Catholic, Crowdfunded, and Collaborative: A Unique Approach to Newspaper Digitization

Patricia Lawton
Catholic Research Resources Alliance, Notre Dame University, Notre Dame IN, USA.
plawton@nd.edu

Nicholas A. Casas
Center for Research Libraries, Chicago IL, USA.
ncasas@crl.edu

Jeff Moyer
Reveal Digital, Saline MI, USA.
jmoyer@revealdigital.com

Frederick Zarndt
Digital Divide Data and IFLA Governing Board, Coronado CA USA.
frederick@digitaldividedata.com, frederick@frederickzarndt.com

Abstract:

Like other newspaper digitization program, the Catholic Newspapers Program (CNP), initiated by the Catholic Research Resources Alliance (CRRA), aims to provide access to newspapers, in this case to all Catholic newspapers published in North America, both newspapers in the public domain and in copyright. But the CNP takes a very different approach to implement its goal.

In collaboration with Reveal Digital, Digital Divide Data, CRRA's forty-four institutional members, and thirty digitizing partners, CRRA has developed a cost model and project plan that focuses on five major cost elements including rights clearance, data conversion, hosting and delivery platform, *project management and outreach. Unlike most newspaper digitization programs, CNP is not grant-funded, but crowd funded. According to Peggy Glahn, Reveal Digital's Program Director, the library crowdfunding model "challenges the traditional approach to scholarly publishing. It requires librarians to think more like active investors and publishers and less like consumers."*

This paper focuses on the following five cost elements and related matters.

1. *Outreach and marketing: how to design a crowdfunding campaign for a newspaper digitization project.*
2. *Newspaper collection acquisition and rights clearance: cooperation and collaboration of CRRA members.*
3. *Digital collection features and technical requirements: an overview of the reasons for the features and requirements and the choice of digitization services provider (Digital Divide Data)*
4. *Reasons for the choice of a cloud-hosted delivery platform and for the choice of Veridian as the platform.*
5. *Metadata model for the Center for Research Libraries ICON newspapers database.*
6. *Reasons and decisions made to ensure CNP's long-term economic sustainability.*
7. *Long term digital preservation considerations.*

About CNP key partners:

CRRA (http://www.catholicresearch.net/) was established in 2008 and is a non-profit membership organization of forty-four Catholic colleges, universities, archdioceses, seminaries, and religious congregations collaborating broadly to deliver projects and services in support of its mission to provide enduring global access to Catholic research resources in the Americas.

Reveal Digital (http://www.revealdigital.com/) was founded in 2013 and works in partnership with libraries in order to bring unique content into the digital world using a cost-recovery business model. The cost recovery approach ensures that resulting collections are affordable, relevant, and sustainable for libraries.

Digital Divide Data (http://www.digitaldividedata.org/) was established in 2001 in Cambodia and has since established operations offices in Laos, Kenya and the US. DDD delivers Business Process Outsourcing solutions to clients worldwide. Customers receive high-quality competitively priced services. At the same time, DDD's innovative social model enables talented youth from low-income families to access professional opportunities and earn lasting higher income. This model, established by DDD in 2001, is now called "Impact Sourcing" and has been implemented by dozens of firms around the world. DDD is a signatory of the Lyon Declaration (http://www.lyondeclaration.org/) and supports the UN Global Compact Initiative (https://www.unglobalcompact.org/).

Keywords: collaborative, crowdsourcing, rights clearance, outreach, marketing, Impact Sourcing.

1. **Introduction**

While no online union list of Catholic newspapers currently exists, collaborative efforts combined with advances in technology may facilitate such a listing in the future.--Charlotte Ames, University of Notre Dame, 1997. -- (http://archives.nd.edu/cathnews/cathnint.htm)

Almost twenty years later, Charlotte Ames' vision may be realized, and extended. Advances in technology have made it possible for a network of Catholic archivists and librarians to build a portal of Catholic resources and then to strategize and begin construction of a comprehensive Catholic Newspapers Directory, thus fulfilling Charlotte's vision of a union list of Catholic newspapers. But wait, there is more. In line with the CRRA mission to "provide enduring, global access to Catholic research resources in the Americas," CRRA is also building a Catholic News Archive of aggregated digital Catholic newspapers on the robust Veridian platform. With no funding and just two full-time staff, CRRA is realizing its goal for providing enduring access to Catholic newspapers through a collaborative and supportive network of members, archives, publishers, vendors, sourcing partners, and more. CRRA embraces and makes manifest the African proverb, "If you want to go fast, go alone. If you want to go far, go together."

2. **The project**

About CRRA

Established in 2008, the Catholic Research Resources Alliance (CRRA) is a nonprofit membership organization of forty-four Catholic colleges, universities, seminaries, archdioceses and religious congregations collaborating broadly in support of its mission to provide enduring global access to Catholic research resources in the Americas. CRRA has two primary projects in advancing this mission: the Catholic Portal and the Catholic Newspapers Program. The Catholic Portal provides access to and discovery of rare, unique and uncommon Catholic scholarly resources. In 2011, CRRA launched the Catholic Newspapers Program (CNP) "to provide access to all extant Catholic newspapers in North America." CRRA accomplishes its goals through a distributed and highly collaborative organization, coordinated by just two full-time staff.

