Digitization in the Real World

Lessons Learned from Small and Medium-Sized Digitization Projects

Edited by
Kwong Bor Ng & Jason Kucsma

Metropolitan New York Library Council

Published in the United States of America by
Metropolitan New York Library Council
57 East 11th Street, 4th floor
New York, NY 10003-4605
p: (212) 228-2320 f: (212) 228-2598
Web site: http://www.metro.org

ISBN: 978-0-615-379998-2

Cover Design: Jason Kucsma *(illustration by Smartone Design, licensed via iStockphoto.com)*

Reviewers Committee: Mark F. Anderson, Jill Annitto, Anna Craft, Jody DeRidder, Renate Evers, Wei Fang, Maureen M. Knapp, Sue Kunda, Mandy Mastrovita, Ken Middleton, Emily Pfotenhauer, Mark Phillipson, Alice Platt, Mary Z. Rose, Stacy Schiff, Jennifer Weintraub, Andrew Weiss.

The views expressed in this book are those of the authors, but not necessarily those of the publisher.

Collaborating for Success: A Cross-Departmental Digitization Project

Sue Kunda (Oregon State University Libraries)

Abstract

In 2007 Oregon State University Libraries acquired the personal collection of Gerald "Jerry" Williams, a native Oregonian and former national historian for the U.S. Forest Service. The collection contains personal papers, historic images, serials, monographs, oral histories, maps, moving images, political posters, ephemera, and artifacts of the U.S. Forest Service and Civilian Conservation Corps. Four library units collaborated to catalog, digitize, and make available online more than 1700 (and counting) items from the collection. In this chapter are descriptions of the planning process, workflows, policies and procedures. The author documents technical and procedural obstacles, methods used to overcome these difficulties, and lessons learned in working through the barriers.

Keywords: Academic libraries, Collaboration, CONTENTdm, Copyright manifest, Data dictionary, Digitization projects, DSpace, Dublin Core, Gerald W. Williams, Institutional repositories, ScholarsArchive@OSU, Workflows.

Introduction

In 2007 the Oregon State University Libraries (OSUL) acquired the personal collection of Gerald "Jerry" Williams, a native Oregonian with a lifelong passion for Pacific Northwest history. Williams began his career with the U.S. Forest Service in 1979, working as a sociologist with the Umpqua National Forest located in southern Oregon. After stints with the Willamette National Forest and later the Pacific Northwest Regional Offices, Williams was appointed, in 1999, national historian for the U.S. Forest Service. Throughout his career Williams was a prolific author and avid collector of Pacific Northwest historical documents, images, and artifacts.

The Gerald W. Williams Collection includes 35 years of Williams' personal papers related to his more than 75 publications, 3100 monographs and serials, and over 6000 copies of documents from the papers of Gifford Pinchot, the first U.S. Forest Service chief. In addition, Williams collected more than 24,000 historic photographs, including photographic prints, postcards, sterographic images, and glass lantern slides. The voluminous collection also includes oral histories, maps, moving images, political posters, and ephemera from the U.S. Forest Service and the Civilian Conservation Corps. (2008). After inspecting the collection and consulting with on-campus stakeholders about its value to the OSU academic community, OSUL purchased everything but several works of art and a group of miscellaneous artifacts.

Once the collection was transported to OSU library administrators agreed digitization of the textual items and historic images would take precedence over the digitization of other components. Many of the monographs and serials had no copyright restrictions and the OSUL Digital Production Unit (DPU) already had a well-established digitization workflow for printed matter. The historical photographs were compelling and had considerable potential value for researchers. In addition, University Archives and DPU worked together on a daily basis to digitize and make available online photographs and other images.

Digitization Project

Project Pre-Planning

With the help of Digital Access Services staff and student workers, the subject librarian for Forestry went through Williams' shelf lists and notes for the library portion of the collection to ascertain which titles were currently owned by OSUL and/or which titles were possible candidates for digitization. The Forestry subject librarian created a detailed spreadsheet documenting this information and a more general digitization matrix indicating the number of holdings at OSUL and number of items eligible for digitization. These documents were shared with an administrative team (University Archivist, Digital Access Services Head, Cataloging Unit Head, DPU Head, Digital Production Librarian) in order to gauge library resources. The Cataloging Unit Head was assigned project manager while the University Archivist and Digital Production Librarian were given oversight of the digitization of the images and textual items, respectively.

