Against the Grain

Volume 27 | Issue 2 Article 40

2015

Curating Collective Collections: Double Dipping Using Digitization Workflows to Acquire Print Preservation Data

Bob Kieft
Occidental College, kieft@oxy.edu

Amy Wood
Center for Research Libraries, Wood@crl.edu

Curating Collective Collections — Double Dipping: Using Digitization Workflows to Acquire Print Preservation Data

by **Amy Wood** (Center for Research Libraries) <Wood@crl.edu>

Column Editor: **Bob Kieft** (College Librarian, Occidental College, Los Angeles, CA 90041) <kieft@oxy.edu>

Column Editor's Note: *Many of the columns that have appeared in* ***Curating Collective Collections*** *have treated the reasons, procedures, and decision parameters for creating shared collections of print journals and monographs. To a one, participants in such projects acknowledge and sometimes lament their having to rely on incomplete, inconsistent, or inaccurate holdings data or to accept the risks of making retention commitments without being able to verify the condition or existence of the volumes retained. The policy decisions about the items partners will share and the number of copies to be shared, together with the financial, operational, and governance arrangements needed to sustain the retained collection, seem like the hard things to do in making a shared print agreement. But, as anyone who has ever used, let alone maintained the records in, a library catalog knows the devil, angel, or God (depending on their metaphorical preferences) is in the data details.* ***Amy Wood's*** *column raises the magic data curtain on shared print projects by arguing for taking the time to record data in standard forms for action over time and among systems. Like its sibling program for legal materials between* ***CRL*** *and the* ***Law Library Microform Corporation****, CERES is also important as an example of domain-based shared collection building and of the two-way street that projects can walk for digitizing print to increase access and using already-digitized materials to define a print archive. In the CERES context, readers will recall the recent announcement that the* ***National Agriculture Library*** *will affiliate with* ***ASERL*** *on physical journal archiving, thereby adding additional heft to efforts for securing future access to materials in the domain of agriculture. — **BK***

Librarians, scholars, researchers, and patrons live in a world connected by data stored and manipulated in databases called by a seemingly endless variety of names: catalog, discovery system, registry, knowledge base, etc. We need all of these in order to promote discovery, less mediated access, and more resource sharing among institutions. For librarians participating in print archiving or shared print collections, recording granular gap or condition information at the issue or item level often seems an unnecessary luxury, but I want to argue with this case study that the cost of recording the granular metadata is a long-term investment that will improve and ensure access to and management of the collection regardless of current trends of metadata tagging and formatting.

The **Center for Research Libraries (CRL)** has learned from experience managing its general collection and its **JSTOR** print archive that item-level information is essential for knowing precisely what is in the collection, for enabling automated collection comparison and development, for sharing data with multiple catalogs or registries, and for addressing future unknown data needs. Tools that help create an efficient workflow in validating and recording the data make it easier and more cost effective to produce granular gap and condition data for print archives and shared print collections. **CRL's** Project CERES offers a model that can be adapted to a variety of projects for producing and recording granular data.

Project CERES Background

Project CERES[1] is a collaborative effort between the **Center for Research Libraries**[2] (**CRL**), the **United States Information Network**[3] (**USAIN**)., and the **Agriculture Network Information Center**[4] (**AgNIC**); it couples print archiving with digitization for access. The idea of the project was conceived from **CRL's** 2010 Institute of Museum and Library Services[5] grant-funded project, *Cooperative Print Archiving by Discipline: Developing an Infrastructure to Sustain Scholarly Resources.*[6] This two-year project has created a sustainable and scalable plan for cooperative management of legacy print materials at the local, state, regional, and national levels in the field of law as well as agriculture as discussed here.

In 2012, **CRL** began working with the **USAIN** preservation committee to develop Project CERES' goals, governance, and a process for choosing projects on which to work. Two primary goals were established: supporting consensus-based, cooperative archiving of agriculture resources and expanding electronic access to these important resources.

The initial focus of preservation and digitization has been:

- The extensive body of serials and government publications on agriculture, rural life, and home economics published between 1820 and 1975 that have been digitized and/or microfilmed under the **USAIN** program.
- Other agricultural and related trade and industrial journals published in the U.S. and Canada.
- Serial publications published by the U.S. agricultural extension services and experimental stations.