Catholic Newspapers Program

The CRRA Catholic Newspapers Program (CNP) has four primary component activities:

1. Make discoverable records describing full runs of North American Catholic newspapers in the Center for Research Library (CRL) International Coalition for Newspapers (ICON) database
2. Digitize selected Catholic newspapers identified as top priorities ("priority papers")
3. Create the Catholic News Archive on the Veridian delivery platform, a robust platform of aggregated Catholic newspapers
4. Preserve digital Catholic newspapers through MetaArchive.

CRRA conducted an environmental scan conducted in 2011 to determine what Catholic newspapers existed or had once existed and where they were held. The scan yielded the identification of 857 titles, with full runs broadly scattered. In many cases, only pieces of the years were discovered; piecing together full runs would prove to be a challenge. Only a handful of the identified papers were in digital form, and those were largely held behind a paywall. To track and make available Catholic papers digital form, CRRA posted to the website a listing of Catholic newspapers which has grown considerably and continues to grow, thanks to user suggestions. In 2014 the *Catholic News Online (CNO*), consisted of approximately seventy-five titles. As of March 18, 2016, CNO identifies more than 220 titles.

Building the Catholic News Archive and Digitizing Priority Papers

Within the CNP, the CRRA has developed an ambitious 2.6 million dollar plan to 1) implement a Catholic News Archive on the Veridian platform, 2) to digitize and 3) preserve 1.5 million pages of Catholic newspapers from local and national "priority papers." The 1.5 million pages will seed the Archive, which will continue to grow through future digitization projects and from dioceses and publishers alike interested in sharing digital Catholic newspapers within a robust platform of aggregated content.

The priority papers include diocesan newspapers from Chicago, Hartford, Miami, New Orleans, New York, Philadelphia, Pittsburgh, San Francisco, and St. Louis; the national perspective is represented by the *National Catholic Reporter* and the Catholic News Service (CNS) newsfeeds, the Catholic equivalent of Reuters. The CRRA will make available all years of these priority papers, from inception to 2013. Open access is a core CRRA value and the aim of this project.

Why Catholic?

Why Catholic papers? Scholars cite extensive use of Catholic newspapers which are relevant to a broad array of disciplines and topics, including immigration and adaptation to new environments, formation of local charities, social justice, development of school systems and hospitals. Diocesan papers focus on parish life within a particular community and are therefore a unique resource for local historical information and stories about people -- parishioners, school children, lay teachers, priests, nuns, and the communities in which they lived.

In addition, these papers are not easily accessible nor discoverable elsewhere.The newspapers are difficult to access: issues are scattered among a variety of locations, including church libraries or small archives; knowledge of where these newspapers are located is often local; the full extent of collections is not represented in major bibliographic databases such as *OCLC WorldCat*; and few papers are available in digital form. In many cases, Catholic newspapers are owned by Catholic dioceses and congregations that may not have access to bibliographic tools such as *WorldCat*. Although massive in scope, the Library of Congress' *Chronicling America* listed just forty percent of the 857 U.S. Catholic titles identified in the 2011 scan. Many papers may be under-described or simply not described at all. In a word, many are hidden.

Pilots and Phases

As a first digitization project, 1.6 million pages were soon deemed not possible for financial and practical reasons. CRRA and its various committees and partners then identified logical slices, or phases, of the project: Civil War years (1861-1865); European Immigration (1880-1919); Pre-1924; and Vatican II years (1958-1972). Each phase represents selected years from across the priority papers, from inception to 2013. The phases represent periods of scholarly interest and involve participation from multiple institutions holding a mix of source material in print and/or microfilm. The Vatican II years also involve gaining permissions since, these materials are yet under copyright.

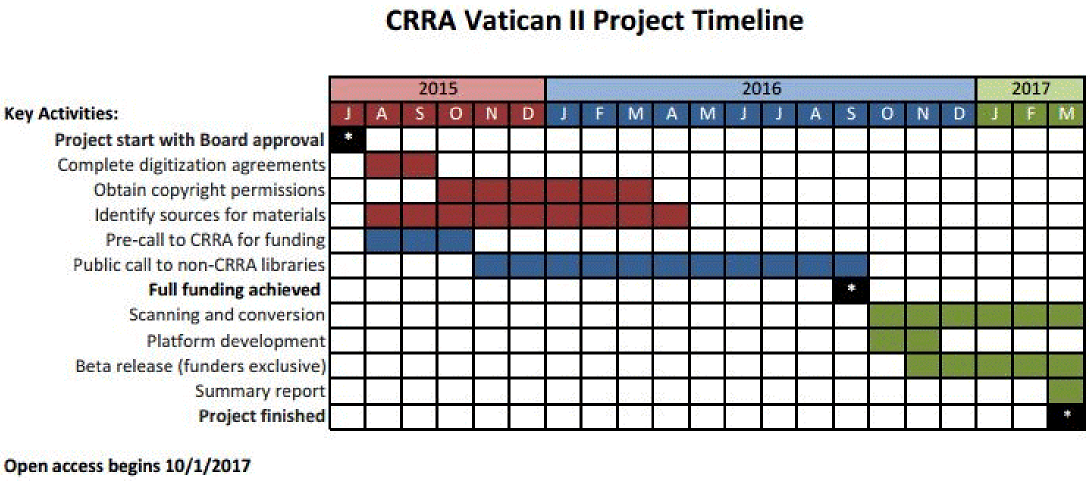
Focusing on one more small, manageable “pilot” projects has been essential for developing workflows, determining costs, establishing timeframes, developing specifications, identifying partners and vendors, and especially, to attract funding. Pilot projects provide the opportunity to establish and test a framework ensuring a viable, functional framework for the larger digitization project as well as for providing prospective donors a proof of concept.

