

Appendix A: Matching Records for Batchload

Last revised September 2010

A.1	Introduction.....	2
A.2	Steps in matching.....	2
A.3	Finding matching candidates	2
A.4	Qualifiers	3
A.5	Comparison elements for unique key matching	4
A.6	Comparison elements for extended matching.....	5
A.7	Multiple matches	7
A.8	Matching institution records	7

Caution: This appendix is an overview; it does not include all of the processing details, anomalies, and exceptions that may occur during record matching. Any machine-matching algorithm, regardless of how sophisticated it is, may not always match records in the same way library staff do. Insignificant differences in records may cause machine algorithms to incorrectly match or not match records.

A.1 Introduction

When bibliographic records you send OCLC for processing (also called “incoming” records) are matched against WorldCat for the purpose of adding or deleting holdings symbols or for adding, replacing or deleting institution records or for enriching WorldCat, records are processed through a variety of matching algorithms. The options for matching 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 may review files you send for processing to 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, there is usually no effect on the matching; records are counted neither as matched nor unmatched. There are some exceptions, however, in which the absence or presence of data causes records not to match.

A.2 Steps in matching

There are two basic steps for matching bibliographic records:

- Looking for candidate record matches with WorldCat records
- Determining whether or not the records match, that is, represent the same manifestation of an item, by comparing data in the records

Several iterations of searching and matching may be attempted to find matching records.

A.3 Finding matching candidates

OCLC uses several kinds of searches to look for matching candidates. Each batchload project may be defined differently to use the most appropriate searches. OCLC database specialists may 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 A.4) may be used to evaluate candidate records retrieved by a search. The options are designed to discriminate between records and eliminate unsuitable candidates. Database specialists can select useful and effective options for each project.

- **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.
- **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 B.5.

- Extended matching.** If a single match has not been found during the unique key matching phase, the database specialist can request extended matching to compare other bibliographic data beyond unique keys. In extended matching, up to five searches are executed using various combinations of terms. Queries combine terms from the following MARC tags in various ways. Records found are added to those found in the unique key matching phase as candidates for matching to the incoming record.

The following table lists data elements and MARC fields/subfields that may be used in searches for extended matching.

Data element	MARC fields/subfields
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 (search label dt=) used for WorldCat searching See Searching WorldCat Indexes for details.

A.4 Qualifiers

Qualifiers are optional specialized comparison elements. Depending on how you order your batchload project, you may initially choose qualifiers, or a database specialist may determine qualifiers to use during a review of your project. Use of qualifiers depends on the kind of data and accuracy of the data in the incoming records and on the purpose for processing the file. Qualifiers may be selected or deselected independently and separately for any matching phase of processing, including OCLC control number matching, unique key matching, or extended matching.

If any of the selected qualifiers do not match, the candidate record is considered unmatched.

Qualifiers are:

- **Language of cataloging.** Matches are determined by matching codes in 040 \$b. If no code is present, the record defaults to **eng** (English).
- **Date of publication.** Attempts many comparisons using dates from the fixed field (008/07-10 and 008/11-14) and from fields 260 \$c, 533 \$d, and 362 \$a.
- **Derived title key.** The matching software derives title keys from fields 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. See [Searching WorldCat Indexes](#) for more about derived title searching.
- **Material Type.** Terms are generated as for the material type (mt=) index. Only terms used to describe the physical manifestation of an item (for example, VHS videotape, microfiche, Braille) are used for comparing the records. General terms (such as juvenile, government publication) are not used. If any of the designated terms match, material type is a match. See [Searching WorldCat Indexes](#) for more about the material type index.

A.5 Comparison elements for unique key matching

The following is a list of “unique” keys used for matching. If any of them match a WorldCat record, the incoming record is considered a match.

Unique key	MARC field/subfield
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

Note: In a few instances, the Other system control number (field 029, \$a) is also used as a unique key.