Project CERES runs on an annual cycle and operates under **CRL's** Global Resources Partnerships.[7] **CRL** provides $50,000 a year in funding for all Global Resources projects combined. CERES is governed by a subcommittee, under the **USAIN** preservation committee, comprising members of **USAIN** and **AgNIC**. The committee guides the priorities within the overall scope, develops the guidelines and process for participating in Project CERES, and chooses how funds are spent each year.

In the first year, August 2013-July 2014, thirteen participants preserved, digitized, and shared metadata for approximately 50 titles composed of roughly 10,500 items. In the current year, eight participants are working on a similar number of titles and items. These are significant numbers considering the first year's participants had a budget of $3,125 each and the average budget of the current year is $5,600. (Each phase had one participant drop out of the project due to staffing changes.)

Project CERES Preservation and Access Data

Data is an important output of Project CERES. **CRL** developed the data and data disclosure requirements for Project CERES to work with existing successive entry cataloging rules,[8] which track major title changes and shared print metadata disclosure[9] standards developed during the **OCLC** Print Archives Disclosure Pilot[10] project. Adhering to industry standards is crucial for optimal sharing of records and information between catalogs and registries that disclose holdings committed to preservation or shared print programs. Participants are required to:

- create title and issue level metadata,
- disclose holdings in **OCLC's** Worldcat and **CRL's** PAPR database,
- provide free access to digital versions via local digital asset management systems and **CRL's** digital delivery service, and
- make the digital versions available for archiving with the **National Agriculture Library**.

Title Metadata

Participants are required to create MARC bibliographic records for both the print and the digital versions. The MARC record for the digital version includes a hyperlink directing users to the digital resource's URL. Participants using digital asset management systems also create metadata records for those systems. No project standards have been set for these records, although participants often used Dublin Core. For the most part, participants are using existing print records from their library catalogs, but if there are no existing records or

continued on page 73

Curating Collective Collections
from page 72

if the library had not previously tracked major title changes, new records have to be created. Existing records also have to be upgraded to current cataloging standards, if necessary. Participants are encouraged to request an International Standard Serial Number (ISSN) from the **U.S. ISSN Center**[11] for each title that does not already have an ISSN.

Granular Metadata

CRL developed a spreadsheet template to capture granular data about completeness and condition of holdings. The spreadsheet was designed using Microsoft Excel, but any software using tables or spreadsheets would work. Each column in the spreadsheet records a single category of information (see entire list below), which helps keep the data clean for aggregation and sharing in a variety of metadata formats. The spreadsheet also minimizes the effort of recording data by requiring entry of a simple yes or no response or page numbers. This approach also helps eliminate inconsistently entered descriptive terms.

Most of the terms for condition have been taken from the Preservation & Digitization Actions: Terminology for MARC21 field 583.[12] Fields included in the spreadsheet are listed in the tables below and in the examples on pg.74.

Metadata Compliance by Project CERES Participants

During the first year, project participants were all able to provide title (bibliographic records) and completeness data. Condition metadata was requested but not required in the first phase, but some participants provided the information. Although some participants were initially intimidated by the amount of data requested, many decided as they input that it was easier than expected and had immediate benefits. One participant reported that the library's archivist was thrilled when the print volumes were transferred to the archives with the metadata spreadsheet because no resources had ever been transferred to the archives with such detailed information. This metadata enabled the archivist to understand what was being transferred and where there might be condition issues to address. This made the process of verifying a complete transfer from library to archive much faster. Another participant found that scanning operators had made decisions about re-ordering pages in the scanned version for easier viewing of images that were meant to be seen in a horizontal layout; filling out the pagination on the metadata spreadsheet helped them catch those changes. Participants also found and recorded variances and inconsistencies with dates and enumeration of issues that were printed on the items.

