digital Newspaper Program (NDNP). The output includes high quality 400DPI color TIFF images, METS/ALTO XML files, JPEG2000s and searchable PDFs—ultimately exceeding petabytes of data. Along with hosting the data for academic research, East View has committed to long-term preservation of all these files and is sharing copies with contributing institutions including Stanford and CRL.

East View works closely with partners that are experts in scanning, digital conversion and platformization. As such, work on a single newspaper travels the world. Originating in Stanford, CA, Chicago, IL, or other partner institutions across the U.S. and as far away as China or Japan, material navigates to Minneapolis, MN and work is supported in Moscow, Kyiv, East Brunswick, NJ, Phnom Penh, KH, and Wellington, NZ, with final files delivered back to Minneapolis, Stanford and Chicago.

With the program in full speed, East View is digitizing over 500,000 pages a quarter. The time to load and ship hard drives around the world cannot keep up with this pace, so East View, with much support from DDD, CCS docWorks, Amazon Web Services, and other partners, developed the first fully cloud-based solution. Materials scanned in Minneapolis, East Brunswick, Kyiv or Moscow, or delivered from a partner or publisher globally, are loaded onto Amazon S3 “buckets” that automatically feed the material into DocWorks stations managed by technical experts in Phnom Penh. Final output is loaded into another Amazon “bucket” and available for QC review by experts at East View. Upon completion of QC checks, files are loaded onto platforms in Wellington, NZ, and moved to other “buckets” shared with Stanford and CRL. Clean accurate metadata, including detailed information on contributing institutions, are used to ensure each partner receives the right content.

**Lessons Learned: Digitization Workflow and OCR**

In spring 2017, East View approached DDD about newspaper digitization for its collaboration with Stanford Libraries and the Hoover Institution Library & Archives. After a small pilot project with input and output data delivered on hard drives, it became apparent that data delivery on hard drives would not scale. Our thoughts turned to the cloud.

DDD has been a user of Content Conversion Specialists (CCS) docWorks software since 2007 and used docWorks for the East View pilot as well as many of its other newspaper digitization projects. East View has long and extensive expertise with Amazon’s AWS. CCS has supplied traditional data center-based digitization workflow software to libraries and service bureaus around the world since the early 2000s but never as a cloud service. Between the three companies, we held our collective breaths and began a pilot workflow using AWS services exclusively.

Held our breaths because a digitization project of this scale had, to our knowledge, never been done in the cloud exclusively. Input images are sometimes up to 400 MB in size with the derivative output data about 2.5x the size of the input images. Moving such large quantities of data into, around, and from the cloud was not usually done, especially when the partners are literally located around the world from Minneapolis to Cambodia.

To our relief, creating a cloud-based workflow with CCS docWorks software on AWS services was trouble-free and uncomplicated. After a short test period, we scaled up the system to full production capacity; this also proved to be easy and fast.

Both AWS and CCS docWorks offer per-use cost models, which keeps fixed costs and investments very low and costs only for capacity used. A major component of the cost proved to be the AWS cost for file storage. The amount of storage needed obviously determined by the time between document are upload and the acceptance of the processed results.

The Stanford-Hoover collection has content from over 125 countries; CRL’s collection and those of CRL contributing members likewise includes content from many countries. The content of these collections is in a variety of languages and scripts, for example, Latin alphabet languages using both Antiqua and Fraktur fonts, traditional and simplified Chinese, Japanese (Hiragana, Katakana, Kanji with and without Furigana), Korean, Russian, Spanish, Arabic, Urdu, Pashto, and many others. There is no shortage of content for which image quality will not be good because the originals have deteriorated or because the microform was poorly made or has deteriorated. For some content, even with exceptional image quality, OCR results will fall short of what one would expect because OCR software for some languages has not reached the accuracies that it has for Western, Latin alphabet languages.

Newspaper formats also present unique challenges. Those familiar with newspapers know that page formats vary widely and, even for the same title, change often. And page formats present other challenges, too, for example, Japanese newspapers are mostly written top-to-bottom but there may be left-to-right text for headlines or illustrations. Article titles in Japanese newspapers may not be at the beginning of an article but somewhere else.

Arabic scripts are particularly formidable because the characters are connected and because some characters have several different forms depending on the character position within the word. Arabic has different diacritics (i'jam, tashkil, harakat, hamza) which are often omitted in modern newspapers but not in classical Arabic works and historic documents. Arabic has many dialects with words particular to each. Furthermore, several languages are written using Arabic script, maybe extended with characters specific to the language, for example, Urdu, Pashtu, Persian, Ottoman Turkish, and Punjabi to name but a few. Some of these languages have no OCR software.

For languages with no OCR software, it is possible to train OCR software – open source Tesseract comes to mind, however, creating training sets for the language is labor intensive and therefore expensive. The Open Islamicate Texts Initiative (OpenITI - <https://iti-corpus.github.io/>) addresses this for some Arabic script languages but certainly not all. OCR for the Tamil language has been on several cultural heritage organizations wish list for many years, but, so far as we are aware, no reliable Tamil OCR yet exists.

Image quality from poorly made or poorly preserved originals or microform hasn’t been mentioned, not because these materials are all in pristine condition – they aren’t! - but because image quality problems have been addressed for other languages and are similar for any language regardless of character set. Members and collaborators in IMPACT project – now the IMPACT Centre of Competence (<https://www.digitisation.eu/>) - built a number of tools, test sets, benchmarks, and lexica for Latin character languages.

