Analysis of CRL Validator Reporting - Bibliographic Records

WEST Unarchived Title Ingest for Cycles 12 & 13 Collections Analysis

Introduction

One of the major projects WEST undertook in 2021 to enhance AGUA was a project to improve metadata validation practices to support more rigorous data analysis by the AGUA Technical Team and to report back findings to WEST members to support local cleanup projects. To accomplish these goals, the AGUA team adopted a Metadata Validator developed by the Center for Research Libraries. WEST uses the CRL Validator to analyze Archivers' disclosure files as well as unarchived holdings files submitted by WEST members ahead of the biennial collections analysis. The Validator reviews metadata in the contributor record and compares it against the OCLC WorldCat database to identify inconsistencies, missing data, and incorrect data. Reports are used by the Tech Team to identify critical errors that will hamper the collections analysis as well as non-critical errors that may be of interest to members.

> *Note on validator reporting*: Before running the CRL Validator on WEST member files, the AGUA Technical Team 'preprocesses' files of bibliographic records, cloning bib records that have more than one location to create an individual record for each unique location that contains that location's holdings. Because of this preprocessing, there may be duplicate entries in the reports produced by the validator. Additionally, holdings fields are consolidated to create a single summary holdings string.
>
> *Note on metadata sources*: The primary OCLC number in the input record is searched in OCLC. If a match is found, data from that OCLC record is compared with the contributor's bibliographic record.

Data

This analysis was performed using the unarchived holdings files submitted by WEST members ahead of the Cycles 12 & 13 collections analysis. WEST received files for 58 OCLC symbols with a total of 1,771,712 records. A total of 494,888 errors were detected by the validator, 307,727 of which were 'critical' (impacting the WEST analysis).

Errors

The Metadata Validator identifies a number of errors, some of which WEST has designated as 'critical' (impacting the WEST collections analysis) and others WEST has designated as 'non-critical' (issues which are reported to members for their information and correction as local circumstances allow, but which do not impact the WEST collections analysis). 'Critical' errors are noted below.

Holdings Have No Years

Critical. This error is generated when no years are detected in the summary holdings string.

- *What this means for WEST*: This error is critical, because the WEST collection analysis tool uses years to determine depth of back-file. If no years are detected, then depth of back-file is equal to zero. Records with this problem will be included in the analysis, but WEST will propose this title for archiving by another member if available.

- *What this means for the data owner*: Members are not asked to update their records. Cataloging practices do not require that years are included in holdings notation unless years are on the piece being described. Some institutions add years in square brackets to aid in discovery, the square brackets indicating that the years are known, but are not on the piece itself.
- *Solution*: WEST should investigate means of calculating volumes which either supplements or replaces year detection.

ISSN Mismatch With OCLC

This error is generated when either a primary ISSN is lacking (022$a) or the value doesn't match with what's expressed in the matching OCLC record.

- *What this means for WEST*: WEST uses the OCLC number from the input record to search OCLC and harvest any relevant control numbers from the matching OCLC record, including ISSN (022al and linking data fields' $x.) Therefore an absent or incorrect ISSN is of no consequence, unless the record also lacks an OCLC number.
- *What this means for the data owner*: It is strongly encouraged that contributors update their records which either lack or have incorrect ISSN numbers, as this will affect local discovery, data control, and record management.
- *Solution*: The data owner should upgrade records to include/correct ISSNs where needed.

No OCLC Number

Critical. This error is generated when the input record lacks a primary OCLC number.

- *What this means for WEST*: WEST uses the OCLC number (either primary or merged value) to harvest an array of control numbers from the OCLC record, including OCLC numbers, LCCN, and ISSNs from both primary fields and from linking data fields. These values are used through various processes to match with Ulrichs to assign a Journal Family ID. Where a primary OCLC number is lacking, WEST is unable to access the full array of possible match points,significantly diminishing the chances that a record will match with a journal family in Ulrichs.
- *What this means for the data owner*: Members are strongly encouraged to update their records lacking OCLC numbers, not just for WEST but for local discovery, local data control, and record management.
- *Solution*: The data owner should upgrade records to include OCLC numbers where lacking.

Holdings Out of Range

Critical. This error is generated when detected holdings (years) fall outside of the matching OCLC record's fixed date fields in the 008.

