OLAQ

OLA Quarterly

Volume 24
Number 4 *Digital Repositories and Data Harvests*

8-1-2019

Sharing Oregon's Cultural Heritage: Harvesting Oregon Digital's Collections Into the Digital Public Library of America

Julia Simic
University of Oregon

Ryan Wick
Oregon State University

Recommended Citation
Simic, J., & Wick, R. (2019). Sharing Oregon's Cultural Heritage: Harvesting Oregon Digital's Collections Into the Digital Public Library of America. *OLA Quarterly, 24*(4), 34-40. https://doi.org/10.7710/1093-7374.1962

OLA Quarterly is an official publication of the Oregon Library Association | ISSN 1093-7374

Sharing Oregon's Cultural Heritage:
Harvesting Oregon Digital's Collections Into the Digital Public Library of America

by Julia Simic
Assistant Head of Digital Scholarship Services, Digital Production and Preservation
University of Oregon Libraries
jsimic@uoregon.edu

and

Ryan Wick
Analyst Programmer, Oregon State University Libraries and Press, The Valley Library
ryan.wick@oregonstate.edu

Julia is the Assistant Head of Digital Scholarship Services at the University of Oregon Libraries. Her primary areas of responsibility include the management of all stages of the digital lifecycle, and participation in initiatives related to digital collections and scholarship projects. Julia assisted in developing training for the Orbis Cascade Alliance's LSTA grant to become a DPLA Service Hub and is the Institutional Representative to the Unique & Local Content Team. She holds a BA and MLS from Indiana University.

Ryan is an Analyst Programmer at Oregon State University Libraries and Press, with both the Special Collections and Archives Research Center and the Emerging Technologies and Services department. Beginning as a student worker in 1999, he has been involved in many digitization and digital collections projects, including publication of The Pauling Catalogue, OSU Sesquicentennial Oral History Project, ScholarsArchive@OSU, and Oregon Digital. He is also active in the Samvera and Code4Lib communities.

Oregon Digital, the library digital collections platform of Oregon State University and the University of Oregon, joined the Mountain West Digital Library (MWDL) and the Digital Public Library of America (DPLA) in 2016 to increase the visibility of our collections. This article discusses the process of becoming participants in the hub-network structure of the two organizations, remediating metadata in compliance with best practices, and modifications to the digital collections platforms, both locally and at MWDL, to successfully harvest over 100,000 items into DPLA.