**Lessons Learned: Publishers and Copyright**

The Stanford-Hoover collection includes mostly 20th century content from over 125 countries. This poses significant challenges for digitization and distribution as the majority of the content remains in copyright. While there is the capability to digitize some content exclusively for the contributing institution as preservation reformatting, the ultimate goal of the program is to make the material widely available for academic research. Beyond legal risk, East View respects the rights of copyright holders. Moreover, librarians as stewards of copyright expect aggregated database providers to secure rights for copyrighted material. This is precisely the key role that database providers such as East View play in bridging the relationship between publishers and the academic community.

East View commissioned research on international copyright law at a global scale and by individual country. Copyright rules vary based on where and when the publications were produced but also where the distributor is located and what countries the material is being sold into. With over 125 countries and even more global distribution, the task or researching and securing copyright is one of the greatest challenges to the Global Press Archive program. While most countries have adopted the Berne Convention, many have implemented the rules differently and signed additional agreements that go beyond the initial 1886 treaty. Additionally, the year that the individual country signed onto the treaty also impacts how rights are protected prior to and after that date. International copyright continues to evolve with new treaties such as those signed among members of the European Union as well as those signed as part of the United States-Mexico-Canada Agreement. Therefore, copyright is a constant concern and a quickly evolving landscape.

Ultimately, every title—over 2,000 from Stanford-Hoover and over 10,000 from CRL—needs to be independently researched to determine what protections are in place. Many titles are straightforward, having ceased publication a century ago with no succeeding titles, and can be distributed as part of Open Access collections. Many more are titles that clearly continue to be in publication, making it possible for East View to negotiate rights from the publisher in exchange for royalties earned from commercial distribution. There is however a very complicated middle ground of “orphaned” titles where the publication ceased and there are no known surviving owners. This is especially true for rare, oftentimes short-run publications that spoke out against the ruling government or ended up on the losing side of a civil war. This is probably the most critical material for academic research and requires deep consideration of balancing the risk of infringing rights versus the value to research. East View ultimately makes such decisions on a case-by-case basis with deep diligence into determining rights and, if necessary, compensates publishers for the digitization of ceased publications, such as *Novoe russkoe slovo* and *Moscow News*.

**Lessons Learned: Funding and Access**

As the above logistical, OCR, and copyright challenges clearly demonstrate, massive newspaper digitization programs cost massive amounts of money. And, yes, even Open Access collections cost money. This makes funding one of the greatest challenges to digitization. Many aggregated database publishers focus on the most widely circulated current titles that have the greatest profit opportunities or focus on out-of-copyright material that does not require publisher royalties and ensures greater profit margins.

The above approaches greatly limit academic research and skew results toward popular Western viewpoints that ignore more than half the world, with massive gaps in available research material covering the bulk of the 20th century. East View aims to correct this imbalance through partnerships to not only fund a massive newspaper digitization program but also make it as widely available to academic audiences as possible.

By partnering with CRL, East View has committed to making large amounts of out-of-copyright orphaned material available to the world. CRL members have graciously agreed to fund the digitization of 3 million pages of content between six Open Access collections. As part of the broader CRL Alliance program, East View is also digitizing 1.5 million pages of in-copyright content to be made available to all CRL members, regardless if they committed to the funding of the program.

Individual publishers will also drive the availability and bundling of content. Publishers expect to get royalties for the distribution of their copyrighted content and in many instances rely on that extra revenue to stay afloat. Many publishers have a high expectation of their worth and demand large continuous royalty payments. They also are often hesitant to agree to be included in larger collections with potential competitors. This requires many titles to be delivered as stand-alone commercial products. Ultimately it is the publisher that defines the distribution model for their content. None will permit their material to be provided for free to the world, a few may agree to be included in broad collections distributed through consortia but most will expect to have their content distributed as a single product with stature above the rest.

**Conclusion**

This paper demonstrates that deep partnerships with commercial and academic entities are required to support massive newspaper digitization programs. Commercial vendors like East View play a critical role in bridging the gap between academic institutions and publishers. Moreover, commercial vendors such as Digital Divide Data and Content Conversion Specialists also bring expertise and support for the logistical aspects of such a program. Ultimately, it is academic institutions that hold material and fund digitization programs, either through individual institutional purchases, consortia-wide cost sharing or funding of Open Access collections. Individual academic institutions, in most cases, lack the funding and logistical skills to undertake a massive program like this by themselves and have limited ability to secure broad distribution rights with publishers, with the end result being either limited accessibility to just their institution or a restricted focus on only out-of-copyright material. Only through close collaboration among individual institutions, consortia and commercial vendors like East View can such a grand program as the Global Press Archive succeed. The future of this program, with the support of its many partners, has much more to share and will continue to transform academic research for decades to come.

Acknowledgments

The authors would like to thank the hundreds of librarians and collection development specialists that spent decades collecting, preserving and cataloguing the world’s newspapers. The authors would also like to thank those in the past, present and future that fund preservation efforts to support research.

References

Stanford Libraries <https://library.stanford.edu/>

Hoover Institution Library & Archives <https://www.hoover.org/library-archives>

Center for Research Libraries <http://www.crl.edu/>

East View Global Press Archive <https://www.eastview.com/resources/gpa/>

Digital Divide Data <https://www.digitaldividedata.com/>

CCS Content Conversion <https://www.content-conversion.com/>

IMPACT Centre of Competence <https://www.digitisation.eu/>

Open Islamicate Texts Initiative OpenITI <https://iti-corpus.github.io/>)