Collections as Data Facets

This document follows the Collections as Data Facets structure outlined by the Always Already Computational - Collections as Data project (https://collectionsasdata.github.io/facet7/). It describes the people, services, practices, technologies, and infrastructure used to create the Periodicals in the US-Mexico Border Region research portal and aims to assist others planning similar efforts.

1. Why do it

Periodicals in the US-Mexico Border Region is a bilingual research portal that provides access to digitized periodicals from Arte Público Press, Recovering the US Hispanic Literary Heritage (Recovery) Digital Archive in collaboration with the US Latino Digital Humanities Center (USLDH) located at the University of Houston. The portal also includes digital copies of newspapers housed at the Instituto Municipal de Arte y Cultura (IMAC) in Tijuana, Baja California, Mexico and the Archivo Municipal del Saltillo in Saltillo, Coahuila, Mexico. The digitized materials are invaluable to scholars, educators, and the general public. This project has the potential for a broad disciplinary and public memory impact as it offers first-hand access to reporting on major events that shaped the US border and allows greater recognition of the important diversity within Latino/a experiences.

2. Making the Case

The Recovery program is one of the premier research centers focusing on the understanding of the experience of Latino/a people and their written legacy. Recovery has three decades of experience in locating, acquiring and preserving primary source materials and making them accessible online. The Periodicals in the US-Mexico Border Region portal is significant as it provides access to primary source materials documenting the history of the border and its cultures. Periodicals cover the crucial period of US annexation of Mexico's northern lands, the period of the Mexican Revolution and the migration of some one million Mexican nationals to the United States, among other topics. These events also include coverage of decades of social and labor movements. Even more importantly, Recovery has the infrastructure and archival expertise to support and maintain this digitization project. They launched the first Center for

Latino Digital Humanities and have a complementary relationship with the University of Houston Libraries and Technology support.

3. How you did it

People

Project Leadership
Nicolás Kanellos, Ph.D., Recovery Program Director
Carolina Villarroel, Ph.D., Recovery Director of Research
Mikaela Selley, CA, Recovery Program Manager

Web Development
Anneliese Dehner, Independent Digital Library Developer

Preservation & Reformatting
Jerrell Jones, Former Digitization Lab Manager, University of Houston Libraries. Now Digital Initiatives Coordinator, Rice University Fondren Library
Bethany Scott, Head of Preservation & Reformating, University of Houston Libraries

Research Assistants
Ayann'ah Batiste, Undergraduate Student Intern
Sloane Davis, High School Student
Carlos Alberto Flores, University of Tijuana
Javier R. Franco, Graduate Student, Recovery Research Assistant
Adrián Alexis García, University of Tijuana
Kai Gomez, High School Student Intern
Valeria Gonzalez, Undergraduate Student Intern
Yanina Hernández, Graduate Student, Recovery Research Assistant
Kathleen Ortiz, Undergraduate Student Intern
Perla Ortiz, Graduate Student, Recovery Research Assistant
Camilo Rodriguez, Graduate Student, Recovery Research Assistant
Jacqueline Torres, Undergraduate Student Intern
Celeste Uribe, Undergraduate Student Intern

Alaíde Ventura Medina, Graduate Student, Recovery Research Assistant

Strategic Planning, Editing, and Translations Assistance
Dr. Gabriela Baeza-Ventura, Arte Público Press Deputy Director
Dr. Lorena Gauthereau, Digital Programs Manager

Collections

This project includes materials housed in the following Recovery Archival Collections.
Alonso S. Perales Papers
Amira Mejía Collection
Recovery Reference Collection
Religious Thought Collection
Saltillo Periodicals Collection
Tijuana Periodicals Collection

For more information on Recovery Archives, visit https://artepublicopress.com/archives/

Selection Criteria

Our selection criteria for this project is materials published in a state along the US-Mexico border with a focus on our earliest holdings.

Of our total holdings (approximately 1,500 periodicals) over 50% come from the US-Mexico Border States. The large majority are on the US side. Over half pre-date 1950 and most of those are within 200 miles. This left us with close to 400 periodicals to choose from. We then narrowed this down further by removing from our list those issues in other institutions, and periodicals partially digitized as part of other projects such as exhibits and research databases. We also looked at content and prioritized the publications that include local news and works by Hispanic writers and publishers, as opposed to items such as translated world news publications.

Equipment & Technologies

ABBY FineReader (for OCR process)

Canon MS-800 (Microfilm Scanner that malfunctioned halfway into the project)

GIMP Software (Image editing: rotation, crop, color correction tools)

IrfanView Software (Scanning Software, also used for bulk conversion to jpg and pdf as needed)

Konica Minolta ScanDIVA (Book & Large Format Scanner)

Omeka S (web-publishing platform)

ScanPro 3500 (Microfilm Scanner)

Scanning Utility 800 (Scanning Software for Canon MS-800)

