

BIBLIOGRAPHIC QUALITY CONTROL IN WISCAT

Duplicate bibliographic records, authority control processing, and bibliographic quality control are three very separate issues that are often problematic in any large union catalog. The WISCAT union catalog brings together records from a variety of sources. Each contributing library generally has its own separate and sometimes conflicting cataloging and authority control decisions in place. Different procedures and considerations need to be in effect for a union catalog than might be needed in a local or shared automated environment.

Duplicate Bibliographic Records. The most difficult aspect to the art of database management is merging bibliographic records that originate from disparate systems into a single database without letting through too many duplicate records and yet being careful not to merge together editions or formats that should remain separate. To this end, WISCAT staff has worked very hard refining the matching algorithms used to match and merge bibliographic records.

Although WISCAT procedures empower local library staffs to take responsibility for their own holdings maintenance directly online, no local libraries are permitted to input or edit MARC bibliographic records directly online. All bibliographic record maintenance is done offline in a batch method by the WISCAT professional database management staff.

Each and every new file of MARC records that is received by the WISCAT staff, whether the file is a snapshot replacement file or includes just a few records on a diskette, is matched against the existing WISCAT database by OCLC number, by ISBN number, by ISSN number, by LCCN number, by music publisher's number (028), and then by a number of bibliographic data keys (title, publisher, format, date, etc.) to determine if a bibliographic record for that title already exists in WISCAT. If a matching bibliographic record exists in WISCAT, the holdings statement(s) on the incoming record are added to the existing master record in the union catalog. For those incoming records that cannot identify a matching record, a "no-hit" file is created. No-hit records are generally not automatically added to WISCAT without further clean-up and manipulation. Depending on the source, most incoming MARC data files tend to produce about a 90-92% hit rate with the WISCAT catalog. The no-hit files are then run through a variety of additional filter programs in an attempt to isolate acceptable records that can be added to the database.

The most effective way to eliminate duplicates from infiltrating a database that contains a large number of records from OCLC libraries is to use the OCLC control number found in the 001 field. For this reason, the OCLC control number is always the first and the most effective match point used to merge incoming MARC records to WISCAT. Since most of the OCLC libraries are larger libraries with dedicated professional cataloging staff, there is very little filtering of those records to monitor the quality of incoming records that contain OCLC numbers. However, deduping on OCLC control number field alone does not mean that all duplicate records will always be prevented from being introduced into the union catalog. Anyone who has used OCLC knows that it is not uncommon or incorrect for the same edition of a work to be cataloged on multiple records. Because these records have different OCLC control numbers, deduping on

control number will not identify them. Any duplicate records existing in OCLC can come unhindered into WISCAT.

A further area of concern in record control number deduping is caused by the dynamic nature of the OCLC Union Catalog. OCLC has defined the MARC 019 field to hold control numbers of duplicate records merged from the online system. While displaced or obsolete records are not retained in the OCLC online system, they continue to reside on many libraries' local files which are subsequently sent to WISCAT for periodic snapshot uploading. When this occurs, duplicate records are often reintroduced into WISCAT. WISCAT staff has worked with Auto-Graphics staff to enhance the processing utility software to recognize and merge control numbers in 001 with 019 fields. This utility upgrade should happen in the next few months. Even with 001/019 inconsistencies and issues, the OCLC control number remains the best and most effective deduping method, and for the most part results in accurate merging of bibliographic records.

Other control number deduping keys are also used to eliminate multiple occurrences of the same bibliographic title. For example, WISCAT uses both LCCN and ISBN keys to identify and eliminate duplicates. These additional control number keys are supplemented with information taken from the title field and other bibliographic data fields. This practice of adding other bibliographic data to the number keys is designed to reduce false matches.

Additionally, for records that lack any matching control numbers, it is necessary to adopt a non-numeric deduping key. Non-numeric deduping relies on the creation of a composite identification key to use in matching and merging incoming records. It combines fixed and variable field information, including data extracted from the title, imprint, media type, etc. The more bibliographic elements that are included in the match key (i.e. the more points that must match exactly), the greater the probability that incoming records will not match and will go to "no-hit" status; and since "no-hit" files can be added directly to the database as new records, this increases the possibility of introducing duplicate records that should have been merged with existing records. Conversely, if fewer bibliographic elements are used in the match key, then it is likely that more incoming records will be incorrectly merged with existing records and the possibility of merging different formats, large type, or microform records together exists. Non-numeric deduping means making trade-offs between precision and record volume and is not as effective as control number deduping. It is only used in merging records that have failed to match on one of the standard library control number fields.

In addition to the filtering and merging routines done on new and incoming records, a number of efforts are underway to merge and de-duplicate existing records that are already in the union catalog.

Authority Control. Retrieval is the "name of the game" in authority control processing. The purpose of authority control is to provide concise intellectual access to the records in the database. Sound authority control processing has little or no effect on duplicate database records. It is done to bring together records under consistent headings and to provide users with a roadmap to alternative and related headings through the cross-reference structure, but it seldom serves to remove duplicate records from a database.

