Inside Look:
Digitizing a Historic Card Index

by Sarah Cunningham
Cataloging Assistant,
State Library of Oregon
sarah.cunningham@slo.oregon.gov
@cunningedge

and

Angela Jannelli
Oregon Documents Specialist,
State Library of Oregon
angela.jannelli@slo.oregon.gov

and

Heather Pitts
Cataloging Services Librarian,
State Library of Oregon
heather.pitts@slo.oregon.gov

Sarah Cunningham is a Cataloging Assistant at the State Library of Oregon. She has been with the State Library for 12 years and held several positions in Technical Services. She's looking forward to completing her MLIS in December 2020 from San Jose State University School of Information. In her free time, she enjoys crafting, building LEGO sets, and playing video games.

Angela Jannelli is the Oregon Documents Specialist at the State Library of Oregon. She has worked at the State Library for 10 years and holds an MLIS from the University of Washington. Outside of work, she loves gardening, playing the harp, and raising foster kittens.

Heather Pitts is the Cataloging Services Librarian at the State Library of Oregon. She has worked at the State Library for 11 years. She holds an M.L.S. from Emporia State University. She enjoys reading, watching cooking shows, and knitting or crocheting.

Introduction

Oregon Index Online (https://digital.osl.state.or.us/islandora/object/osl:or_index) is a resource for discovering information about the news, events, and people who shaped Oregon. It builds on the decades of work that went into creating the physical Oregon Index. This article reviews the methods library staff took to digitize and process nearly 800,000 cards to make the Oregon Index available online.

The physical Oregon Index at the State Library of Oregon.

Brief History of the Oregon Index

The Oregon Index is made up of nearly 800,000 hand-typed index cards housed in 657 wooden drawers in the second-floor catalog alcove in the State Library Building. It is the joint effort of several institutions and groups including the State Library of Oregon, volunteers, and local organizations such as the Daughters of the American Colonists—Mahonia Chapter.

The Oregon Index was physically divided into two parts: Biography (indexing of Oregon people—196 drawers) and Subjects (indexing of Oregon topics other than people—461 drawers).

The cards were created between 1913 and 1987. Newspaper indexing covers articles from roughly 1900 to 1987. Book indexing covers people and topics from the beginning of Oregon history, such as early travelers on the Oregon Trail.

This is a citation index. For example, index cards for newspaper articles include basic citation information (article title, newspaper source, date, and pages). The full-text of articles, books and other sources of information is not included.

The majority of the cards cover selective indexing of articles published in the major daily newspapers of Portland (*Oregon Journal, Oregonian*) and Salem (*Oregon Statesman, Capital Journal,* and *Statesman Journal*).

The Oregon Index also covers specialty newspapers such as the *Oregon Farmer* and *Willamette Week;* statewide and local magazines, journals and newsletters such as the *Oregon Voter;* and selected books, microfilmed historic documents and visual materials included in State Library collections when the index was created (Hegeman, 2019).

Scanning Collaborations

For decades, the way to access the Oregon Index was in person at the State Library building. Since at least 2002, library staff wanted to provide online access to the Oregon Index, but figuring out the method and tools needed—as well as actually doing the work—took considerable time. The scanning was accomplished through two partnerships over several years. The processing and preparation of the cards once scanning was complete took even more time than the scanning itself.

Sweet Home
Swan Island
Survey
SURPLUS PROPERTY
Pool cues to cars on state 'for sale' list. Statesman-Journal. 7 N 85. p. 1B, c. 5.
JAVENS, DAREL
State sale called hughe success. Statesman-Journal. 8 N 85. p. 11A, c. 1-4.

One of the physical Oregon Index drawers with the cards and dividers.

The first step to being able to put the Oregon Index online was scanning the Biography section, which was completed through a project in partnership with the company Ancestry. An agreement was signed in February 2010. The materials for this project came from several State Library collections, including the cards from the Biography section of the Oregon Index, which was the top priority for the project. Ancestry provided a digital SLR camera mounted on an overhead copy stand and hired a contractor to perform the scanning of the cards on-site. The scanning began in April 2010 and was completed in 2011. The output from the Biography section was 244,044 color TIFF images.

The next step was scanning the larger Subjects section. Alice LaViolette, a State Library reference librarian, read about a project at the Natural History Museum of London to digitize the Global Lepidoptera Names Index (https://tinyurl.com/y3lvnpub), containing 300,000 cards. The museum partnered with the University of Essex Department of Computer Science to create the VIADOCS project. The cards were scanned using a desktop check scanner in 61 days (Beccaloni et al., 2003). After reading about this project, Alice posted a question to a mailing list for finance personnel at Oregon state agencies asking if any had a high-speed check scanner and if they thought such a project would be feasible. We received a favorable response from the Oregon Department of Revenue (DOR) that they would be able to fit this project in. DOR staff would have the opportunity to learn more about the features of their scanner, improve their scanning skills, and tighten workflow processes. Scanning began in June 2014. Library staff prepped the cards, loaded drawers of index cards into boxes and onto book trucks, wrapped the book trucks in plastic, and rolled the book trucks two blocks over to the DOR building in batches. Using their high-speed check scanner, the Department of Revenue provided both TIFF and JPEG images of the front and back of each card. They also printed the date scanned, the drawer number, and the sequential card number on the back of each card. The scanning was completed in September 2014. The output was 1,290,295 grayscale TIFF images, including card backs and divider cards.