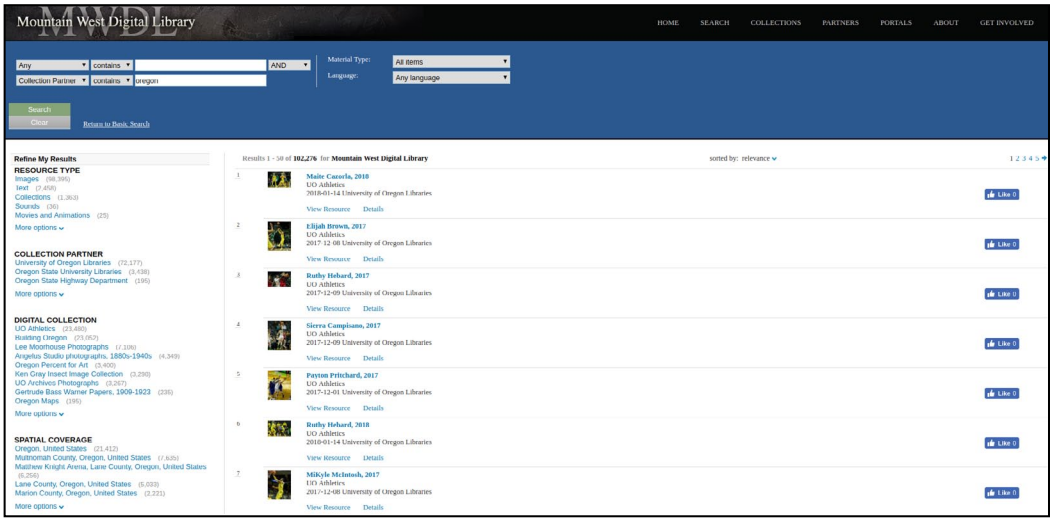
Background

Oregon State University and the University of Oregon have a longstanding and successful collaboration in providing access to unique digitized cultural heritage materials. Utilizing expertise from both institutions, Oregon Digital (n.d.) was launched in 2009 as a joint project on CONTENTdm, and later migrated to the Samvera (formerly Hydra) (n.d.) platform. Oregon Digital provides a single point of access for over 450,000 items in 116 discrete collections. Among our regional partners are the Greater Western Library Alliance, the Oregon State Historic Preservation Office, the Oregon Arts Commission, the Oregon Historical So-

ciety, and many others. In 2015, about a year after we migrated our collections to Samvera, we began to explore participation in the Digital Public Library of America (DPLA) (n.d.) as a way to share and promote our collections beyond the state of Oregon. DPLA member repositories make their digital collections metadata available for OAI-PMH harvesting. DPLA aggregates that metadata and makes it public and searchable through their interface. Actual content, such as images and documents, are not harvested; users clicking on an item in the DPLA interface are redirected to the original repository item. Partnering with the Mountain West Digital Library (MWDL) (n.d), a service hub for DPLA, was a natural extension of the collaborative spirit of both institutions.

Preparing Collections for Harvest

Although we began reviewing our digital collections for compliance with DPLA and MWDL content and metadata standards in late 2015, Oregon Digital officially joined MWDL as a single member repository in 2016. MWDL, based at the University of Utah, has a long relationship with DPLA, participating in the foundational Digital Hubs Pilot Project between 2012 and 2015, and has built partnerships with over sixty cultural heritage institutions. Their expertise as an established metadata harvester for DPLA was invaluable in assisting the Oregon Digital team through the technical challenge of making MWDL's Primo-based harvester work with our Samvera OAI-PMH output. Oregon Digital was, in fact, their first attempt at providing service to the Samvera platform, as most of their member repositories used CONTENTdm for delivering digital collections.



MWDL search results showing content from both UO and OSU.

In preparation for harvesting, metadata specialists at OSU and UO identified collections (or Sets in Oregon Digital) that could be contributed to DPLA and compared the Oregon Digital Metadata Dictionary (n.d.) to MWDL's Dublin Core Application Profile (Mountain West Digital Library, 2011), each documenting the metadata standards and fields that would be used in harvesting. Metadata remediation was necessary to meet MWDL/DPLA requirements. Some fields, such as Description and Subject were required by MWDL, but were not used always used in Oregon Digital. Other fields had incompatible data formats.

Building Oregon (Oregon Digital. Building Oregon, n.d.), one of the most popular collections from UO, contains over 4700 images photographed by former Dean of the UO School of Architecture Marion Dean Ross. Scanned from 35mm slide film, the only metadata we had about the images was what was written on the slide mount itself and what could be gleaned from its filing position in the physical collection. They lacked information appropriate for inclusion in the Description and Subject fields necessary for harvest into MWDL. To address this and similar complications in other collections, we had to take into account the subject matter and availability of staff who could add missing metadata, and the needs of the users of Oregon Digital and how they would discover the items through searching and browsing. In the end, most items were given "boilerplate," or generally applicable Descriptions and Subjects that took minimal staff time and required little Quality Assurance.

Inconsistent metadata, particularly in the Date field, also needed to be addressed. Agreement on a single input standard between collections, even within institutions, was non-existent. Once we decided on the machine-readable Extended Date Time Format (EDTF) specification (2018) and the level of support we would provide for it, scripts were written to search out and correct the formatting with little human intervention.