The Cataloging Unit Head facilitated a meeting with Cataloging and DPU staff to hammer out workflows and responsibilities. Attendees brainstormed ideas, devised two separate cataloging workflows, and agreed to do a one-week pilot test of each set of procedures to determine which was more efficient. Workflow #1 required Cataloging staff to organize the library portion of the Gerald W. Collection into the four groups in the Digitization Matrix (Public Domain/Not Held at OSU, Public Domain/Held at OSU, Non-Public Domain/Not Held at OSU, Non-Public Domain/Held at OSU) before beginning the cataloging process. Workflow #2 allowed Cataloging staff to begin cataloging without organizing the collection according to the Digitization Matrix. DPU staff also developed digitization workflows that coordinated with cataloging workflows.

The University Archivist and Head of Special Collections perused the library portion of the collection and flagged items deemed rare and/or valuable. After processing, these items would reside in the OSUL Special Collections rather than the general circulating collection.

Digital Repository Platforms

OSUL currently supports two digital repository platforms: DSpace and CONTENTdm.

DSpace: an open-source institutional repository platform developed by MIT and Hewlett-Packard, was originally designed to manage, archive, and provide access to textual documents making it an obvious choice for the monographs and serials in the collection. DSpace provides full-text searching through the Lucene search engine, which is especially conducive for print materials, and its contents are routinely crawled by search engine spiders. DSpace also offers permanent URLs, licensing, flexible metadata, and multiple digital preservation features.

CONTENTdm: a proprietary-based digital content management system first developed at the University of Washington and later purchased by OCLC Online Computer Library Center, was originally conceived to store and provide access to digital images. The software automatically generates thumbnails, JPEG, and JPEG2000s from TIFF digital files. Users can perform basic editing functions (resize, rotate, sharpen, crop), pan and zoom JPEG 2000 files, bookmark favorites, and create personal slideshows. Because of these more image-based capabilities, OSUL chose to use CONTENTdm to store and display the historic photographs, slides, and artifacts from the Gerald W. Williams Collection.

Metadata

Textual Items

The Digital Production Librarian created a data dictionary for the Gerald W. Williams Collection's textual items, which was appropriate for both monographic and serial titles. The dictionary specifies the metadata elements, their definitions as well as their use policy (e.g., required or not, repeatable or not, etc.) Most of the metadata elements are derived from the qualified Dublin Core.

These approximately 900 titles would be housed in ScholarsArchive, OSUL's instance of DSpace. Dspace supports the Qualified Dublin Core Metadata Schema. Most descriptive metadata for the collection was fairly standard (creator, title, date, etc.) but several adaptations were made to align this collection with OSUL unique needs.

While the Gerald W. Williams Collection's titles were grouped under a collection by the same name in ScholarsArchive, OSUL also used the description field (dc.description) to add a note designating them as Gerald W. Williams Collection titles. By doing this, users finding an item through a search engine rather than entering directly through the ScholarsArchive user interface, would see the item came from the Gerald W. Williams Collection.

The Digital Production Librarian also added metadata elements specific to Oregon Explorer, OSUL's natural resources digital library and its associated portals. Oregon Explorer metadata generally consists of geographic and spatial information relevant to the state of Oregon, but a relation field (dc.relation) also allows items to be pulled into the appropriate Oregon Explorer portal. For example, "Bohemia Mining District: A Brief History" was assigned the relation field "Explorer Site::Land Use Explorer", allowing it to be included in search results in the Land Use Explorer.

Dspace automatically attaches administrative and structural metadata.

Images

The images from the Gerald W. Williams Collection are housed in CONTENTdm, which also supports the Dublin Core Metadata Schema; therefore, the metadata for this collection is similar to the textual metadata. Standard descriptive metadata (creator, title, and date) was used, but OSUL also added geographic metadata such as Rivers and Streams, Hydrologic Unit Codes (HUC's), and Longitude Latitude Identification (LLID).

Technical metadata describing the digitization process is found in the Transmission Data field and Administrative Metadata for this collection is found in the Restrictions and Contributing Institution fields.

Digitization Specifications

Textual Items

Textual items that could be debound were scanned in both tiff (600 dpi master files) and pdf (300 dpi access files) formats on a Canon DR-9080C. This production scanner has a feed capacity of 500 sheets and scans at rates of up to 90 pages-per-minute. Tiff files were scanned as individual files while pdf files were scanned as a single document. This allows future manipulations of single tiffs, if necessary, and provides a pdf file more appropriate for web viewing.