Processing and Uploading the Files

Once the scanning at the Department of Revenue was complete, the first step in image processing began. We first organized the files for the Subject cards into 461 folders, corresponding to their drawer numbers in the physical index. DOR had scanned the backs of the cards and divider cards, but since these were blank, we decided they should be deleted. We needed a way to identify and delete these files as efficiently as possible. We ended up using Adobe Bridge (a companion to Photoshop) where we could arrange by file size, delete the bulk of the blank pages because of their smaller file size, and then scroll through the rest to find the stragglers.

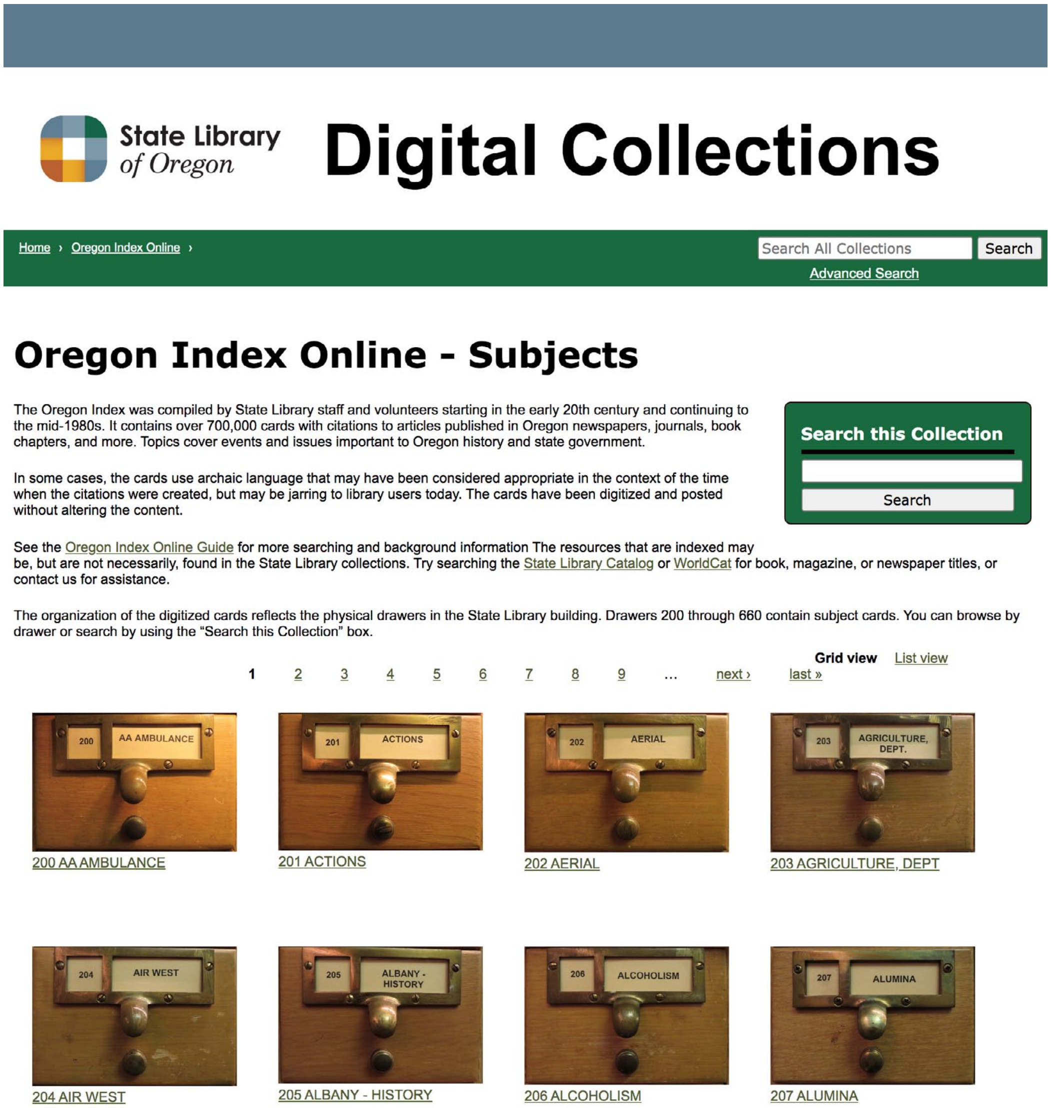
The scans of the Biography cards from Ancestry required some additional processing. They were extremely large-sized files in color with irregular black borders. Using Adobe Bridge and Photoshop, we converted them to grayscale, reduced their file size, and cropped the images. The edited TIFF files were then organized into 196 numbered folders as in the physical index.

