

# Matching Records When Batchloading to WorldCat

**Note:** This document is an overview; it does not include all of the details, anomalies, and exceptions that may be invoked during the record matching process. Any machine-matching algorithm, regardless of its sophistication, may make matching decisions that do not mimic the decisions humans might make when comparing records. Software does not always match records in the same way library staff do—some records are matched to records for disparate items while others are not matched to records for identical items due to insignificant differences in the records.

## I. Introduction

When matching bibliographic records against WorldCat for the purpose of adding or deleting holdings symbols, adding, replacing or deleting institution records, or for enriching WorldCat, records are processed through a variety of matching algorithms. The options selected for loading a file vary from one project to another depending on the fullness and accuracy of the bibliographic records and the purpose for which the records are being processed. Database Specialists review the files sent for processing and determine the most effective matching strategy.

### General rules as they apply to matching include:

- If a single match is found, no other searches are done to find others.
- If multiple matches are found, various additional criteria are used to discriminate between the records and to select the ‘best’ records.
- If data is present in one record but not the other, it usually has no effect on the matching. It counts neither as matched nor unmatched. There are some exceptions, however, in which the absence or presence of data causes records not to match.

## II. Steps in matching

There are two basic steps when matching bibliographic records:

- (1) Looking for candidates—i.e., searching WorldCat for possible matches
- (2) Determining whether or not the records match—i.e., comparing data to determine if the records represent the same manifestation

Several iterations of searching and matching may be attempted until a matching record is found.

### III. Finding matching candidates

OCLC uses several kinds of searches to look for matching candidates. Each project is defined differently to use the most appropriate searches. The Database Specialists analyze the contents of the file to determine the most efficient search strategy. Generally, once a single match is found, no additional searches are performed; if multiple matches are found, the matching software continues through other comparison elements to find the best match. If multiple matches remain after all matching is completed, additional software selects the best record from the set of matching records.

Optional qualifiers (see section IV) may be used to evaluate the records retrieved for each search. The options are designed to discriminate between records and eliminate unsuitable candidates. The Database Specialists select options they deem most useful and effective for each project.

- (1) **OCLC number matching.** If the record contains a validly constructed OCLC number, it is used to find a match if the Database Specialist has selected it in the project definition.
- (2) **Unique key matching.** Numbers that identify a manifestation or a record may be used to find matching candidate records. If more than one key is in the record and the Database Specialist has specified it be used, it is OR'd with other keys to retrieve the most comprehensive set of candidate records. Unique keys are described further in Section V.
- (3) **Extended matching.** If a single match has not been found during the unique key matching phase, extended matching is invoked if the Database Specialist has requested it. Extended matching looks at bibliographic data beyond unique keys to identify matches and to eliminate inappropriate candidates. In extended matching, up to five searches are executed using various combinations of terms. Terms from the following MARC tags are combined in various ways to find other candidates for matching. Any records found are added to those found in the unique key matching phase. The queries identify records that are candidates for matching to the input record. The following table lists data elements and MARC tags that may be used in searches during the extended matching phase.

Author	100 \$a
	110 \$a and \$b
	111 \$a
	130 \$a
	700 \$a
	710 \$a and \$b
	711 \$a
	730 \$a
	720 \$a

Title	245 \$a \$k \$b \$f \$n and \$p
Publisher	533 \$c
	260 \$b and \$f
	261 \$a, \$b, and \$e
	262 \$b
Date	008/07-10
Material Type	Subset of terms from the Format/Document Type index (dt=) created for WorldCat searching

#### IV. Qualifiers

Qualifiers are specialized comparison elements and each is optional—they may not be used for every file. Whether to use qualifiers is at the discretion of the Database Specialist who reviews the submitted file and determines which qualifiers are likely to be most useful in identifying matches. Decisions are based on the kind of data and accuracy of the data in the incoming records and on the reason for processing the file. Qualifiers may be selected or deselected independently and separately for any of the matching phases—OCLC control number matching, unique key matching, or extended matching. If any of the selected qualifiers do not match, the candidate record is rejected.

**(1) Language of Cataloging**

Determined by code in 040 \$b. If no code is present, the record defaults to 'eng' (English). If Language of Cataloging is selected for matching and does not match, the candidate from WorldCat is rejected as a match.

**(2) Date of Publication**

Attempts many comparisons using dates from the fixed field (008/07-10 and 008/11-14), field 260 \$c, 533 \$d, and 362 \$a.

**(3) Derived Title Key**

The matching software derives title keys from tags 245, 246, and 247 and compares them to each other. If any of them match, the Derived Title Key is considered a match. If they mismatch, more extensive title comparisons are performed.

**(4) Material Type**

Terms are generated as for the Material Type (mt=) index. Only terms used to describe the physical manifestation (e.g., VHS videotape, microfiche, Braille, etc.) are used for comparing the records. General terms (such as juvenile, government publication, etc.) are not used. If any of the designated terms match, Material Type is a match.

## V. Comparison Elements (Unique Key Matching)

In unique key matching, the following list of keys are considered 'unique keys' for matching. If any of them match, the record is considered a match.

OCLC	035 \$a
LCCN	010 \$a
ISBN	020 \$a 020 \$z (if no 020 \$a)
ISSN	022 \$a
CODEN	030 \$a
URL	856 \$u (if indicator 1 = '4')
Publisher Number (Scores and sound recordings only)	028 \$a  262 \$c
Other System Number	024 \$a
Report Number	027 \$a 027 \$z 088 \$a 088 \$z
National Bibliographic Agency Control Number	016 \$a

In a few instances, the Other System Control Number (field 029, \$a) is also used as a unique key.