Textual items that could not be debound were scanned on a Bookeye 2 Planetary Book Scanner located in the Interlibrary Loan Department. Unlike the Canon DR-9080C, which can scan a document in either single-page or multi-page format, the Bookeye 2 can only scan documents as single images. In addition, this scanner does not have the capacity to scan at 600 dpi, the resolution OSUL normally requires for master (tiff) files. Because these files are used to recreate an access copy, if necessary, and not considered the preservation copy, a lower standard is acceptable. All textual pdf files were compressed and ocr'ed using cvision's pdfcompressor™ 4.0.

Images

A University Archives volunteer used an HP4370 ScanJet to scan all images (photos, postcards, etc.) currently included in the Gerald W. Williams Collection. Items equal to or smaller than 3 x 5 inches were scanned at 800 dpi while items larger than 3 x 5 inches were scanned at 600 dpi. The volunteer created both grayscale and color tiff files for each item and the University Archivist chose the image he felt was of the highest quality. The volunteer used Adobe Photoshop Elements 4.0 to make any necessary edits.

Workflows

Cataloging

Cataloging staff had devised two possible workflows for processing the collection and planned to spend one week following each set of procedures to determine the most efficient workflow. It quickly became apparent that the more expedient workflow was the second one, which did not require any further organizing of the collection before cataloging. The first workflow was dropped.

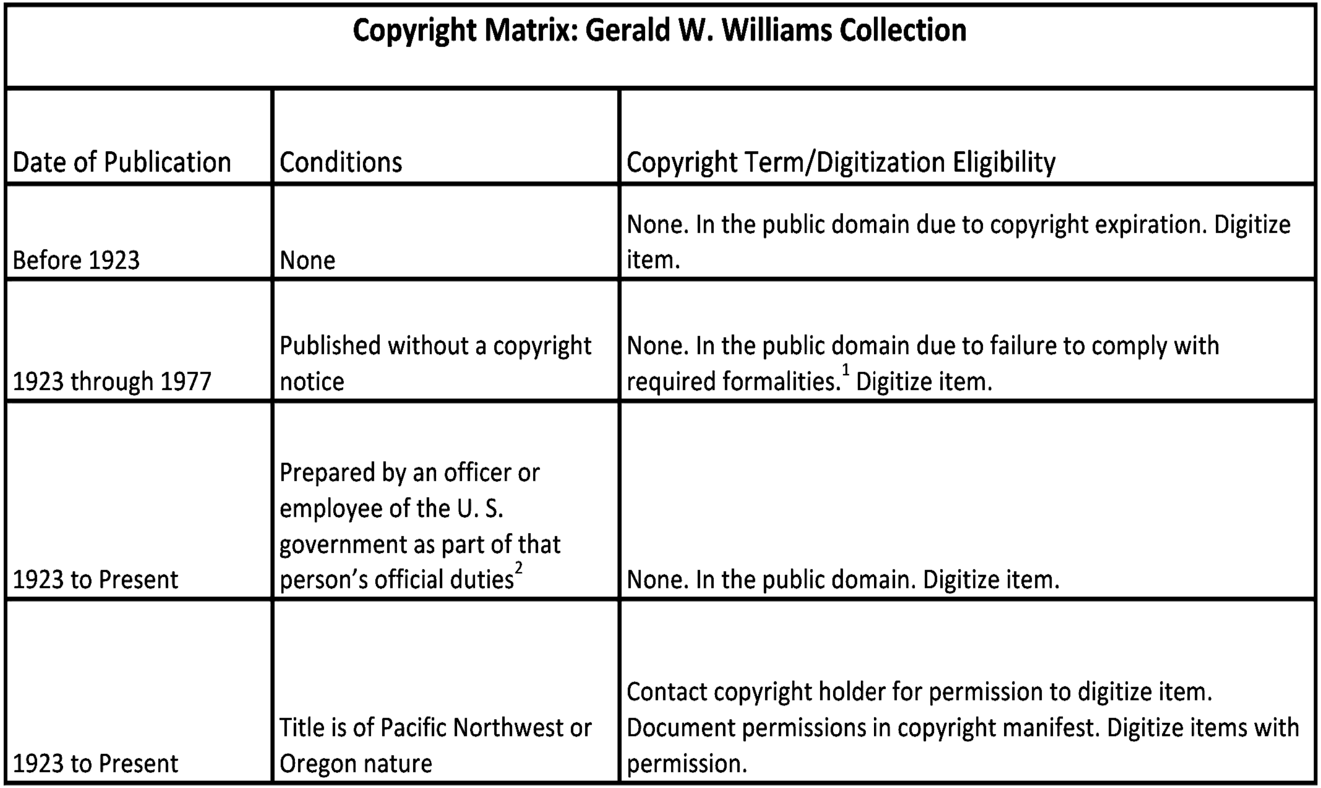
Library technicians responsible for adding monographs and serials to the library collection started by retrieving items, one book cart at a time, from University Archive compact shelving. Working from the book cart the cataloging technician checked the library OPAC to see whether or not the item was currently held at OSUL and, if not, cataloged the item. Additionally, the cataloging technicians used a variety of online sources (Google Books, Internet Archives, organizational websites) to ascertain whether or not the title had already been digitized elsewhere.