In order to allow full-text access, the TIFF files needed to be converted into searchable PDFs. The Digital Collections Workgroup at the State Library had been using ABBYY FineReader for OCR (optical character recognition) of documents, which has a tool called ABBYY Hot Folder that allows for batch processing of files. Four to five folders of files were batch converted to OCR'd PDFs each evening. Evenings were chosen to make sure that the processing wouldn't interfere with daily work on staff computers. Each batch would create a log, which included the percentage of low confidence characters. While this wasn't exact since it is the software's best guess at its own errors, it gave us an estimate of how well the OCR performed and most fell within 12–20 percent low-confidence characters. We decided that access to the cards as-is was a greater benefit than waiting to edit the OCR of 785,667 cards, so the OCR correction would be the final step in the project.

At this point, the project hit a two-year pause, since the State Library was transitioning from a home-grown repository to a hosted option of Islandora, an open source Digital Asset Management System (DAMS). The migration to Islandora (and post-migration clean-up) needed to happen before we could figure out how to upload the Oregon Index.

While this work was happening, we made several design decisions. We wanted to keep the experience of the physical index, with the capability of browsing drawers as well as specific searches. We selected the Internet Archive book content model since it gave the user the experience of flipping through cards. Combining all of the cards from one drawer into one "book" seemed like an unruly size because each drawer contained anywhere from 750 to 1,400 cards. To solve that problem, we came up with 250 cards per book as a guiding number. This gave each online drawer three to six books of cards. To further support the design mimicking the physical index, we took photos of each drawer to be the image of each collection folder.

During the book creation, staff watched for errors from the batch OCRing process, like images being flipped upside down, split into two, or file corruption. We replaced these cards with newly created PDFs from the original TIFFs. The card files were combined using



Oregon Index Online - Subjects

The Oregon Index was compiled by State Library staff and volunteers starting in the early 20th century and continuing to the mid-1980s. It contains over 700,000 cards with citations to articles published in Oregon newspapers, journals, book chapters, and more. Topics cover events and issues important to Oregon history and state government.

In some cases, the cards use archaic language that may have been considered appropriate in the context of the time when the citations were created, but may be jarring to library users today. The cards have been digitized and posted without altering the content.

See the Oregon Index Online Guide for more searching and background information The resources that are indexed may be, but are not necessarily, found in the State Library collections. Try searching the State Library Catalog or WorldCat for book, magazine, or newspaper titles, or contact us for assistance.

The organization of the digitized cards reflects the physical drawers in the State Library building. Drawers 200 through 660 contain subject cards. You can browse by drawer or search by using the "Search this Collection" box.

Oregon Index Online made to mimic the physical index.

Adobe Acrobat Pro, then renamed manually with the title of the first and last card to give the range within each book. With 657 drawers and 785,667 cards, this was no small feat.

Uploading of the books was done overnight so as not to interrupt the daily work activities of staff using Islandora. The number of books uploaded per night was dictated by how many the system could handle without crashing or making all the books error out. Through trial and error, this ended up being 10 books for the subject cards and five books for the biography cards. Sometimes batches wouldn't upload for various reasons and books would have to be uploaded again.

Two Technical Services staff members, Sarah Cunningham and Angela Jannelli, worked on this process for three and a half years, creating books and uploading them into Islandora. In the end, we created 3,304 books.

OCR Correction

The OCR from the automatic processes described in the previous section is a great start, and fairly adequate for many of the cards, but we know that overall the OCR is imperfect and needs to be manually reviewed and corrected. There are lined cards that are particularly problematic, but any of the cards could have errors in the OCR.

We got a start on this clean-up project with library staff. First, starting in March 2019, one staff member piloted correcting the OCR file contained within Islandora. Then when the COVID-19 pandemic hit in March 2020 and many State Library staff started working from home, a few additional individuals began to work on the project in sections. To date, one drawer out of 657 has been completely reviewed and six other drawers are in-progress.

Our next step is to involve volunteers in the OCR correction. At the time of this writing, we are setting up a training in July 2020 for one volunteer who had been working on an in-person project in the State Library building that had to be placed on hold when the building closed in March 2020. In the future, we would like to structure the project to be able to recruit multiple volunteers.

Conclusion

Oregon Index Online is the culmination of years of work to make the indexing widely available. Through partnerships, and with the right tools, staff labor, and persistence, we were able to successfully bring this valuable information online for researchers and history enthusiasts.

References

Beccaloni, G., Scoble, M., Kitching, I., Simonsen, T., Robinson, G., Pitkin, B., … & Lyal, C. (Eds.). (2003). About the VIADOCS project and the construction of LepIndex. Retrieved from https://www.nhm.ac.uk/our-science/data/lepindex/aboutproject/

Hegeman, D. (2019). Oregon Index Online: a guide. Retrieved from https://libguides.osl.state.or.us/oregonindexonline

Acknowledgments

We thank the Oregon Department of Revenue staff, in particular Joseph Nations, Operations & Policy Analyst, who were willing to collaborate on an interagency project. We thank Alice LaViolette, who discovered the check scanner method and forged the partnership with the Oregon Department of Revenue.