A standard annual authority control run against the Library of Congress authority file has been a part of WISCAT processing routines since the microfiche days in 1989. In the fall of 2001 when the database was converted to the new web software, Auto-Graphics performed its first authority control process and cleanup of the WISCAT database. Another full authority control process was scheduled for early in 2003. That processing was postponed due to the more pressing need to add many snapshot replacement files in a short period of time following the discontinuation of OCLC archival files. Authority control processing has now been rescheduled before the end of 2003. WISCAT staff met with Auto-Graphics technical staff in September 2003 to plan the upcoming authority control processing.

The WISCAT union catalog contains records contributed from a variety of sizes and types of libraries, and thus it contains subject headings from a number of different subject heading schemes. Some records contain Library of Congress subject headings, some contain Sears Subject Headings, some contain medical MESH Subject Headings, and many contain subject headings that adhere to more than one scheme. All subject headings are indexed for searching purposes. Existing processing routines remove local, non-standard subject headings (MARC 690 fields) from all incoming bibliographic records. This has not always been the practice. There are local subject headings in the catalog that are scheduled to be removed with a future global batch run in 2003. There are also surely misspelled, obsolete, inaccurate and "just plain wrong" subject headings on some records that have been contributed to WISCAT. These subject headings are generally not corrected by routine subject authority control. Authority control will only convert subjects that are included as "see from" entries in a subject authority file. Subject headings that fall into this category are best identified and fixed through general bibliographic quality control processing.

Auto-Graphics employs a standard name authority control against the Library of Congress Name Authority File; and standard subject authority control against the Library of Congress Subject Heading (LCSH) Authority File, the Library of Congress Subject Headings for Children's Literature, the Canadian Subject Headings (CSH) Authority File; and the National Library of Medicine Subject Headings (MeSH) Authority File.

Authority records are not maintained within the WISCAT database. Thus, unlike some local and shared automated catalogs, all authority control processing in WISCAT is done in an offline batch on the full existing file rather than making attempts at authorizing records as they are being added to the database for current cataloging. Wisconsin's contract with the union catalog vendor allows for annual authority control processing.

WISCAT authority control processing automatically generates "See" and "See also" cross-references in the catalog based on 4xx and 5xx field data in the Library of Congress authority records. Authority control offers the catalog user a reference structure, leading through cross-references to related headings appropriate to research goals. "See" cross-references send users from headings they think the catalog might use to headings the catalog actually does use. "See also" cross references lead users from headings they locate in WISCAT to other related headings that may also be useful. Cross references are displayed in WISCAT within the Alpha Browse searching mode.

Bibliographic Quality Control. Bibliographic quality control of the database has been an ongoing effort since the early 1980's. Typically quality control processing includes review of existing bibliographic records for errors, misspellings, duplicate records, etc. In earlier years, systematic quality control checking was implemented on the WISCAT catalog. At this time, there is not adequate staffing to do systematic quality control clean up of the database. Quality control checking does continue as needed on problematic records as they are identified. It has always been the policy that if a cataloger "touched at record" for any quality checking, they should review and correct all problems in the record, not just the specific problem or misspelling they were trying to correct.

WISCAT staff has worked closely with Auto-Graphics technical staff to identify some problem data in WISCAT records and will be performing some global processing on the database to remove problematic fields (e.g. fields 090 and 092 for local call numbers, field 590 for local notes, etc.). Additionally, records for NetLibrary and other e-book or remote titles will be removed, and a normalization of General Materials Designations (GMDs) will be performed at the same time. Global cleanup processing of this sort is scheduled to be performed annually.

With the Auto-Graphics union catalog software that will be implemented this fall, there is a new feature called Union Database Maintenance Module (UDMM) available to the participating library staffs. This will allow local library staffs to participate in the quality control of the union catalog. They can use an online form directly tied to specific database records to report errors, duplicate records and other issues directly to the WISCAT staff for immediate correction.

WISCAT staff has always had access to client software that has allowed them to correct individual errors and do general cleanup of bibliographic records in the WISCAT database. This fall, WISCAT staff will have access to a new web-based bibliographic record editor product from Auto-Graphics that will facilitate database cleanup of individual problematic records. This software is not available to local library users of WISCAT. A small contract is maintained with a cataloging processing center (SALPC) that allows for experienced catalogers to assist in the review of existing records and initiate data fixes or merges to any identified records

Future Plans for WISCAT Bibliographic Integrity. The Division's plan for general quality control of the WISCAT union catalog is to continue with annual batch authority control processing against the Library of Congress name and subject and other standard authorities; to continue with control number (OCLC, LCCN, ISBN, etc) and bibliographic key filtering of incoming MARC files for merging of duplicate records; to continue to evaluate no-hit records on a library by library basis prior to inclusion; to continue to allow individual users to identify duplicate and/or inaccurate records as candidates for manual merging or cleanup; and to continue running periodic and specific de-duping runs of the full file based on standard control numbers or other factors.

Regardless of the variety of quality control issues and complexities, WISCAT remains one of the most bibliographically accurate large union catalogs in the country.