Current efforts focus on the Vatican II years (1958-1972). This phase has an estimated cost of $419,644 and CRRA is working with Jeff Moyer of Reveal Digital in implementing a unique crowdfunding model to fund the project, (described later in this paper).

The Vatican II years were selected due to their importance to scholars as well as the fact that they are under copyright. CRRA would need to face the challenge of obtaining permissions, so why not now? As a proof of concept, CRRA is able to demonstrate the deep support of its community in providing source material *and* in signing legal agreements granting permissions to make the images available. Gaining permissions is often one of the most challenging aspects of digitization projects and why CRRA wanted to take it on.

The years 1958-1972 represent the years preceding and following the convening of the Second Vatican Council (also known as “Vatican II”) from 11 October 1962 – 8 December 1965. Vatican II ushered in significant changes within the Catholic Church aimed at greater dialogue with the modern world. Visible and memorable changes brought about included greater lay participation in the liturgy, celebration of Mass in vernacular languages instead of Latin, the ability to celebrate Mass with the celebrant facing the congregation, changes in nuns’ habits from the traditional garb to more contemporary attire, and more. Scholars are deeply interested in reporting from local and national perspectives before, during, and after the Council meetings.

Figure [ ] shows the timeline for the Vatican II years project, begun with board approval in January 2015 with a targeted completion date of March 2017. To date, CRRA has spent two years developing the plan and cost model, identifying source material, drafting permissions agreements, drafting digitization specifications, and outlining a workflow. CRRA’s digitization vendor, Digital Divide Data (DDD), and Frederick Zarndt in particular, have been instrumental in drafting specifications and setting up our workflow. Frederick has been exceedingly generous, responsive - and patient, in guiding us through this process.



CRRA Vatican II Project Timeline

Key Activities:	2015						2016												2017		
	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M
Project start with Board approval	*																				
Complete digitization agreements		🟥	🟥																		
Obtain copyright permissions				🟥	🟥	🟥	🟥	🟥	🟥												
Identify sources for materials		🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥											
Pre-call to CRRA for funding		🟦	🟦	🟦																	
Public call to non-CRRA libraries					🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦						
Full funding achieved															*						
Scanning and conversion																🟩	🟩	🟩	🟩	🟩	🟩
Platform development																🟩	🟩				
Beta release (funders exclusive)																	🟩	🟩	🟩	🟩	🟩
Summary report																					🟩
Project finished																					*

Open access begins 10/1/2017

Figure [ ]

Project Vendors

CRRA vendors are essential and valuable collaborators. Reveal Digital was an early partner, and has provided crucial project guidance and a viable cost recovery model. Jeff Moyer has been most generous with his time and expertise in shaping the project.

To identify vendors, recommendations were sought from members and colleagues. Several prospective vendors were suggested, yet when it came to METS/ALTO and article segmentation, all roads led to Digital Divide Data (DDD). Scanning vendors of merit were outsourcing the METS/ALTO to DDD, so why not engage directly with them? Working with DDD has enhanced the quality of this project, elevating it from a U.S. project to one of international standing. Finally, DDD complements CRRA's core value of service. DDD is "doing good" with an award-winning and innovative social model that enables talent from underserved populations to access professional opportunities and earn lasting higher income, including youth from low-income families in developing countries, as well as military spouses and veterans in the USA.

CRRA experienced a series of unsatisfactory tests from prospective imaging vendors. DDD put CRAA in touch with Master Enterprises Inc. (MEI) and President, Lisa Stasevich. Lisa has been gracious and generous in providing a variety of samples, and offering guidance on processes and workflow. DDD and MEI work closely together and with DL Consulting. The vendors understand and determine among themselves issues of workflow and design. This has proven to be an unanticipated and positive benefit.

To determine the platform for the Catholic News Archive, CRRA convened The Repository Working Group, chaired by Betsy Post of Boston College. The Working Group identified DL Consulting's Veridian software as the best-in-class platform for the delivery of digital newspapers. It is a turnkey solution that can be hosted locally or by the vendor; for CRRA, a hosted solution was critical.

Veridian generously established a demonstration platform, complete with CRRA branding and a text correction module, and in the test instance CRRA could view the samples that MEI and DDD processed. The image quality, zoning, and OCR was solid, far superior to what were presented from previous vendors.

With just two paid full time staff, this plan would have been inconceivable without member volunteers and a team of vendors who have generously supported and guided the project every step of the way.