If selected, qualifiers are used to deselect records from the set of records retrieved via unique key matching. When a single record matches based on one of these unique keys **and** the qualifiers match, matching is complete. If no matches are found or if there are multiple matches, the candidate records may undergo extended matching. If there are too many candidates, no attempt is made to match the record.

A.6 Comparison elements for 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 of an item. 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 of the item. In the same way, the matching software also rejects candidate records that have different publishers.

Comparing candidate records

Once candidates are identified, the records are compared in WorldCat. While there are some specialized rules for various kinds of materials (such as maps and sound recordings), most of the comparison elements are treated the same for all types of material. In addition, database specialists can also apply qualifiers (see section A.4).

Normalizations and comparisons

Each comparison element is normalized and compared in a variety of ways, tailored to the specific comparison element. Normalization (that is, treating two similar elements as the same) 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. Because the comparison rules for some elements can be complex, several different comparisons are attempted.

The following general descriptions of comparison points include only the most general high-level rules:

Comparison element	MARC fields/subfields	Notes
LCCN	010 \$a for serials 010 \$z	If both the incoming record and WorldCat record are Library of Congress (LC) records, the numbers must match exactly, including both the prefix and the numeric portion.
Title	245 \$a, \$b, \$n, \$p, \$f, \$k 222 \$a 246 \$a	<ul style="list-style-type: none"> • 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 for matching. • When both records have generic titles such as “Journal,” the author data is also compared.

Comparison element	MARC fields/subfields	Notes
Publisher	533 \$c 260 \$b 261 \$a or \$b or \$e 262 \$b	<ul style="list-style-type: none"> • Only the first field/subfield found is used for matching. • Abbreviations are ignored in the comparison. • Equivalency tables have been built for comparing some publishers (for 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 ignored.
Place of publication	533 \$b 260 \$a 261 \$f 262 \$a	<ul style="list-style-type: none"> • Only the first field/subfield 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	250 \$a	
Author	100 \$a 110 \$a and \$b 111 \$a 130 \$a and \$p 700 \$a 710 \$a and \$b 711 \$a 720 \$a 730 \$a and \$p	<ul style="list-style-type: none"> • Authors are compared under only a few specific conditions, such as when two records have a generic title such as "Journal," in order to distinguish one record from another. • The author is also compared when the record has very little other data for matching.
Extent of item	300 \$a	<ul style="list-style-type: none"> • Not compared for serials. • Single-part items and multipart items do not match each other. • The largest number of, for example, pages or volumes, is compared.
Size	300 \$c 305 \$c	<ul style="list-style-type: none"> • Not compared for serials. • Only the numeric portion is compared.
Type of score designation	300 \$a 250 \$a 254 \$a 130 \$s 240 \$s 6XX \$x and \$v	<ul style="list-style-type: none"> • Determines the kind of score (for example, vocal score, miniature score). • 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	255 \$a 507 \$a.	<ul style="list-style-type: none"> • Compared only for maps. • Only the numeric data is compared.

Comparison element	MARC fields/subfields	Notes
Music parts	300 \$a and \$e	<ul style="list-style-type: none"> • Compared only for scores. • Compares only the number of parts. • If one record has parts and the other does not, the records are considered to be unmatched.
Music publisher number	028 \$a	Compared only for scores and sound recordings.

A.7 Multiple matches

The goal of record matching is to find a single match or no match at all to WorldCat records. Sometimes the matching software finds more than one matching record in WorldCat. The software removes all records from the set of matching records except the five 5 most likely matches. For example, records with more data to compare are preferred over those with less data. The software attempts to find the “best” record among the reduced number of records. The “best” record is defined by a combination of the source of record, the encoding level, and the presence or absence of authentication codes in field 042. Records are ranked at fifteen levels, from full Library of Congress records at the highest level to minimal-level vendor records at the lowest level. The WorldCat record with the highest rank is chosen as the matching record.

A.8 Matching institution records

Institution records are matched in two phases. The first phase is to find the WorldCat master record as described in sections 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:

- OCLC control number assigned to the institution record in WorldCat
- Local system number in the incoming record
- Research Library Group (RLG) control number for institution records (if present)