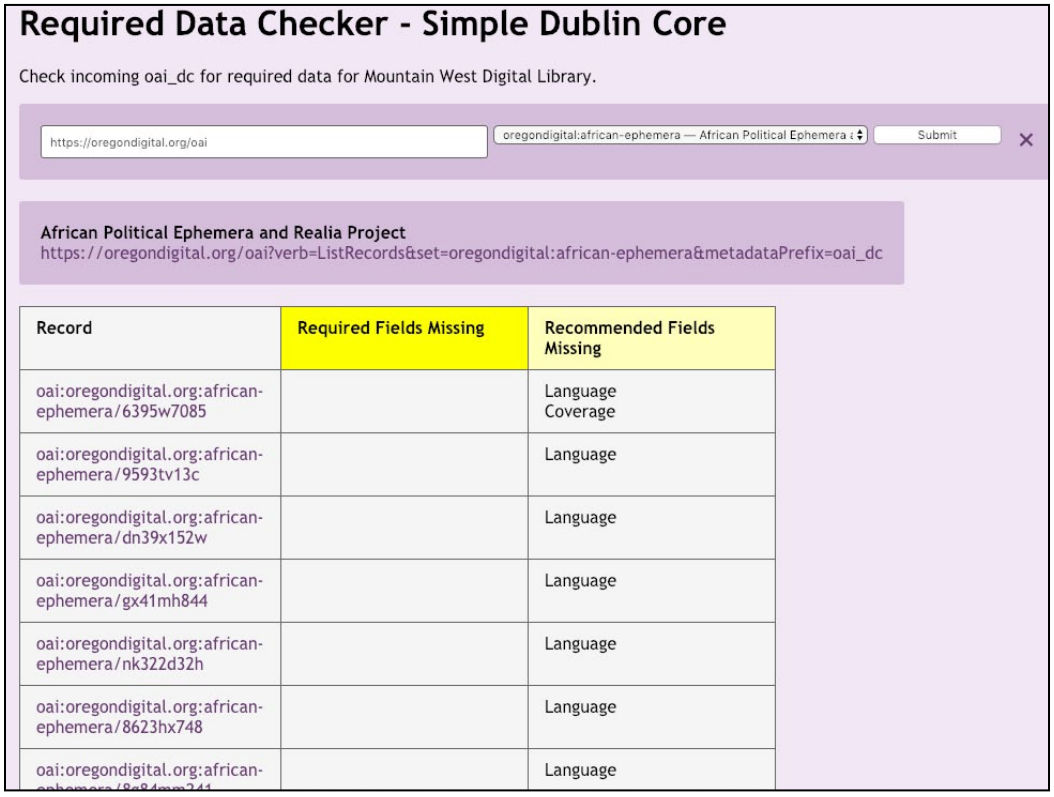
Several collections had items that used separate Earliest Date and Latest Date fields with values specifying a date range. For OAI output, these were collapsed into a single range value. EDTF date ranges gave us more flexibility, and MWDL agreed to adjust Primo to handle the ranges and parse them out for date values and facets. MWDL also knew that other partners were interested in using EDTF and Oregon Digital could serve as a pilot effort. This proved to be more involved than first anticipated, partly due to staff transitions, but ultimately was resolved with data normalization rules.

Compound or Complex Objects were also a challenge. These were used heavily by UO to represent physical archival folders, and at OSU to display items such as oral histories, audiovisual materials, and sometimes documents such as scrapbooks that have individual page descriptions. They manifested in Oregon Digital as parent metadata records to which child item records with content files were related. In early harvest tests both the parent and the child records were taken, resulting in some confusion in MWDL's Primo instance and their public search interface. After conversations with MWDL, we decided to make only parent records available for harvest by adding a metadata field, Primary Set, which functioned directly as the OAI Set and was applied to records selected for harvesting.

Oregon Digital makes heavy use of RDF and linked data. Fields such as Type and Rights were recorded in Oregon Digital metadata records as Uniform Resource Identifiers (URIs) that needed to be translated into text for the MWDL harvester. Record text labels are not stored in Fedora, so they were instead pulled from the Solr index and returned in OAI records. In a few cases, our label formatting was different than what was expected by MWDL. Our Region and Location labels, built from GeoNames (n.d.), separated the hierarchical levels with '>>' (i.e. Corvallis >> Benton County >> Oregon >> Pacific Northwest), but MWDL wanted commas as separators to match DPLA's metadata requirements. In our OAI specific code we could adjust the labels after they came out of Solr and leave Solr data as it was, not affecting the main Oregon Digital site.

