Colecciones como facetas de datos

Este documento explica qué personas, servicios, prácticas, tecnologías e infraestructura fueron utilizadas para la creación del portal de búsqueda de periódicos de la región fronteriza entre los Estados Unidos y México, con el objetivo de ayudar a la planeación de esfuerzos similares. Hemos seguido la estructura de “Colecciones como facetas de datos” planteada por el proyecto Always Anyway Computational para las colecciones como datos (https://collectionsasdata.github.io/facet7/).

1. ¿Por qué hacerlo?

Periódicos de la región fronteriza entre los Estados Unidos y México es un portal bilingüe de investigación que permite el acceso a periódicos digitalizados de las colecciones de Arte Público Press, el programa de recuperación de la herencia literaria hispana (Recovery) y el centro de humanidades digitales de los latinos estadounidenses (USLDH) de la universidad de Houston. El portal también incluye copias de periódicos hospedados en el Instituto Municipal de Arte y Cultura (IMAC) de Tijuana, Baja California, México, y en el archivo municipal de Saltillo, en Coahuila, México. Estos materiales digitales poseen un valor inigualable para académicos, profesores y público general. Pueden ser útiles para varias disciplinas e impactar en la memoria histórica, ya que ofrecen información de primera mano sobre sucesos relevantes para la región fronteriza México-Estados Unidos, al tiempo que promueven el reconocimiento de la diversidad de experiencias de los y las latinas.

2. Justificación

El programa Recovery es un centro pionero de investigación enfocado a la comprensión de las experiencias de los y las latinas en Estados Unidos y su legado escrito. A lo largo de tres décadas, Recovery ha localizado, adquirido y preservado materiales de primera mano y los ha puesto al acceso del público. La enorme relevancia del portal de periódicos de la región fronteriza entre los Estados Unidos y México radica en que permite el acceso a materiales que documentan la historia de la frontera y sus culturas. Los periódicos abarcan los períodos cruciales de la cesión mexicana de las tierras del norte, la revolución mexicana y la migración de más de un millón de mexicanos a Estados Unidos, entre otros. También, la cobertura de

movimientos sociales y laborales durante varias décadas. Lo más importante es que Recovery cuenta con la infraestructura y el conocimiento archivístico necesarios para habilitar y mantener este proyecto. Recovery fue el primero en lanzar un centro de humanidades digitales y actualmente mantiene una colaboración cercana con las bibliotecas y la red de apoyo tecnológico de la universidad de Houston.

3. Cómo lo hicimos

Personas

Líderes de proyecto
Dr. Nicolás Kanellos, director del programa Recovery
Dra. Carolina Villarroel, directora de investigaciones del programa Recovery
Ac. Mikaela Selley, coordinadora del programa Recovery

Desarrollo web
Anneliese Dehner, desarrolladora independiente de bibliotecas digitales

Preservación y cambio de formato
Jerrell Jones, antiguo coordinador del laboratorio de digitalización de la biblioteca de la universidad de Houston y actual coordinador de iniciativas digitales de la biblioteca Fondren de la universidad de Rice
Bethany Scott, líder de preservación y cambio de formato de la biblioteca de la universidad de Houston

Asistentes de investigación
Ayann'ah Batiste, estudiante subgraduada en pasantía
Sloane Davis, estudiante de bachillerato
Carlos Alberto Flores, universidad de Tijuana
Javier R. Franco, estudiante de posgrado, asistente de investigación de Recovery
Adrián Alexis García, universidad de Tijuana
Kai Gomez, estudiante de bachillerato en pasantía
Valeria Gonzalez, estudiante subgraduada en pasantía
Yanina Hernández, estudiante de posgrado, asistente de investigación de Recovery

Kathleen Ortiz, estudiante subgraduada en pasantía
Perla Ortiz, estudiante de posgrado, asistente de investigación de Recovery
Camilo Rodriguez, estudiante de posgrado, asistente de investigación de Recovery
Jacqueline Torres, estudiante subgraduada en pasantía
Celeste Uribe, estudiante subgraduada en pasantía
Alaíde Ventura Medina, estudiante de posgrado, asistente de investigación de Recovery

Planeación estratégica, edición y apoyo en traducción
Dra. Gabriela Baeza-Ventura, directora adjunta de Arte Público Press
Dra. Lorena Gauthereau, coordinadora de programas digitales

Colecciones

Los materiales de archivo incluidos en este proyecto provienen de las siguientes colecciones de Recovery:
Papeles de Alonso S. Perales
Colección Amira Mejía Collection
Colección de referencia de Recovery
Colección de pensamiento religioso
Colección de periódicos de Saltillo
Colección de periódicos de Tijuana

Para más información sobre los archivos de Recovery, visita
https://artepublicopress.com/archives/

Criterios de selección

Para este proyecto, nos enfocamos en los materiales publicados en los estados fronterizos de México y Estados Unidos, con un énfasis en su antigüedad.