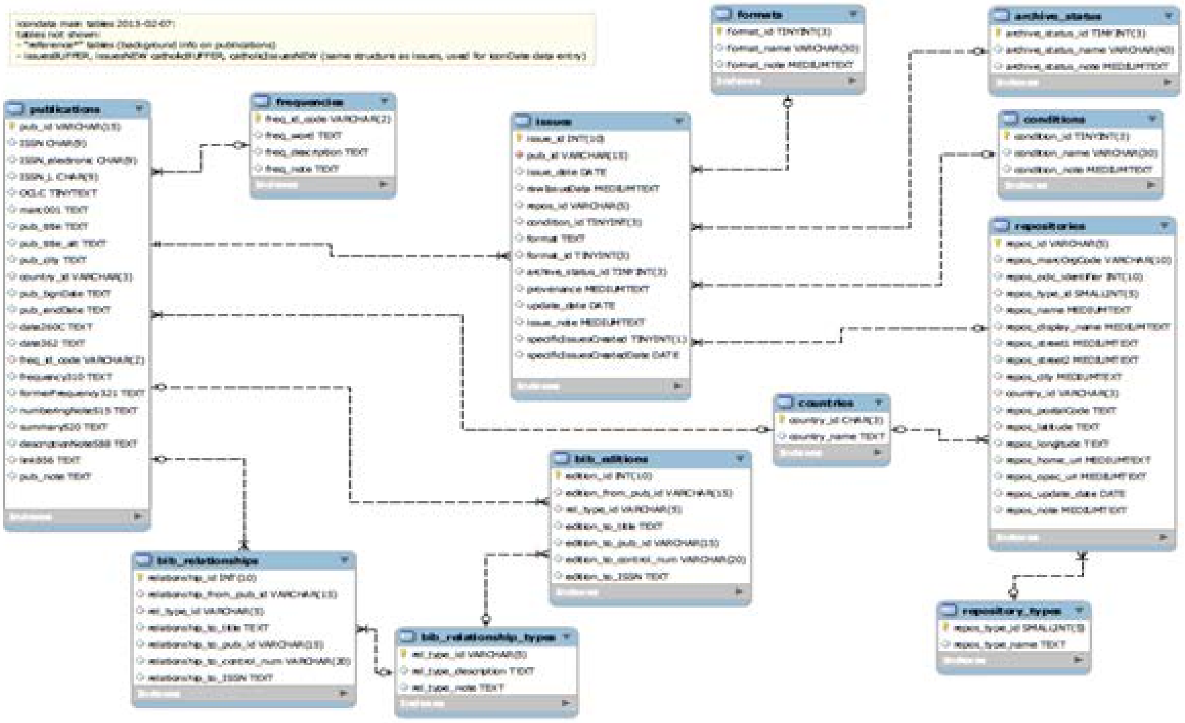
ICON

The International Coalition on Newspapers (ICON) established an international newspaper directory in 2002 called the ICON database. Modeled in part after the U.S. Newspaper Project's effort to comprehensively identify and describe newspapers published throughout America, ICON database was designed as to inform library decisions on the development, management, and preservation of collections of international newspapers. The ICON database was expanded in 2013 to include titles published in the U.S., to incorporate holdings information for digitized newspapers, and to accommodate issue-level holdings for greater granularity. To date, the ICON database contains an estimated 47 million issue-level records for more than 170,000 newspapers from around the world. Because of its level of specificity for issue-level records, it is the largest database of its kind. The ICON database is free and open access for anyone to view newspaper records, and libraries and archives are free to contribute their own newspaper holdings to the database as well. Institutions can use the ICON database to assess their own collections, provide specialist reference services and interlibrary loan to users, and especially plan newspaper digitization projects.

What makes the database so attractive to newspaper digitization efforts is its attention to detail on issues in the ICON database newspaper records. A typical record is divided between the publication information and holdings information. Publication information displays bibliographic information of a title including name, publisher, publication location, identifiers, and more. The holdings information displays dates of specific issues held by organizations in all formats using an interactive calendar and timeline. Users can click through a timeline and calendar to see which libraries and archives hold particular issues in different formats such as paper, microfilm, digital, and more. This is a powerful tool for digitization projects because it shows what source materials are available, keeps record of what newspaper issues exist (and where they are held), and, if available, provides data on what specific issues have already been digitized.

CRRA in collaboration with CRL is using the ICON database for the CNP to improve visibility of Catholic newspapers on a global scale and to assist with pre- and post-digitization. CRRA digitizing partners are invited to submit their issue-level newspaper data to the ICON database. Using the interactive calendar and timeline for pre-digitization, CRRA can see which specific issue dates digitizing partners have and see where gaps need to be filled. For post-digitization, ICON records can be updated to show the specific issue dates in Catholic newspapers that have been digitized by the CNP.

The ICON database is built on a relational database management system (RDMS) that uses SQL to retrieve, insert, remove, and manage newspaper data from CRL and third-party organizations' newspaper collections (called datasources). The model contains several tables that are in relation to one another and work together for user queries. Some examples of these tables include publications, issues, formats, archive statuses, and more. As tables are populated with newspaper metadata by datasources, the information is displayed on ICON's interface for each record. An illustration of ICON's tables is shown below.



(Illustration credit to Andrew Elliott)

To display the publications information on an ICON record, general metadata of a newspaper is populated on the ICON interface using the publications table. Datasources provide basic metadata such as identifiers (OCLC numbers, LCCNs, and so on), title, publication date and location, frequencies, and more. Because these tables are based on, but not restricted to MARC format, any organization can be a datasource including archives, publishers, corporations, and beyond.