If a digital copy existed, and it met OSUL general quality standards, the cataloging technician inserted the appropriate information into the record and either sent the piece out to the circulating collection (items not held by OSUL) or the gift shelves (items already held by OSUL). The cataloging technician then placed items not found online or found online but of a very low quality, on a shelf marked "Gerald W. Williams Collection" outside the Digital Production Librarian's office.

Either the Digital Production Librarian or a trained library technician checked each item's copyright information to determine if it was eligible for digitization. Figure OREG-1 (next page) outlines the matrix used to determine each item's digitization eligibility.

The Digital Production Librarian created a copyright manifest to track copyright decisions and contact information. This Excel file was kept on a shared server so anyone involved with the project could access it at anytime.

Items not eligible for digitization were returned. Items eligible for digitization were placed on one of two shelving units. If the title was new to OSUL, and could not be debound because it was going to be added to the collection, it was placed on a shelf for scanning on the planetary book scanner in ILL. Items already in the OSUL collection, and could therefore be debound, were placed on a shelving unit reserved for sheetfed scanning.



Copyright Matrix: Gerald W. Williams Collection

Date of Publication	Conditions	Copyright Term/Digitization Eligibility
Before 1923	None	None. In the public domain due to copyright expiration. Digitize item.
1923 through 1977	Published without a copyright notice	None. In the public domain due to failure to comply with required formalities.[1] Digitize item.
1923 to Present	Prepared by an officer or employee of the U. S. government as part of that person's official duties[2]	None. In the public domain. Digitize item.
1923 to Present	Title is of Pacific Northwest or Oregon nature	Contact copyright holder for permission to digitize item. Document permissions in copyright manifest. Digitize items with permission.

Figure OREG-1. Digitization eligibility matrix

Textual Digitization

A DPU student scanner retrieved items that could be debound from the appropriate shelf. After debinding the item the student scanned the text block on the Canon sheetfed scanner and saved it as both individual tiff files and a multi-page pdf file. Hardback covers were also scanned and saved as individual tiff and pdf files. The student used Adobe Acrobat 9 Pro to combine the covers and text block to make one complete pdf file. The pdfs were then compressed and ocr'ed using cvision's pdfcompressor™ 4.0.

An ILL student scanner retrieved items that could not be debound and needed to be scanned on the ILL planetary scanner from the appropriate shelf. The student scanned each single page and double-page spread into one tiff file and used BScan ILL 2.0 software (packaged with the Bookeye scanner) to split double page spreads, deskew any crooked pages, crop the borders, and remove any thumb images captured during the scanning process. The student checked the quality of the resulting tiff file, made any necessary modifications, and used Adobe Acrobat Pro to convert it to pdf. Both files were saved and placed into a folder on a DPU server.

Textual items, both bound and debound, were added to ScholarsArchive using the customized DSpace metadata. Because each item had already been cataloged and added to the library's OPAC the student simply copied and pasted the information from the catalog into the metadata textboxes and attached the corresponding compressed and ocr'ed pdf file to the record.

After submitting the digital copy to ScholarsArchive the student returned the physical piece to the cataloger who had performed the initial work on the item. The cataloger checked the student submission, uploaded the item to the Internet, and added the ensuing URL to both the library's OPAC and WorldCat catalog. Debound items were recycled while bound items were added to the library's collection.

As of this writing all items destined for scanning on the sheetfed scanner have been digitized and added to the collection. Items that need to be scanned on the ILL planetary scanner continue to be added at a rate of approximately two per week.

Image Digitization

Each week the University Archivist pulled eight to fifteen images from the Gerald W. Williams Collection prior to the volunteer scanner's arrival on Wednesday. He completed a Word table for each weekly session, which included the image file number (written on the back), scanner settings, and descriptive metadata. The volunteer created two or three digital files of each image based on the provided specifications, saved all images to a jump drive, and placed it on the University Archivist's desk.