Más de la mitad de nuestros materiales, que comprenden alrededor de 1500 periódicos, proviene de los estados fronterizos de México y Estados Unidos, y la gran mayoría proviene del

lado estadounidense. También, más de la mitad de los materiales es anterior a 1950 y proviene de un radio de doscientas millas. Esto nos deja con 400 materiales disponibles. Para delimitar aún más la muestra, dejamos fuera los ejemplares que otras instituciones y proyectos ya han comenzado a digitalizar para sus proyectos, exhibiciones y bases de datos. Finalmente, revisamos el contenido y priorizamos las publicaciones que incluyeran noticias locales y trabajos escritos y editados por hispanos e hispanas por sobre las traducciones y las noticias globales

Equipo y tecnología

ABBY FineReader (para el procesamiento de OCR)

Canon MS-800 (escáner de microfilm que dejó de funcionar a mitad del proyecto)

Software GIMP Software (edición de imagen: rotación, recorte y corrección de color)

Software IrfanView Software (software para escaneo, también empleado para convertir bloques enteros de jpg a pdf cuando fue necesario)

Konica Minolta ScanDIVA (escáner de libros y gran formato)

Omeka S (plataforma de publicación web)

ScanPro 3500 (escáner de microfilm)

Scanning Utility 800 (software de escaneo para Canon MS-800)

Retos

Nos enfrentamos a diversos retos relacionados con el equipo y la tecnología. No habíamos previsto que el escáner dejaría de funcionar, ni que la compra de uno nuevo nos tomaría largos meses. El otro reto significativo fue generar transcripciones de calidad con OCR. La naturaleza

bilingüe de nuestro proyecto, y las condiciones de algunos materiales (periódicos frágiles y dañados), nos dificultaron la obtención de transcripciones fidedignas. Utilizamos ABBYY FineReader y trabajamos con un bibliotecario de tecnologías de la información que nos animó a realizar varias pruebas para determinar si alguna de nuestras acciones al escanear estaba mermando la calidad del OCR. Ahora, nosotros te animamos a ti a realizar estas pruebas de imagen hasta asegurarte de alcanzar la calidad deseada. Por ejemplo, nuestro escáner de microfilm permite cambios en la configuración que afectan sutilmente la calidad de la imagen; si bien esto no se refleja en la calidad visual, sí en la legibilidad del OCR. Tuvimos que probar varias configuraciones para entender cómo se iba modificando el resultado, antes de embarcarnos en el escaneo de miles de imágenes. Para algunos periódicos no fue posible obtener una transcripción exacta, pero decidimos que contar con una perfectible era mejor que no contar con ninguna. Nuestro sitio permite acceso a todos los contenidos con OCR.

Al enfrentarnos a estos retos, también aprendimos mucho. Lo más importante fue entender mejor nuestras necesidades de tecnologías de información para proyectos futuros de digitalización a gran escala. Ahora sabemos qué preguntas debemos hacerle a nuestros socios de tecnologías de la información, y entendemos el rol tan crucial que desempeñan los desarrolladores web en estas conversaciones.

4. Compartir los documentos

Metodología

Buenas prácticas de USLDH https://artepublicopress.com/digital-humanities/

Estándares de escaneo

Recovery escanea periódicos impresos originales para generar documentos .tiff sin comprimir a 300dpi (24-bit profundidad de color) y microfilm para generar documentos .tiff sin comprimir a 300dpi (8-bit profundidad en escala de grises). Además, genera documentos derivados para producir JPEG, PDF, y documentos TXT desde cada documento TIFF.

Estándares de metadatos

Recovery genera metadatos usando el vocabulario aceptado por la biblioteca del congreso. Si la biblioteca no cuenta con los encabezados, se emplea vocabulario específico local. Todos los metadatos creados por Recovery son bilingües español e inglés.

5. Navegación del portal

Este portal de investigación contiene cerca de 25,000 páginas escaneadas provenientes de 190 periódicos y es totalmente bilingüe. Un menú desplegable en la página principal ofrece la posibilidad de visitar el sitio en inglés o español. Los usuarios pueden buscar por título de periódico o acomodar los contenidos alfabéticamente, por lugar de publicación, por fecha y por idioma. En el mapa de la página principal pueden identificar la ubicación original de cada periódico y, agrupadas en círculos, aparecen las concentraciones más grandes de publicaciones por área. Pueden dar clic en los círculos y ver los marcadores individuales; cada uno mostrará los periódicos de forma individual.