Completeness metadata:	
Title and piece identification	**Page specific information**
• Journal Title • Additional Title (monograph titles), • Print OCLC# • Print ISSN • Series • Volume • Issue • Part • Publication Date	• Number of Pages • Title Pagination • Missing Pages • Covers, foldouts, etc. • Scanned Front Covers • Scanned Back Covers • Additional Pages

Condition metadata:		
• Highlighting/Underlining • Insect Damage • Loose (pages, covers, bindings) • Marginalia • Mold Damage • Obscured Text Block • Rebacked • Rehoused poorly	• Repaired poorly • Repaired soundly • Tight Binding • Torn • Warped/Cockled • Yellowed/Browning pages • Reprint • Binding pattern variations	• Printing Errors • Acidic Paper • Alkaline Paper • Brittle Paper • Faded • Foxed.

Additional fields to capture administrative metadata are also included to help manage the projects.

Colorado State[13] was one participant that incorporated the metadata gathering into the quality control steps of the overall workflow. Although filling out the gap and condition metadata was not something they had done for other digitization projects, they were able to exceed their expected preservation goals for the project by 22%. In their project proposal, they listed 100 items that would be preserved and digitized. They completed the digitization and metadata recording for 122 items within the project's single year timeline.

Model of Metadata Capture for Collective Print Archives

There are many elements of the project that can be adapted to other projects. It is important in a library environment to use MARC bibliographic records because that is what **OCLC's** Worldcat database and library catalogs and discovery systems use now. It is important to encourage participants to request unique ISSNs because a unique internationally recognized ID that transcends individual MARC records and possible duplicates is a key element in sharing data among databases and systems. Once the MARC record and ISSN are in place, the focus can be on recording granular metadata elements of enumeration variations, publication history, and gaps and condition in a flexible format that allows data to be easily transformed into a variety of formats for sharing. This will enable libraries to respond more quickly to system innovations of the future.

Using spreadsheets to record and manage data during the project gave participants the most flexibility and potential for accuracy with minimal training. Most library staff are familiar with using spreadsheets or tables at the level of entering data, and the format requires little training even if staff do not use tables or spreadsheets frequently. Part-time student workers often completed the metadata worksheet and did so with consistency. There are no tagging or field codes or data formatting and punctuation rules to learn (and re-learn each time the data is entered). Questions that surfaced when entering data were about inconsistencies recorded on the pieces themselves such as an incorrect enumeration or date printed on an issue. Resolutions to data problems encountered by one participant were easily shared among all participants via email. With everyone using the same spreadsheet, there were no additional software-specific data entry requirements that necessitated additional instructions tailored to the software. The spreadsheet has also helped **CRL** aggregate all of phase 1 participant data.

CRL is still in the process of aggregating the data for the first phase. Steps include: loading the MARC records to the **CRL** catalog, adding records to **CRL's** digital delivery system registry, creating MARC holdings records with 583 fields for commitment, gaps, and conditions according to **OCLC's** recommendations for disclosing print archive holdings, and loading the issue-level data into a database that stores the granular data at an item level. The granular metadata in the spreadsheet and existing tools enable us to do all of that.

Conclusion

There are many successful print archiving, shared print programs and collaborative

continued on page 74

Little Red Herrings
from page 71

such quellenforschung is also better done in print than in a myriad of distracting hyperlinks.

Of course, it isn't that digital natives or anyone else refuse to read online. Many love the ability to define words (though they likely forget them immediately), or to do quick key word searches. Some, though I admit to reading between the lines, also prefer being able to do searches in books they haven't read for materials they may need for a paper. Science materials, too, tend to be online favorites.

So, what are we to make of all this? As I have written elsewhere, it's part of the transition. In no way do I believe that this spells the end of online materials. Publishers, who

continued on page 75

collection management and programs upon which to model new projects. Project CERES offers a unique model in the capture of metadata that can be reproduced in other projects coupling digitization with preservation or a high level of validation without digitization. The flexible format for capturing individual elements of data in separate fields lends itself to modification based on data needs of a project producing even minimal validation. The focus of working with existing standards but storing the data in a format-agnostic database enables data and resource sharing. The ability to dip into the data well multiple times for multiple purposes is a major gain in efficiency and also lays the foundation for working with any future standards that may be developed.