Test Load

Five test collections were submitted to MWDL's Required Data Checker (Mountain West Digital Library, 2014) after metadata remediation. This tool, first provided as part of the DPLA OAI Aggregation Tools and modified to meet the requirements of MWDL's Dublin Core Application Profile, gives item-level feedback on the presence of metadata in required fields. We used this feedback as the final step of Quality Assurance for metadata remediation of the test collections, and cleaned up anything we missed earlier. When that was completed, an initial harvest of these collections was performed, and technical difficulties, including with the OAI provider response, could be addressed. Simultaneously, remediation began on more collections for harvest.



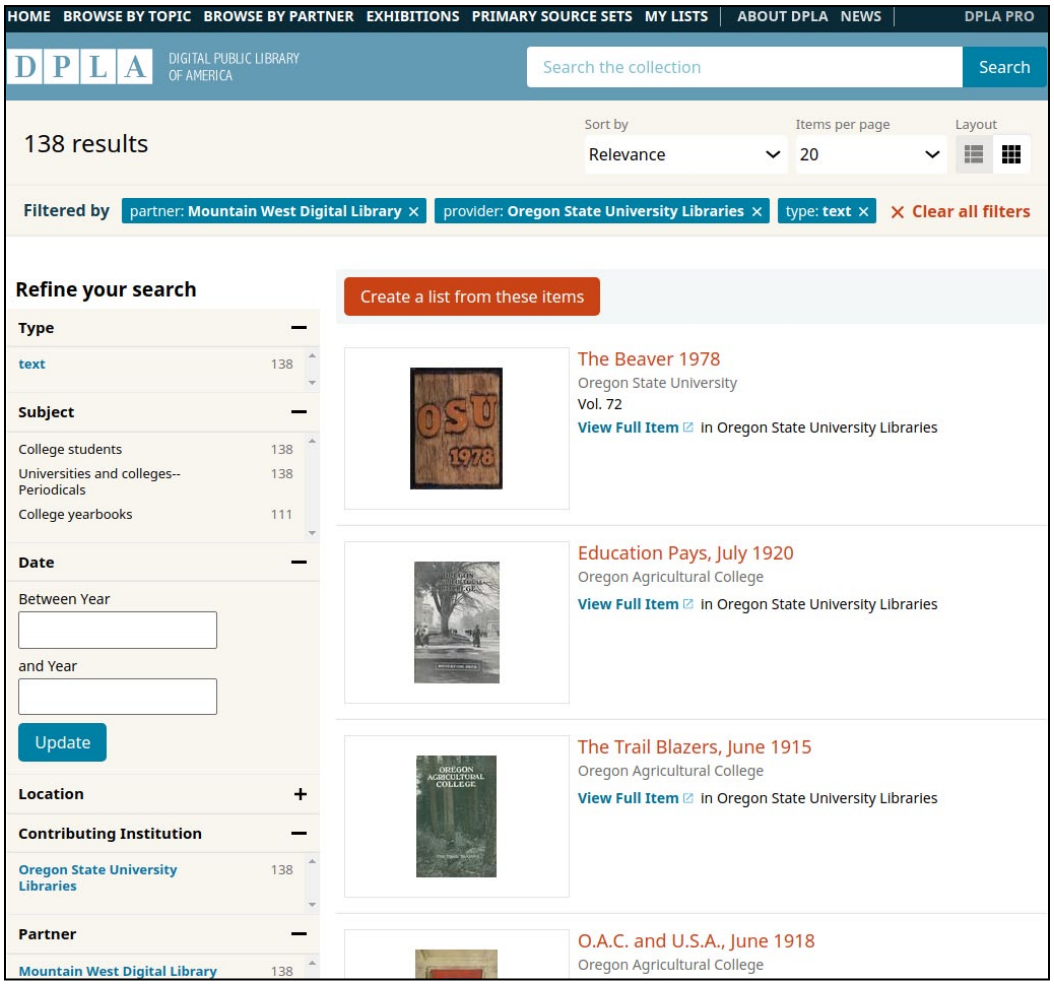
MWDL's Required Data Checker reviewing one of our collections.