Qualifiers, if selected, are used to deselect records from the set of retrieved records. Once a single record is found to match on one of these keys and the qualifiers match, matching is complete. If no matches are found or if there are multiple matches, the candidates may continue to extended matching. If there are too many candidates, no attempt is made to match the record.

## VI. Comparison Elements (Extended Matching)

When a record is not matched in the unique key phase, it is sent to extended matching if the Database Specialist has selected the extended matching option. In extended matching, several parts of the bibliographic records are compared in an attempt to find the best matches and to eliminate records for similar manifestations. These comparison elements resemble those that library staff use to distinguish records from each other. For example, if a cataloger sees that two records have different publishers, the cataloger determines they are not a match because they do not represent the same manifestation. In the same way, the matching software also rejects candidates with different publishers.

### **Comparing candidates**

Once candidates are identified, the records are compared. While there are some specialized rules for various kinds of materials (such as maps, sound recordings, etc.), most of the comparison elements are treated the same for all types of material. In addition, qualifiers (see Section IV) are applied if the Database Specialist has selected them.

### **Normalizations and comparisons**

Each comparison element is normalized and compared in a variety of ways, tailored to the specific comparison element. Normalization usually includes changing the text to all uppercase or lowercase, eliminating most punctuation, and eliminating some common words (such as 'a', 'an', and 'the'). Each comparison element has its own rules for normalization and comparison. The comparison rules for some elements can become quite complex and, for most of them, several different comparisons are attempted.

Comparison points are described below. Please note that this is a high level summary of how each element is compared and contains only the most general rules. For any record, several dozen comparisons may be attempted.

- **LCCN**  
Uses 010 \$a; for serials, also uses 010 \$z  
If both the input and WorldCat record are LC records, the numbers must match exactly, including both the prefix and the numeric portion.
- **Title**  
Uses 245 \$a, \$b, \$n, \$p, \$f, \$k; also tags 222 \$a and 246 \$a. All of the titles are compared against each other. Single character typographical differences are considered a match. Some words and phrases, such as 'a novel', may be ignored and not affect matching. When both records have generic titles (like "Journal"), the author data is also compared.
- **Publisher**  
Uses 533 \$c, 260 \$b, 261 \$a or \$b or \$e or 262 \$b. Only the first of these found is used for matching. Abbreviations are ignored in the comparison. Equivalency tables have been built for comparing some publishers (example: "Charles Scribner" will match "Scribner"). Single character typographical differences are considered a match. Many words (such as "company", "press", "editorial") are ignored in the comparison. Publisher statements such as "S.n." or "n.p." are treated as though the publisher statement was not in the record.
- **Place of publication**  
Uses 533 \$b, 260 \$a, 261 \$f, or 262 \$a. Only the first of these found is used for matching. Single character typographical differences are considered a match. If the software does not find a match on the textual data, the Country of Publication code in the fixed field (008/15-17) is compared. Two records published in the same country are considered a match.

- **Edition**  
Uses 250 \$a.
- **Author**  
Uses 100 \$a, 110 \$a and \$b, 111 \$a, 130 \$a and \$p, 700 \$a, 710 \$a and \$b, 711 \$a, 720 \$a, and 730 \$a and \$p. Authors are compared only under a few specific conditions, such as when two records have a generic title like “Journal”. In such a case, the author is necessary to distinguish one record from another. The author is also compared when the record has very little other data for matching.
- **Extent of Item**  
Not compared for serials. Uses 300 \$a. Single part items and multi-part items do not match each other. The largest number of pages, volumes, etc., are compared.
- **Size**  
Not compared for serials. Uses 300 \$c and 305 \$c. Only the numeric portion is compared.
- **Type of score designation**  
Uses 300 \$a, 250 43a, 254 \$a, 130 \$s, 240 \$s, and 6XX \$x and \$v to determine the kind of score (e.g., vocal score, miniature score, etc.) If one record has a specialized kind of designation, the other record must also have the same designation. If only ‘score’ is found in both records, they are considered a match.
- **Cartographic scale**  
Compared only for maps. Uses 255 \$a and 507 \$a. Only the numeric data is compared.
- **Music parts**  
Compared only for scores. Uses 300 \$a and \$e and compares only the number of parts. If one record has parts and the other does not, the records are considered not to match.
- **Music publisher number**  
Compared only for scores and sound recordings. Uses 028 \$a.

## **VII. Multiple Matches**

While the goal of record matching is to find a single match or no match at all, sometimes the matching software finds more than one matching record in WorldCat. Records are removed from the set of matching records to reduce the set to fewer than 5 matches. The criteria for removing records are designed to retain only the most likely matches. For example, records with more data to compare are preferred over those with less data. Even after removing some matching records, it is possible that multiple matches will remain. The software attempts to find the “best” record among the remaining matching records. The “best” record is defined by a combination of the source of record, the Encoding Level, and the presence or absence of codes in field 042 (Authentication Code). Records are labeled as one of fifteen ranks, from full Library of Congress records, to minimal-level vendor records. The WorldCat record with the highest rank is chosen as the matching record.

## **VIII. Matching Institution Records**

Matching institution records is done in two phases. The first phase is to find the WorldCat master record as described above. The second phase is to match the institution record to existing institution records. Control numbers in the incoming institution record are compared to control numbers in institution records already in WorldCat using the following keys:

- a. OCLC control number assigned to the institution record in WorldCat
- b. Local system number in the incoming record
- c. RLG control number for institution records containing one