One of the most unique features of the metadata model is the issues table. It is responsible for populating the interactive calendar and timeline used on the interface of ICON records. Datasources are given spreadsheet structures that reflect the issues table called issues skeletons. They populate these skeletons by inserting their organizations' holdings, including exact issue date, the formats they carry, and (optional, yet effective) condition status of each issue. The skeletons are then ingested into the database and the calendar and timeline is therefore populated for users. CRL stresses accuracy when datasources provide specific issue dates and formats. Any disruption of the overall flow of the timeline and calendar can cause confusion for users.

The table below shows the specific metadata the ICON database requires from datasources.

Title-level metadata (publications skeleton)

Element	Description	Required?
Title ID	Unique identifier for Title assigned by provider	Required
OCLC	OCLC Record number(s)	At least one is required. Any/all such numbers for the given title are desired.
LCCN	Library of Congress Control Number	
ISSN	International Standard Serial Number	
Publication Title	Title assigned by provider—uniform, family or title for entire publication.	Required
City	City of publication	Required
Country	Country of publication	Required
Publication Date One	Beginning publication year (YYYY)	Required if present
Publication Date Two	Ending publication year (YYYY)	Required if present
Frequency	Publication frequency or pattern of publication. Include frequencies (current, former) and date ranges for when each frequency applies. Corresponds to MARC 310 and 321 fields	Required if present
Other information about dates of publication	Information from MARC 362 or 310 fields, or other sources of dates of publication [beginning date, end date, or variances in publication]. Submit as separate date fields	Required if present

First Date Held	First issue held by provider (YYYY-MM-DD)	Required
Last Date Held	Last issue held by provider (YYYY-MM-DD)	Required
Collection	Name of product or collection	Required

Issue-level metadata (issues skeleton)

Element	Description	Required?
Issue ID	Unique identifier for Issue assigned by provider	Required
Masthead data	Data recorded from keyed masthead [including title as published on given Issue; other fields?]	Required if present
Date	YYYY-MM-DD	Required
Volume Number	Volume number	Required if present
Issue Number	Issue number	Required if present
Edition	Edition designation	Required if present
Number of pages	Number of pages included in issue.	Required
Source*	Source of the film/print from which the issue was digitized. May include “Provenance” (source from which Issue was procured), or “Reproducer” (production details for the filmer of given Issue)	Required if present
Additional Source Information	Cataloging or identifying information known about source	Required if present
Completeness	Any information about incomplete issues due to missing or damaged pages.	Required if present

For general inquiries about the ICON database, please contact James Simon, Vice-President, Collections and Services for CRL (jsimon@crl.edu) and Amy Wood, Director of Technical Services for CRL (awood@crl.edu).

2.x Technical specifications background

CRRA is digitizing newspapers based on a foundation laid during the past 10+ years of newspaper digitization projects in North America, Europe, and Asia-Pacific. In particular, CRRA specifications draw on newspaper metadata and technical precedents set by the Library of Congress / National Endowment for the Humanities National Digital Newspaper Program (NDNP) in the USA, the Australian Newspapers Digitisation Program (ANDP), Europeana Newspapers, the Singapore National Library Board's NewspaperSG, the California Digital Newspaper Collection (CDNC), Utah Digital Newspapers, and the Boston College digital newspapers.

These collections are a mixture of page-level and article-level digital newspapers. Articles, even those that continue across pages, are easy to "clip" from article-level digital newspapers; clipping is important for digital newspaper users, especially for family historians. For digital newspaper collections that enable crowd-sourced OCR text correction, it is easier, less frustrating, and more satisfying to correct the text of a single article rather than correcting the text of an entire page.

But there is a compelling argument to be made for page-level newspapers and that is cost. Generally, page-level digitization costs are about half of article-level digitization costs. This means that for the same amount of money, one can digitize twice as many newspapers. Which would users of a digital newspaper collection rather have: Twice as many pages to search and browse or a better user experience?

Commercial digital newspaper collections are a mixture of article-level and page-level. Proquest's Historical Newspapers are article-level. The Newspapers.com collection is page-level. Both have plenty of paid subscribers. Cultural heritage newspaper collections too are a mixture of article-level and page-level, sometimes within the same collection. California and Utah, both NDNP program participants, chose to digitize newspapers hosted with their own software systems to article-level while delivering page-level digitized newspapers to Chronicling America.

Nearly every cultural heritage organization which digitizes newspapers uses METS XML ALTO XML, TIFF archival images, and usually JPEG2000 access images. Sometimes PDF image over text files are created for issues and/or pages, but sometimes they are not. The technical specifications and workflow for page-level and article-level digitization is very similar: TIFF and JPEG2000 images can be identical, ALTO XML files can be identical, and much of the metadata contained in the METS XML file is the same. The difference between them is in the article metadata. That metadata may be contained in the METS XML file or in a separate XML file and typically consists of article title, body, and type. Other fields are possible but not so common, for example, sub-title and byline (author).

In spite of its cost premium over page-level digitization, CRRA chose article-level digitization because of its superior user experience. CRRA will publish its specifications for use by its members and anyone else. Although the specifications are now complete, publication of the specifications will be delayed until some of the collection is online in order to permit the inevitable last minute specification changes.

2.x+1 Workflow

Broadly viewed, the digitization workflow has 6 distinct steps. These are not waterfall-type steps, but, after initial use rights negotiation and materials collection, all steps will be done in parallel. Newspapers will be digitized in batches of issues, for example, a batch will be comprised of N or more issues of one or more titles[1].