Endnotes

1. Project CERES description on **CRL** Website: *http://www.crl.edu/collections/global-resources-partnership/global-resources-agriculture-partnership*.
2. **Center for Research Libraries** Website url: *http://www.crl.edu/*.
3. **United States Agriculture Information Network** Website url: *http://usain.org/*.
4. **Agriculture Network Information Center** Website url: *http://www.agnic.org/*.
5. Institute of Museum and Library Services Website url: *http://www.imls.gov/*.
6. **CRL's** Archiving by Domain: Agriculture Webpage url: *http://www.crl.edu/node/7371*.
7. **CRL's** Global Resources Partnerships Webpage url: *http://www.crl.edu/collaborations/global-resources-partnerships*.
8. CONSER's Cataloging Manual: 31.18, Changes that require a new record *http://www.itsmarc.com/crs/mergedprojects/conser/conser/module_31.18._changes_that_require_the_creation_of_new_records.htm*.
9. **OCLC's** Web page on shared print management: Detailed Metadata Guidelines: *http://www.oclc.org/services/projects/shared-print-management/metadata-guidelines.en.html*.
10. Final Report of the **OCLC** Print Archives Disclosure Pilot: *https://www.oclc.org/content/dam/oclc/productworks/OCLCPrintArchivesDisclosurePilotFinalReport.pdf*.
11. **U.S. ISSN Center** Website url: *http://www.loc.gov/issn/*.
12. Standard Terminologies for the MARC 21 Actions Note Field Webpage url: *http://www.loc.gov/marc/bibliographic/583terms.html*.
13. The author would like to thank **Beth Oehlerts**, Metadata Management Librarian, **Colorado State University Libraries**, for supplying the following data.

Sample Detail from Metadata Spreadsheet 1

METADATA FOR COLORADO STATE UNIVERSITY'S CERES PROJECT TITLES

JournalTitle	Additional title (monograph titles)	Print OCLC#	Print ISSN	Series	Volume	Issue	Part	Publication Date	Number of Pages
Press bulletin (Colorado Agricultural Experiment Station)	Seepage or return waters on the Uncompahgre River	5165452	n/a			15		1902	2
Press bulletin (Colorado Agricultural Experiment Station)	The prairie dog as a range pest : and methods of extermination	5165452	n/a			16		1903	2
Press bulletin (Colorado Agricultural Experiment Station)	Trials of macaroni wheat by dry farming : 1902	5165452	n/a			17		1903	3
Press bulletin (Colorado Agricultural Experiment Station)	Grasshoppers : their habits and remedies	5165452	n/a			19		1903	3
Press bulletin (Colorado Agricultural Experiment Station)	Plant lice and their remedies	5165452	n/a			20		1903	2
Press bulletin (Colorado Agricultural Experiment Station)	Spraying for plant lice and the codling moth	5165452	n/a			21		1904	2
Press bulletin (Colorado Agricultural Experiment Station)	A co-operative experiment in tree planting	5165452	n/a			22		1905	4
Press bulletin (Colorado Agricultural Experiment Station)	Fall handling of potatoes to lessen injuries from insects and fungi	5165452	n/a			23		1904	3
Press bulletin (Colorado Agricultural Experiment Station)	Formalin treatment of seed grain for smut	5165452	n/a			24		1906	2

Sample Detail from Metadata Spreadsheet 2

Title Pagination	Percentage overall of images (estimate)	Color Images	Photographs	Plates	Other Images	Missing Pages, Covers, foldouts, etc.	Scanned Front Covers	Scanned Back Covers	Additional Pages	Reprint
		No	No	No	No	Missing front and back cover	No	No		No
		No	No	No	No	Missing front and back cover	No	No		No
		No	No	No	No	Missing front and back cover	No	No		No
		No	No	No	No	Missing front and back cover	No	No		No
		No	No	No	No	Missing front and back cover	No	No		No
		No	No	No	No	Missing front and back cover	No	No		No
		No	No	No	No	Missing front and back cover	No	No		No
		No	No	No	No	Missing front and back cover	No	No		No
		No	No	No	No	Missing front and back cover	No	No		No