- *What this means for WEST*: Detected depth of holdings may be deeper for that title than what is actually accurate. Volumes held by related titles (succeeding and preceding) may be lacking altogether, or less than what is accurate.
- *What this means for the data owner*: Local discovery of volumes held by individual titles will be compromised. Succeeding and preceding titles may be lacking from the catalog. While these titles may be added to a "primary" record for discovery purposes, the catalog will lack the added richness of these individual records.
- *Solution*: The data owner should review holdings and move volumes to existing records or new records as needed.

OCLC Merged Number

This error is generated when the input primary OCLC number matches with the correct record, but has matched on a merged OCLC number.

- *What this means for WEST*: This has no effect on WEST processes. WEST captures the entire OCLC number "set" in order that both primary and any merged (cross-reference) OCLC number can be used to match with the correct OCLC record.
- *What this means for the data owner*: The data owner need not make any changes to their records for WEST purposes. However, they may want to update these records with a fresh OCLC copy. A merged primary number may indicate that the record is old, and fresh copy would provide significant enhancements. If OCLC record ingest in a local catalog is not set up to consider both primary and merged numbers for match purposes, a merged primary may signal a need to review to be sure that the catalog does not contain "duplicate" bibliographic records that should be merged.
- *Solution*: None needed.

Serial Type Not Periodical

This error is generated when the matching OCLC record is coded for something other than periodical (not blank, \, m, p, -, |) See 008/21, continuing resources.

- *What this means for WEST*: We do not ask that materials be excluded by serial type.
- *What this means for the data owner*: Since we do not ask that materials be excluded by serial type, there is no action required.
- *Solution*: None needed.

Invalid Form of Item

This error is generated when the matching OCLC record is coded for something other than physical print, (blank, \, r, -) See 008/23, continuing resources.

- *What this means for WEST*: We ask that only records for physical print materials be submitted with unarchived title files. Form being something other than physical print, i.e., microform, electronic, online, etc., indicates that the record isn't a physical print description record. This may affect Ulrichs journal family matching, or other control number matching processes down the line, especially if these same records are used with disclosed WEST archived holdings.
- *What this means for the data contributor*: Multi-format records should not be used as they adversely affect downstream processes in organizations like WEST and HathiTrust. In some respects they make discovery easier for end-users, but in others they do not. For example, material type filters employed by end users may exclude online, microform, or print holdings depending upon the type desired. Print records, on the other hand, may merely be badly coded either locally, or also in OCLC. Note that correctly coded, non print records may have been submitted with non-print holdings, in which case there is no error. These holdings will generally be excluded using location criteria, so can be ignored.
- *Solution*: Data owners should correct badly coded records both locally and in OCLC where needed, and/or pick apart multi-format records and either move holdings to existing records, or to new records. Where records are non-print with non-print holdings attached, consideration can be given to excluding these in future.

Invalid Carrier Type

This error is generated when MARC 338 contains a carrier type that is not consistent with a hard copy print record. Carrier type and form of item are generally in sync. When they are not, it indicates a miscode in one or the other. When in sync, it will indicate that a non-print record was used with print holdings, or that non-print holdings were submitted.

- *What this means for WEST*: WEST asks that only records for physical print materials be submitted with unarchived title files. Carrier type indicating something other than physical print, indicates that the record isn't a

physical print description record. This may affect Ulrichs journal family matching, or other control number matching processes down the line, especially if these same records are used with disclosed WEST archived holdings.

- *What this means for the data contributor*: Multi-format records should not be used as they adversely affect downstream processes in organizations like WEST and HathiTrust. In some respects they make discovery easier for end-users, but in others they do not. For example, material type filters employed by end users may exclude online, microform, or print holdings depending upon the type desired. Print records, on the other hand, may merely be badly coded either locally, or also in OCLC. Note that correctly coded, non print records may have been submitted with non-print holdings, in which case there is no error. These holdings will generally be excluded using location criteria, so can be ignored.
- *Solution*: Data owners should correct badly coded records both locally and in OCLC where needed, and/or pick apart multi-format records and either move holdings to existing records, or to new records. Where records are non-print with non-print holdings attached, consideration can be given to excluding these in future.