Otra página permite la búsqueda por periódico. Aquí, los usuarios pueden ingresar datos y/o filtrarlos por tipo de publicación (periódico o revista), fecha, ubicación, idioma, encabezado y palabras clave. Al seleccionar un ejemplar, accederán a metadatos descriptivos que incluyen un resumen de la historia y contenidos del periódico. Una vez que seleccionen la página, el portal ofrecerá varias opciones de acceso y utilización, incluyendo tres formatos de citación, imágenes descargables (.jpg) y transcripciones (.txt), así como la posibilidad de imprimir, enviar por correo electrónico o copiar el vínculo directo.

Este proyecto fue diseñado pensando en investigadores, académicos y público en general. Decidimos usar Omeka S y contratar una desarrolladora web con experiencia en la creación de sitios amigables para el usuario. El portal es llamativo, claro, fácil de usar y está libre de tecnicismos. Es de gran utilidad para explorar la historia de las publicaciones hispanas a ambos lados de la frontera México - Estados Unidos, así como para investigar momentos relevantes y personajes icónicos. Una característica importante del sitio es la sección que ofrece material educativo para niños de sexto a doceavo grado, que consiste en un muestreo de los periódicos disponibles. El portal es de interés para la investigación y para las dinámicas escolares en disciplinas como la lingüística, historia, religión, estudios latinos, estudios fronterizos, movimientos sociales y laborales, estudios de género, y más.

6. Patrocinadores

Este proyecto es posible gracias al apoyo del centro de humanidades digitales de los latinos estadounidenses (USLDH) y del programa de recuperación de la herencia literaria hispana (Recovery). El portal de periódicos de la región fronteriza entre los Estados Unidos y México es auspiciado por el consejo de recursos bibliotecarios y de información (CLIR) y por el programa de digitalización de colecciones especiales ocultas y archivos, generosamente patrocinados por la fundación Mellon.

7. Cosas que hay que saber

Metadatos

Este proyecto se alimenta de dos tipos de metadatos descriptivos, empleando hojas de cálculo de Excel. El primer set de metadatos es a “nivel periódico” y el segundo es a “nivel ejemplar”. El nivel periódico incluye una fila de datos para cada ejemplar de cada periódico. Esta práctica no es la norma en Recovery, sino que fue implementada específicamente con base en las conversaciones con la desarrolladora web. Contar con dos conjuntos de datos nos permitió desarrollar el sitio en entregas, y así poder probarlo e identificar problemas en cada página (búsqueda por periódicos y búsqueda por ejemplares). Esto también nos permitió alimentar las colecciones más grandes con los datos a nivel periódico más fácilmente, y al mismo tiempo mantener aparte los detalles del nivel ejemplar. El reto de mantener más de una hoja de cálculo es, por supuesto, asegurarse de que ambas estén igualmente actualizadas.

Reconocimiento óptico de caracteres (OCR)

Nuestro sitio ofrece la transcripción descargable en .txt de cada página de cada periódico. Logramos producir estos archivos pasando cada imagen escaneada por un software de OCR. Debido a la naturaleza bilingüe del proyecto, y las condiciones específicas de cada periódico (algunos dañados y muy frágiles), no obtuvimos transcripciones precisas de todos los

periódicos; sin embargo, insistimos en que una transcripción perfectible es mejor que ninguna. Nuestro sitio ofrece acceso a todos los documentos generados con OCR.

Desarrollo web y apoyo técnico

Los retos más significativos que enfrentamos fueron en cuanto a equipo y tecnología, de ahí que enfaticemos en la necesidad de una relación cercana entre la desarrolladora web y el equipo de apoyo técnico. Este proyecto se vio fortalecido una vez que entendimos qué tipo de preguntas debemos hacer a nuestros compañeros de apoyo técnico y de información, y que la desarrolladora web siempre debe participar en estas conversaciones. Puedes asomarte a este pdf para leer la lista de preguntas.

8. ¿Qué sigue?

La desarrolladora web que contratamos para este proyecto nos ha dado instrucciones para el mantenimiento y lanzamiento del sitio, para que Recovery pueda ir agregando más contenido año con año. También, nos hemos propuesto crear una cuenta de Creative Commons para el proyecto y subir información al repositorio digital de la universidad de Houston, repositorio de investigación de acceso libre (ROAR) para garantizar el acercamiento. En Recovery también queremos continuar con la promoción del portal en conferencias y presentaciones dentro y fuera de Estados Unidos.