OAI support was provided by adding the Ruby OAI gem (Code4Lib, 2015) to the Oregon Digital Ruby on Rails application, integrating OAI commands and responses. A few small parts of code from the gem were overridden in our application based on MWDL/DPLA metadata requirements. One instance of this was modifying the OAI record identifier code to return a value that included the collection or Set identifier in it; this is used by Primo to determine the OAI Set in the item record. Another example was modifying the OAI XML result to not include any empty metadata fields.

Our first implementation of OAI in Oregon Digital did a full lookup of items from our Fedora backend when requested, in order to return all of an item's metadata for processing. An OAI ListRecords request to show 100 items could take a minute or more to return a

response. For a full harvest, this would obviously not scale, and harvesting a single collection would take several hours. We changed the code to pull metadata and labels out of our Solr index instead, as this already powered the Oregon Digital public user interface, which was much more performant.

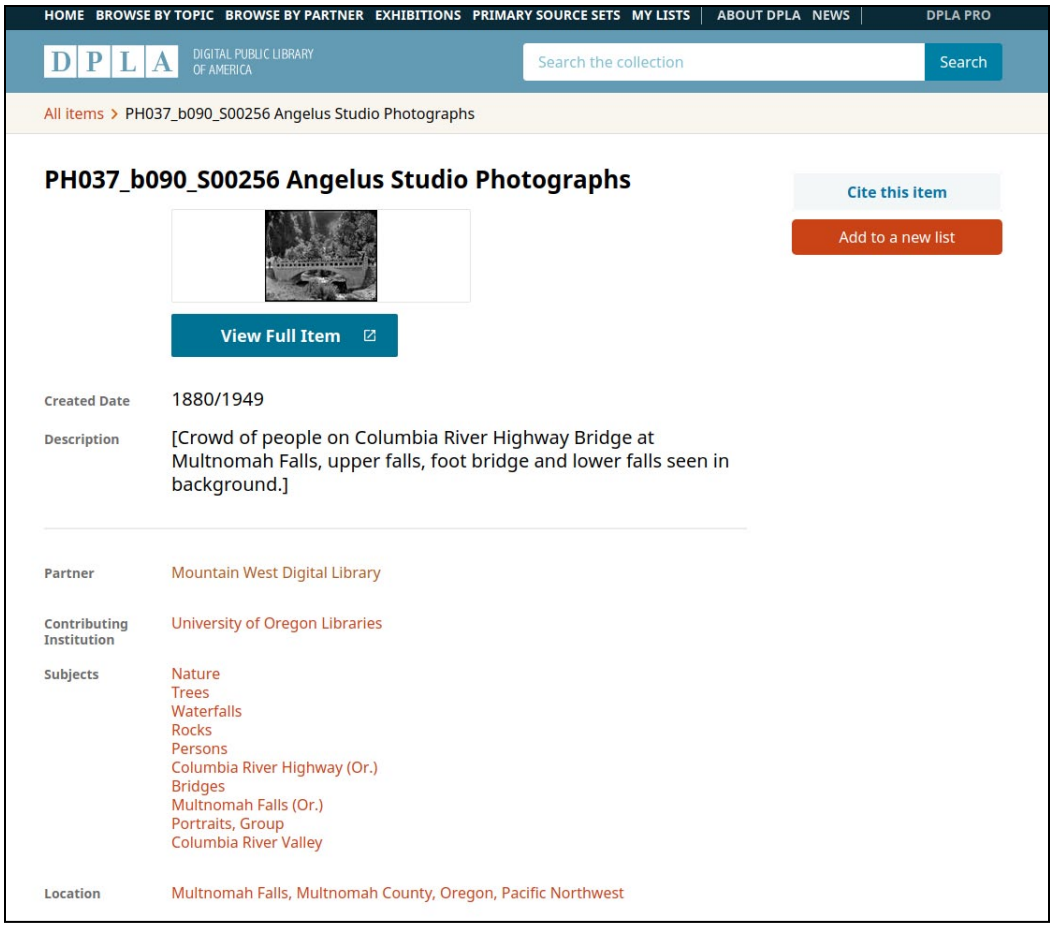
Providing thumbnail images for harvest also required configuration. Our images are stored on disk in folders that are organized based on parts of the item pid (permanent identifier). While this is consistent and reproducible, it didn't make sense to try and implement the folder rules in an external program. We built a new Rails controller in the Oregon Digital application to handle thumbnails when another system only had the item pid value. An image request with the pid value resolves and returns the correct thumbnail URL. This allowed MWDL's Primo to use a thumbnail template for any item harvested.