2.x+1.1. Use rights negotiated and materials collected

User rights and sourcing material are provided via a wide network of "digitizing partners," identified in the legal agreements as "Sourcing Partners." Sourcing Partners are the key to this collaborative endeavor. They are member and non-member institutions alike comprised of universities, colleges, publishers, archdioceses, archdiocesan archives, and seminaries that hold or have access to the pristine copies of sourcing material for the priority papers and/or access to the copyright holder, usually the Bishop or publisher. In collaboration with CRRA, Sourcing Partners have identified the best sourcing material[2] and have identified and negotiated licensing agreements with copyright holders for the priority papers. At this writing, all copyright holders have been identified, two signed agreements are in hand, with others in the pipeline.

2.x+1.2. Source materials dispatched from source repository to image service bureau

In collaboration with members of CRRA, the image service bureau Master Enterprises Inc. (MEI) and Digital Divide Data project managers, CRRA staff schedules the materials for scanning and production. Both microfilm and newspaper hard copy will be scanned at a single location. The materials will be returned to the source repository soon after Digital Divide Data verifies that the source data is complete (no missing pages, bad images, etc).

2.x+1.3. Images sent to production

After the service bureau scans the microfilm or newspapers and does initial quality control on the images, a batch of images is copied to a USB hard drive and sent to one of Digital Divide Data's production facilities in SE Asia (about 5 days transit time). At the production facility, the images from one issue are grouped together and ingested into Content Conversion Specialists docWorks software. Digitization production goes as follows:

Activities	Role	Description
Receive the source images	PM	• Receipt/Acknowledge the input files
Create an	PM	• Generate the image list into excel for inventory

[1] In general production prefers a single title or, at any rate, a small number of titles, in a batch. This reduces the possibility of errors during ingest or by operators during production.

[2] Second-generation (2N) duplicate silver negatives are preferred (for reasons of cost) and where not available, then print.

inventory and status report		report
Check metadata	Operators	• Review the METS template of each issue making sure it matches the correct metadata in print. • Pre-identify the blank page, missing page, target page, duplicate page, and duplicate issues. • Checks for problems and issues that are not covered by the layout, zoning and markup instructions. Report possible action that DDD will take to correct those issues. • Reviewing the input structure before import into DocWorks
Import	PM/PA	• Loads the image and batch files to the system for processing.
Crop and deskew	Operator	• Corrects skew angle so that it is less than 3̐. Crop borders to page edge.
Review zoning	Operator	• Review crop and deskew results before review zoning job, give a comment if error found • To classify the page content element if they are headlines or body text. • Identify the reading direction of each zone.
Review page sequence	Operator	• Check for missing/duplicated pages based on print folios. • If required, set page types (title page, table of content page, cover page, etc.) • To verify and insert issue level metadata
Collect output	Project Admin	• Generate the output list and compare with the input manifest. • Move from the processing server to project storage
Final QC on article headlines	QC	• Extract the headline and author text from the METS file then looking for suspicious words
Quality assurance	QA	• Collect sample 10% random of the total issues from the batch and check against the quality requirements
Cross check	PM	• Make sure all the output files exist in every folder and the amount of page are correct. • Make sure that there are NO corrupt files.
Deliver output	PM	• Send output data to repository via Fedex or UPS

Note that the output data is delivered in bagit data bags. Bagit data bags create a file manifest and generate checksums for each output file, thus ensuring that any file corruption can easily be detected before ingest into the repository.

2.x+1.4. Output data sent to repository host

CRRA chose to use Veridian software as access software for the digitized newspapers, in part because of Boston College's success with Veridian (cf. http://newspapers.bc.edu/). Veridian software is from DL Consulting in Hamilton, New Zealand; it is a child of the New Zealand Digital Library Project at the University of Waikato and its open source Greenstone software (http://www.greenstone.org). Greenstone is also supported by UNESCO and the Human Info NGO. Greenstone was first released in November 2000, and, as such, is one of most mature digital library software packages available.

Veridian's origins are from Greenstone and it shares many features with Greenstone including support for ~60 languages, but it is no longer open source. Veridian software is specialized to support large collections (millions of objects) of page-level and article-newspapers. Veridian software runs on most common operating systems -- Linux (preferred), Windows, Mac OS X, Solaris -- and on Amazon's AWS cloud storage and Elastic Compute Cloud.

Veridian runs in the cloud and DL Consulting provides all support and services for Veridian collections hosted in the cloud. This is especially attractive since CRRA has no data center and no IT staff to support the CRRA digital newspapers collection.

2.x+1.5. Output data ingested into cloud-based repository

Although DL Consulting is headquartered in Hamilton, New Zealand, the output data is shipped on hard drives to a data center in the USA where it is uploaded to the Amazon AWS cloud and ingested into Veridian. Via remote terminal sessions, DL Consulting engineering and support staff ensure that the output data is complete and hasn't been corrupted using bagit file manifests and fixity checks, ingest the output data into cloud-hosted Veridian, configure the collection according to CRRA specifications. Following successful ingest, the hard drives are forwarded by DL Consulting to MetaArchive for deposit into a dark archive for long-term preservation.