After the University Archivist reviewed the scans and chose the best images, a University Archives student worker gave the originals to a DPU staff member and copied the files to a shared drive located on a workstation in the DPU. This allowed the DPU staff member to pull a copy of the file off the server and into CONTENTdm's Acquisition Station. She used the original image when creating the necessary metadata, uploaded the item, returned the original to University Archives, and transferred the file to a permanent server.

Project Challenges, Successes and Lessons Learned

Challenges

It is impossible to take on a project of this magnitude and not encounter obstacles. Listed below are several challenges OSUL encountered during the process and the methods used to overcome those "bumps in the road".

Challenge #1: The DPU had limited experience scanning older monographs such as those found in the Williams Collection. Image printing processes used during the production of the books made it difficult to produce a high-quality copy of both images and text using standard DPU procedures. Trying to balance the desire for high-quality images with a fast-paced production environment – and without appropriate software – frustrated student workers and staff.

Solution #1: Unfortunately, this is still an obstacle when scanning older monographs. Moire patterns appear on some photographic images, making them less than ideal, but rarely does it affect whether or not an item can be viewed effectively. The Digital Production Librarian is now working to purchase scanning software that will lessen, or completely remove, these imperfections.

Challenge #2: Project organizers wanted all books eligible for digitization to be scanned, but the only DPU scanner that could be used for unbound books was a flatbed scanner. It would have been extremely inefficient if used for this purpose, especially with a number of books containing more than 300 pages.

Solution #2: DPU and ILL had a history of collaborating to solve digital access issues and of sharing student workers during calendar breaks. When approached by the Digital Production Librarian about the possibility of using the ILL planetary scanner for the Gerald W. Williams Collection, ILL staff agreed without hesitation. The two units worked out a schedule and workflow that allowed ILL students to digitize the books, edit the digital files, and save them to a DPU server.

Challenge #3: Student workers had limited access to the ILL planetary scanner due to its heavy usage during the workday and its continual malfunction.

Solution #3: Student workers were allowed to scan beyond the eight-to-five workday and on weekends. After repeated repairs the vendor eventually replaced the planetary scanner.

Challenge #4: Because of the complexity, detail, and number of people involved in the project, it was nearly impossible at times to remain consistent with workflow policies and procedures.

Solution #4: The project manager scheduled regular "check-in" meetings to review and, if necessary, update procedures. As the project progressed and staff became more familiar with practices and workflows, meetings were scheduled further and further apart until they were eventually discontinued altogether.

Challenge #5: Original workflow designs did not account for a copyright-checking step between catalogers and student workers nor identify a location for catalogers and students to place processed or digitized items.

Solution #5: Catalogers and DPU staff created additional workflow steps to include copyright checking and assigned shelves for items at various stages in the digitization process.

Successes

Digital Access Services, the department tasked with completing the vast majority of this project, is well known for managing and accomplishing large multi-step endeavors without much fanfare. The Gerald W. Williams Collection was no exception; the work was completed efficiently even as staff continued handling normal day-to-day responsibilities. Other measures of success include:

Success #1: Five library units (Cataloging, Digital Production, University Archives, Interlibrary Loan, and Special Collections), 17 staff members, and nine student workers collaborated to add more than 3,000 valuable, historic volumes to OSUL, 850 digitized monographs and serials to ScholarsArchive, and nearly 1000 images to the online Gerald W. Williams Collection. One would expect difficulties coordinating policies, procedures, and workflows with a collaboration of this size, but other than the complications surrounding the planetary scanner, few project disruptions occurred.

Success #2: Download statistics for the textual items in ScholarsArchive illustrate the amount of interest in this collection, both nationally and internationally. In 2009 the 686 items in the collection received nearly 40,000 downloads from more than 100 countries (Usage statistics retrieved December 10, 2009, from Gerald W. Williams Collection https://ir.library.oregonstate.edu/jspui/handle/1957/9112.)

Success #3: On February 14, 2009, timed to coincide with the 150th anniversary of Oregon's statehood, OSUL made its debut in

Flickr Commons with a launch of close to 125 images from the Gerald W. Williams Collection. The series, focusing on the Civilian Conservation Corps, was an instant hit with more than 8000 views within the first twenty-four hours. As of December 10, 2009, OSUL has added an additional 156 items and plans to continue increasing to the collection on a regular basis.