Challenges

We faced several challenges related to equipment and technology. We did not anticipate a failing microfilm scanner, or the months-long process to make the purchase for a new machine. The other significant challenge was in generating quality transcriptions through the OCR process. Due to the bilingual nature of our periodicals and condition of some of our holdings (fragile and damaged periodicals), it was challenging to achieve a high accuracy rate for transcription. We used ABBYY FineReader and worked with an IT Librarian who encouraged several batch tests to determine if any steps in our scanning process were impacting the quality of the OCR. We encourage you to test several batches of images to ensure you achieve the quality desired for your project. For example, our microfilm scanner offers a variety of settings that change the quality of the scan in subtle ways that do not appear to impact the visual quality, but did impact the OCR readability. It was important for us to test various settings to understand our output options before scanning hundreds of images. Several periodicals did not produce a high accuracy transcription, however, we ultimately decided that any transcription is better than none. Our site provides researchers with access to everything produced through our OCR process. These challenges led to several valuable lessons from this project. The most significant being a better understanding of our IT needs for future, large-scale digitization projects. This project has strengthened our understanding of the kinds of questions we need to ask of our IT partners in the future, and highlights the crucial role of our web developer in these discussions.

4. Share the docs

Methodology

USLDH Best Practices https://artepublicopress.com/digital-humanities/

Scanning Standards

Recovery scanned from original newsprint to create uncompressed .tiff files at 300dpi (24-bit depth, color) and from microfilm to create uncompressed .tiff files at 300dpi (8-bit depth, grayscale). In addition, derivative files were created to produce JPEG, PDF, and TXT files from each TIFF file.

Metadata Standards

Recovery created metadata using controlled vocabulary by the Library of Congress and a local vocabulary where appropriate LOC subject headings were not found. All metadata created by Recovery is bilingual in English and Spanish.

5. Understanding use

This research portal contains approximately 25,000 scanned pages from over 190 periodicals and is fully bilingual. A drop-down menu on the Home Page offers the option to view the site in English or Spanish. Users can browse periodical titles and have the option to sort alphabetically, by publication location, date, and language. A map on the home page features pins marking the location of all periodicals. Clustered circles with numbers represent large concentrations of publications in one area. Click the circles to zoom in and see the individual pins. Click on pins to view individual periodical titles.

A separate page offers searching by periodical issue where users can enter search terms and/or filter by type of periodical (magazine or newspaper), date, publication location, language, subject headings, and keywords. When a periodical or issue is selected, users can access descriptive metadata that includes an abstract summarizing the history and contents of the periodical. Once a specific page is selected, the portal offers several options for access and use including 3 citation formats, downloadable image file (.jpg) and transcription (.txt), printing, emailing, and copying the direct link.

This project was designed with the experienced researcher, educator, and general public in mind. We chose to use Omeka S and hired a web developer with experience designing visually pleasing and user-friendly sites. The website is inviting, clear, easy to use and free of jargon. As a research portal it is useful for exploring the history of Hispanic publishing on both sides of the United States and Mexico border, as well as researching significant historical moments and people. An exciting feature for educators is a page offering lesson plans for grades 6-12 based on a sampling of periodicals on the site. The portal may be of interest to research and classroom topics in linguistics, history, print history, religion, Latina/o studies, border studies, labor studies, social movements, gender studies, and more.

6. Who supports use

This project is supported by the US Latino Digital Humanities Center/Recovering the US Hispanic Literary Heritage Program. Periodicals in the US-Mexico Border Region is funded by the Council on Library Information Resources (CLIR) Digitizing Hidden Special Collections and Archives awards program, which is generously funded by the Mellon Foundation.

7. Things people should know

Metadata

This project pulls from two sets of descriptive metadata using excel spreadsheets. One set of metadata is at the “Newspaper-Level” and the other is “Issue-Level.” Newspaper level metadata includes a row of data for each periodical, while the issue-level spreadsheet includes a row of data for every issue in a periodical. This method is not standard practice at Recovery and was implemented specifically for this project based on conversations with the web developer. Two sets of data allowed us to develop the site in batches so as to test and address issues with each search page (Browse Periodicals vs. Search Issues). This also allows us to ingest the newspaper-level data into our larger collections database easily, while keeping the issue-level details separate. The challenge to more than one spreadsheet, of course, is ensuring that both are updated as needed.

Optical Character Recognition (OCR)

Our site offers a downloadable transcript file in .txt format for every page of a periodical. This .txt file was produced after running each scanned image through OCR software. Due to the bilingual nature of our periodicals and condition of some of our holdings (fragile and damaged periodicals), several periodicals did not produce a high accuracy transcription. However, we decided that any transcription is better than none. Our site provides researchers with access to all files produced through our OCR process.

Web Developer & IT Support

The most significant challenges related to equipment and technology, and thus we emphasize the need for a strong relationship with your web developer and IT support. This project has strengthened our understanding of the kinds of questions we need to ask of our IT partners in the future, and highlights the crucial role of our web developer in these discussions. View this pdf for a list of questions for consideration.

8. What’s next

The web developer hired for this project is providing Recovery with instructions for the website maintenance and upload process, so that Recovery can continue to add content each year. We also aim to create a Create Creative Commons account for the project and upload project information into the University of Houston’s digital repository, Research Open Access Repositories (ROAR) to ensure discoverability. Recovery also plans to continue efforts to publicize the project in conferences and presentations around the country and abroad.