2.x+1.6. Output data deposited into dark archive

Preserving Catholic newspapers in perpetuity is a key project goal and to that end, CRRA intends to deposit a copy of all project files (TIFF archival images, JPEG2000 access images, METS XML, ALTO XML,PDFs), into the dark archive, MetaArchive. Founded in 2004, MetaArchive is a community-owned, community-led initiative comprised of libraries, archives, and other digital memory organizations to create and maintain a secure and cost-effective repository that provides for the long-term care of digital materials – not by outsourcing to other organizations, but by actively participating in the preservation of member content.

3. **Collaboration**

CRRA accomplishes its goals through a distributed and collaborative organization comprised of member and non-member institutions alike. Committees and working groups provide essential services and guidance in all that CRRA does, such as policy-writing, strategic planning, and conducting user studies. For the Newspapers Program, collaboration has been taken to new heights. As described earlier, the Vatican II project is attributable to the collaborative efforts of CRRA committees, through provision of source material and securing copyright permission by Sourcing partners, and by a team of vendors.

Project coordination and oversight has been the role of CRRA's two paid employees, CRRA's executive director, Jennifer Younger and digital projects librarian, Pat Lawton. They have worked behind the scenes to coordinate the activities of these various groups, and to develop project plans, funding models, digitizing partners and page counts, cost estimates, digitization specifications, and legal agreements. In 2015, CRRA contracted with the CRL to hire Nick Casas of CRL for fifteen hours per week to coordinate efforts to add records for Catholic newspapers to the ICON database. With the digitization process on the horizon, Nick has recently been redeployed to work with sourcing partners to identify the best source material available and will provide oversight of the digitization process as content is shipped by the sourcing partner and is eventually returned, along with digitized content.

4. **Library crowdfunding**

An increasing number of libraries are earmarking budget dollars to support open access initiatives, such as: arXiv, SCOAP3, Knowledge Unlatched, and the Text Creation Partnership (TCP). While their content type (journals, scholarly monographs, special collections, etc.) and time periods covered (from current back to the birth of the printed book) varies, all of these initiatives share the common goal of creating open access collections. These and similar initiatives are being viewed by libraries as strategic investments in building digital corpora of open access material.

Reveal Digital was launched in 2012 as a library crowdfunded service to assist libraries in pooling their funding and holdings to create open access digital collections. The goal of Reveal Digital's initial project, *Independent Voices*, is to digitize 1,000 alternative press publications from the later-half of the twentieth century, As of April 15th, 2016, Reveal Digital has raised over $1.1 million dollars from 79 contributing libraries in the U.S., Canada and the U.K. Material for the project has been sourced from dozens of libraries, archives and private collections.

For library crowdfunding to work, there needs to be a well defined cost structure to serve as the basis for defining the funding goal. The costs for the Vatican II project are displayed below and are estimated to be $419K in total. They cover everything from sourcing material, project management, content conversion (scanning and metadata), platform development, outreach and marketing, digital archiving, and general administrative expenses. Behind these summaries are very detailed cost models.

In addition to the first year's development costs, CRRA has included 3 additional years of hosting and digital archiving costs for the Vatican II project to provide some time to develop and implement a longer-term funding plan for continued access.

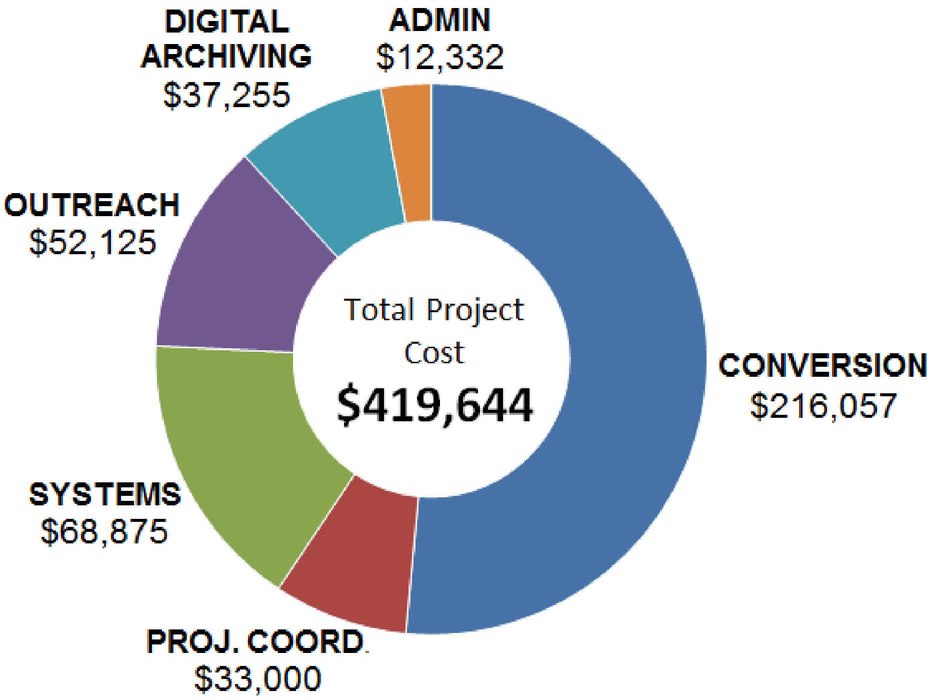


Figure [ ] - Estimated costs by major category for the Vatican II Project.