DPLA search results showing OSU publications.

Conclusions

Preparing and configuring our OAI endpoint and results took more work and time than was initially expected, but we knew it was important and necessary to get right. Furthermore, because we had full control of our application, we could make all of the changes that were needed, including making our legacy content better. Our initial goal was getting content into DPLA and MWDL, but there are other aggregators, including the Orbis Cascade Alliance, that we have worked with in the past and may again in the future.



DPLA item view with thumbnail and metadata.

Our partnership with MWDL has also led to participation in community efforts beyond the Orbis Cascade Alliance. Metadata librarians participated in the Western Name Authority File Project (Myntti & Neatrour, 2016), a pilot for creating linked open data name authorities for regionally significant people, and the Bulk Digitization Interest Group, a place to share standards and technical infrastructure for large-scale digitization projects. Our technical work has also contributed to the Samvera open source community. Developers actively participate in the Samvera Metadata Interest Group, the Applied Linked Data Interest Group, and in Samvera application development.

As members of the Orbis Cascade Alliance, OSU and UO have worked with the Digital Collections in Primo Group and the Dublin Core Best Practices Standing Group of the Unique and Local Content Team, assisting in preparing the Alliance itself to become a DPLA Service Hub. The experience we gained from participation in MWDL and DPLA has greatly benefited us and our sister institutions, and provided a valuable opportunity to grow our knowledge and practice of digital collection building.

References

Code4lib. (2015, May 8). Code4lib/ruby-oai. Retrieved June 10, 2019, from https://github.com/code4lib/ruby-oai/

Digital Public Library of America. (n.d.). Retrieved June 10, 2019, from https://dp.la/

Extended Date/Time Format (EDTF) Specification. (2018, October 22). Retrieved June 10, 2019, from https://www.loc.gov/standards/datetime/edtf.html

GeoNames. (n.d.). Retrieved June 10, 2019, from https://www.geonames.org/

Metadata Task Force of the Digitization Committee of the Utah Academic Library Consortium. (2011, July 20). Mountain West Digital Library Dublin Core Application Profile. Retrieved June 10, 2019, from https://mwdl.org/docs/MWDL_DC_Profile_Version_2.0.pdf

Mountain West Digital Library. (2011, July 11). Mountain West Digital Library Dublin Core Metadata Application Profile. Retrieved June 10, 2019, from https://mwdl.org/docs/MWDL_DC_Profile_Version_2.0.pdf

Mountain West Digital Library. (2014, May). DPLA OAI Aggregation Tools 1.0. Required Data Checker - Simple Dublin Core. Retrieved June 10, 2019, from http://dpla-aggregation.sandbox.lib.utah.edu/reqdata_checker/index_oai_dc.php

Mountain West Digital Library. (n.d.). Retrieved June 10, 2019, from https://mwdl.org/

Myntti, J., & Neatrour, A. (2016, May 17). Western Name Authority File Project. Retrieved June 10, 2019, from https://sites.google.com/site/westernnameauthorityfile/

Oregon Digital. (n.d.). Retrieved June 10, 2019, from https://oregondigital.org/

Oregon Digital. Building Oregon. (n.d.). Retrieved June 10, 2019, from https://oregondigital.org/sets/building-or/

Oregon Digital Metadata Dictionary. (n.d.). Retrieved June 10, 2019, from https://tinyurl.com/y3zypmpr

Samvera: An open source repository solution for digital content. (n.d.). Retrieved June 10, 2019, from https://samvera.org/