Title Mismatch

Critical. This error is generated when the title in the input file differs significantly from the title in OCLC. Titles are considered to be matched if 90% of the characters in whichever is the shorter of the two matches the longer title.

- *What this means for WEST*: This is a critical error, as it indicates that the OCLC number in the input record is not the correct OCLC number for that title. Because of WEST's other processes in harvesting control numbers based on the input primary OCLC number, this will adversely affect matching with Ulrichs.
- *What this means for the data owner*: These should be reviewed and corrected as soon as possible by the data owner as this has such a profound effect on downstream processes for organizations such as WEST and HathiTrust, or any batch automation data exchange with OCLC.
- *Solution*: Review and correct any OCLC numbers and other bibliographic data associated with title mismatches as soon as possible.

Bib Level Not Serial

This error indicates that Leader byte 07, bibliographic level was something other than "s" serial in the OCLC record.

- *What this means for WEST*: WEST asks that data contributors only send serial titles. WEST does not drop these from loading or processing, however, as they tend to be miscoded rather than actual monographic titles.
- *What this means for the data owner*: Miscoded records may affect data discovery and data management. Filtering results sets for serials, for example, may not work correctly if records are miscoded as monographs.
- *Solution*: These records should be reviewed both locally and in OCLC and corrected where needed.

No OCLC Match

Critical. This error indicates that no match was found for the OCLC number given in the input record with OCLC.

- *What this means for WEST*: Since WEST relies so heavily on OCLC number, this is considered to be a critical error. It generally indicates that the number contained punctuation or other characters which kept it from matching. These may also be old Institution Record numbers (assigned to the local copy in OCLC). These records/numbers no longer exist.
- *What this means for the data owner*: Records lacking correct OCLC numbers in local systems can affect discovery, data control, and database management.
- *Solution*: The data owner should upgrade records to include/correct OCLC numbers where needed.

Invalid ISSN

This error indicates that the input ISSN failed the algorithm for checking valid ISSNs.

- *What this means for WEST*: Since WEST looks up ISSNs for each bibliographic record with a valid input primary OCLC number, this isn't an issue; WEST harvests the current ISSN set from OCLC. Furthermore, an "invalid" ISSN may be the currently assigned and matching/legitimate ISSN in the OCLC record.
- *What this means for the data owner*: While the ISSN may fail validation, it may be the correct ISSN in the OCLC record. Still, these should be reviewed and corrected locally and in OCLC where needed. Use of the ISSN Database and referencing the hard copy may be needed in order to do so.
- *Solution*: The data owner should review and correct ISSNs both locally and in OCLC where needed.

Rec Type Not Language Material

This error indicates that Leader byte 6 is not equal to a, so is not language material.

- *What this means for WEST*: WEST asks that data contributors only send print, language, serial titles. WEST does not drop these from loading or processing, however, as they tend to be miscoded rather than actual non-language materials.
- *What this means for the data owner*: Miscoded records may affect data discovery and data management. Filtering results sets for language material (as opposed to maps or notated music), for example, may not work correctly if records are miscoded as non-language.
- *Solution*: These records should be reviewed both locally and in OCLC and corrected where needed

Appendix: Summary of errors in Cycles 12 & 13 data set

Measure	Number of Records	% of Total Records
Total Records	1,771,712	100%
Total Errors	494,888	27.93%
Critical Errors	307,727	17.37%

Table 1: Summary of total records submitted and total and critical errors detected

Error Type	Records with Error	% of Total Records with Error
Holdings Have No Years	146,840	8.29%
ISSN Mismatch w/OCLC	99,869	68.01%
No OCLC Number	80,696	80.80%
Holdings Out of Range	67,201	83.28%
OCLC Merged Number	26,041	38.75%
Serial Type Not Periodical	20,799	79.87%
Invalid Form of Item	19,105	91.86%
Invalid Carrier Type	18,598	97.35%
Title Mismatch	12,003	64.54%
Bib Lvl Not Serial	1,444	12.03%
No OCLC Match	987	68.35%
Invalid ISSN	919	93.11%
Rec Type Not Lang Material	386	42.00%
Total Errors	**494,888**	**27.93%**

Table 2: Breakdown of error types, with total records with each error and the percent of total records they represent (critical errors highlighted in red)