The next step is to set the investment levels for library funding. CRRA has adopted a tiered approach to setting these levels for the Vatican II project. The tiers attempt to provide equitable funding levels based first by type of library, and then for academic libraries, by ARLs, and for non-ARL libraries by highest degree granted in history or religious studies.

CRRA looks to match the number of libraries (by tier) with the suggested investment levels to try to find that balancing point between price and units that results in the lowest investment levels that will generate enough funding to reach the cost recovery goal (see Figure [ ]). If Reveal Digital and CRRA are successful in their outreach activities, then the total funding raised will equal the total project costs. Once funding goal is achieved, the sourcing, conversion and development stages of the project will start.

Institution Type	Target Count of Funding Libraries	One-time Commitment Level	Funding Raised (Target)
Diocesan/Religious	10	$200	$2,000

Theological/Special	23	$500	$11.500
Bachelors	26	$3,400	$88,400
Masters	27	$4,250	$114,750
Other Doctorate	13	$5,100	$66,300
ARL	12	$8,500	$102,000
Large Publics	8	$4,250	$34,000
TOTAL	**119**		**$418,950**

Figure [ ] - Tiered funding levels for the Vatican II Project with estimated counts of the number of funding libraries.

4.1 Outreach plan

The outreach plan starts with CRRA members and partners. This is the group with the shared vision and mission for the project. Currently, we are in the process of asking CRRA member libraries to make a commitment at the proposed investment levels – if they all did, then about one-third of the project's target goal would be raised. To date CRRA members have contributed about 33% to that goal. Daily progress to goal, as well as a listing of the contributing libraries, is tracked on the CRRA website. Early support from CRRA members is critical to attracting funds outside of CRRA. Beyond CRRA, Reveal Digital will be extensively marketing to libraries individually and through consortia. The next funding call will be going out to non-CRRA Catholic institutions and other religious libraries. That call will be followed by a general funding call to all other non-Catholic libraries on the target list.

It's important to provide funding institutions with a distinct set of benefits, something for them to point to, to help strengthen their decision about investing in the project. For Vatican II this includes: early access to the collection while it's under development (plus a 6 month bonus access period), free MARC records, and support for mass text downloading. There's also the collective benefit of making Catholic newspapers an accessible archive for all, which may be the most important benefit of all.

5. **Sustainability**

The first step in defining the long-term strategy is to actually build the repository. This involves setting up a platform, adding content, and putting a digital archiving solution in place (CRRA has selected Meta Archive for the Vatican II project). The funds CRRA hopes to raise for the Vatican II project will cover all of these costs for that initial three-year period.

Beyond this initial period, CRRA will be exploring longer-term solutions and options. For example, for the long-term hosting for Vatican II – this could stay on the Veridian platform or could involve another hosted solution more cost effective for long-term access. Ideally, this would be a trusted provider, perhaps even one of CRRA's academic partners.

There are also the funding issues associated with digitizing and adding new content, both additional years of coverage for the initial set of newspapers but also adding new titles, which may include newspapers that have already been digitized. There is no single solution to funding, rather CRRA sees a mix a funding sources & strategies, which will likely include continued library crowdfunding, but will also look to grants, sponsorships, content licensing and other options.

For this project to be a success, CRRA needs to build that community of libraries and archives who see the value in digitizing and archiving Catholic community newspapers and who are willing to financially support the project going forward. The larger CRRA can make that community the more equitably the costs can be spread. As CRRA moves beyond the Vatican II project, they'll be encapsulating these options and strategies in a formal long-term plan.

References

Australian Newspaper Digitisation Program. http://trove.nla.gov.au/newspaper
and https://www.nla.gov.au/content/newspaper-digitisation-program

Boston College digital newspaper. http://newspapers.bc.edu/

California Digital Newspaper Collection. http://cdnc.ucr.edu/

Catholic Newspapers Online. http://www.catholicresearch.net/cms/index.php/catholic-newspapers/catholic-newspapers-online/

Center for Research Libraries, ICON. https://www.crl.edu/programs/icon

Chronicling America. http://chroniclingamerica.loc.gov/

CRRA Catholic Portal. http://www.catholicresearch.net/cms/index.php/catholic-portal/olic Portal

CRRA Digitizing Partners. http://www.catholicresearch.net/cms/index.php/about/crra-groups/catholic-newspapers-committee/crra-digitizing-partners/

Digital Divide Data (DDD). http://www.digitaldividedata.com/

DL Consulting Veridian. http://www.veridiansoftware.com/about/

Europeana Newspapers. http://www.europeana-newspapers.eu/

International Coalition on Newspapers (ICON) Database. http://icon.crl.edu

Master Enterprises, Inc. (MEI). http://www.yourdigitalsource.net/

MetaArchive. https://www.metaarchive.org/

National Digital Newspaper Program. https://www.loc.gov/ndnp/, http://chroniclingamerica.loc.gov/, and https://www.loc.gov/ndnp/guidelines/

OCLC Worldcat. https://www.worldcat.org/

Priority Papers (detailed list). http://www.catholicresearch.net/cms/index.php/catholic-newspapers/priority-papers/

Reveal Digital. http://www.revealdigital.com/

Singapore NewspaperSG. http://eresources.nlb.gov.sg/newspapers

Utah Digital Newspapers. http://digitalnewspapers.org/