Success #4: Oregon Public Broadcasting twice visited to access historic pictures for two episodes of Oregon Experience. Civilian Conservation Corps chronicles the story of the New Deal program by the same name, and The Logger's Daughter explores the history of African American loggers in northeast Oregon.

Lessons Learned

Every large digitization project brings with it a unique set of challenges and obstacles. Adhering to a few basic principles can often mitigate the impact of these barriers:

Lesson #1: Pre-plan, pre-plan, pre-plan. While it may be tempting to skip the pre-planning phase, especially if an organization has previously undertaken numerous digitization projects, this step cannot be overemphasized (see Lesson #5). For most sizable digitization projects OSUL organizes and completes an inventory of the items in the collection, creates one or more data dictionaries, establishes workflows, and conducts meeting to discuss policies and procedures with appropriate personnel.

Lesson #2: Involve those people responsible for completing the work in the pre-planning phase. Students and staff on the front lines often have a better understanding of, and a more efficient method for, completing tasks. Their insights can be invaluable.

Lesson #3: To identify potential obstacles, process a few "test" items through the entire workflow prior to formally starting a project. It is much less disruptive to modify policies and procedures before the project gets fully underway.

Lesson #4: Do not start a project of this magnitude at a time when key decision-makers are unavailable. There are always questions and workflows modification early in the process, and not having the person(s) with the authority to make those decisions not only disrupts progress but can also frustrate workers.

Lesson# 5: Be (and stay) flexible. No two digitization projects are the same and each one has its own idiosyncrasies. Be prepared for policies and procedures from previous projects to fall short when applied to another. When obstacles do arise – and they will – encourage others' suggestions and keep an open mind when considering possible solutions.

Lesson# 6: Ask for help. Many cultural heritage organizations have considerable experience with digitization projects and most are willing to share their strategies and techniques with others.

Future Plans

As of this writing, OSUL has approximately 125 volumes to scan and add to the Gerald W. Williams Collection in ScholarsArchive, which will bring the total number of items in the institutional repository to just less than 1000. The image collection in CONTENTdm now houses 910 images. The digitization of textual items will most likely be finished in 2010. OSUL will continue digitizing the more than 24,000 historic photographs for many years to come. As time permits the University Archives staff will pull together groupings deemed of interest based on professional judgment and patron requests, and place them in the standard digitization queue.

Another possible future digitization project from the Gerald W. Williams Collection include Williams' own personal papers. His collected working papers, unpublished manuscripts, articles, and conference papers are valuable research documents in and of themselves. Finally, with Williams still collecting and now donating many of his acquisitions to OSUL, it appears the Gerald W. Williams Collection could be growing for a long, long time.

Supplementary materials can be found in ScholarsArchive@OSU at http://hdl.handle.net/1957/16758

References

Copyright Information Center. (2009). *Copyright term and the public domain in the United States, 1 January 2009.* Retrieved November 2, 2009, from: http://copyright.cornell.edu/resources/publicdomain.cfm

Oregon State University Libraries, University Archives (2008). Guide to the Gerald W. Williams Collection, 1855-2007. Retrieved November 27, 2009, from http://digitalcollections.library.oregonstate.edu/cdm4/ client/gwilliams/index.html

Peterson, K. (2009). *OSU enrollment jumps more than 8 percent to nearly 22,000. Retrieved November 28, 2009,* from http://oregonstate.edu/ua/ncs/archives/ 2009/nov/osu-enrollment-jumps-more-8-percent-nearly-22000

Simmons, T. (2006). *OSU recognized as Oregon's leading research university.* Retrieved November 28, 2009, from: http://oregonstate.edu/events/newsevents/carnegie.html

Springer, M., Dulabahn, B., Michel, P., Natanson, B., Reser, D., Woodward, D. et al. (2008). *For the common good: The Library of Congress Flickr pilot project.* Retrieved November 27, 2009, from http://www.loc.gov/rr/print/flickr_report_final.pdf

Stauth, D. (2006). *OSU College of Forestry viewed as number one in North America.* Retrieved November 28, 2009, from: http://oregonstate.edu/dept/nce/newsarch/2006/Oct06/forestryrank.html

The Commons. (2009). Retrieved November 27, 2009, from flickr Web site: http://www.flickr.com/Commons?GXHC_gx_session_id_=6afecb2055a3c52c

United States Copyright Office. (2009). *§101. Definitions, Copyright law of the United States of America and related laws contained in Title 17 of the United States Code.* Retrieved November 2, 2009, from http://www.copyright.gov/title17/92chap1.html#101