Backstage LIBRARY WORKS

DATA DEDUPLICATION PROFILE GUIDE

NOTES

Introduction

OVERVIEW

Our Deduplication service provides a variety of options to reduce the number of duplicate or similar records within your database. You and your staff determine the criteria that makes sense and Backstage makes it happen.

The purpose of this guide is to assist you in filling out the profile. Explanations about key parts are included to keep you as informed as possible about your options.

For more information about any part of this guide or the profile itself, please contact your Project Manager.

7 STEPS

The Deduplication service consists of these 7 Steps:

1. **Record Format:** Format to process & return records
2. **Record Stamp:** Determine whether to include Backstage field
3. **Merge Fields:** Specify which fields to retain into matched records
4. **Basic Match Criteria:** Several different options as to which record to keep
5. **Advanced Match Criteria:** Keep records based on data within specific fields
6. **Hit Criteria:** Establish initial pool of matches
7. **Verify Criteria:** Refine pool of matches to determine best possible match or matches

- *Other output options are available, please speak with your Project Manager for more details*

REPORTS

To eliminate confusion about the processing, our reports enable you to see what happens with your data and address any questions you may have. Fields in our reports are color-coded to designate **Hit**, **Verify**, and Merge criteria, giving side-by-side comparisons to allow for easy reference.

We hope this guide helps you find what you're looking for and informs you along the way.

At your service,

The Backstage Automation Team

Step 1: Record Format

MARC-8 FORMAT

STEP 1

MARC-8 has been the standard format for MARC-21 records since 1968. Nearly every system that can export records in MARC format can do so in MARC-8 format.

There is an inherent limitation built into MARC-21 format such that no record can exceed 99,999 characters. Also, no field can exceed 9,999 characters. If a record exceeds the field or record size limits, there may be truncation or loss of data.

During the deduplication process, Backstage will notify customers in the event that a record cannot be processed due to field or record length.

UTF-8 FORMAT

STEP 1

UTF-8 has been in use since early 1993. The main difference between MARC-8 and UTF-8 is that UTF-8 allows for more character types to be used within the records.

Since UTF-8 can represent many more characters than MARC-8, the files tend to be larger in size. Each character in UTF-8 is between 1 - 4 bytes (whereas MARC-8 is only 1 byte in length).

If your system uses UTF-8, please also let us know whether the characters are in precomposed or decomposed format. Precomposed characters use combined diacritics (e.g., n & ~ are combined to form: ñ). Decomposed format separates the characters.

MARCXML FORMAT

STEP 1

MARCXML was developed by the Library of Congress and is based on the MARC-21 format.

The number one advantage with MARCXML format is that there are no limitations to either the field or record size of the data. While both MARC-8 and UTF-8 are constrained by the field and record limits, MARCXML conveniently circumvents that.

MARCXML is typically in UTF-8 encoding, though it depends on the utility used at the time of export.

The default is to send the records back in the same format as how they arrived.

NOTES

NOTES

Step 2: Record Stamp

040 FIELD — STEP 2

Some customers prefer to have a stamp added to records that a vendor has modified in some way. These stamps make it easier to identify which records have already undergone processing and which records may still need to be submitted.

Backstage can add an 040 $d UtOrBLW to all or only merged records.

ALL RECORDS — STEP 2

All will add this field to all records, regardless of whether the record deduped into another record. This option is useful for customers that wish to designate every record that was processed by Backstage

MERGED ONLY RECORDS — STEP 2

Merged will only add this stamp to records where fields were retained from a matched record. Note that this option will only add the stamp to records where some other field was merged into the matching record. So if a record is deduped (where the original record found a matching record), yet no fields were retained during the dedupe, the 040 $d UtOrBLW will not be added in this case.

The default is to add 040 field for merged records only.

Step 3: Merge Fields

MERGE FIELDS INTO MATCHED RECORDS — STEP 3

There may be certain fields your institution would prefer to retain when one record is deduped into another, better match. It could be a field that represents your holdings or items, or a control number, or subject headings, etc. Entire ranges of fields (i.e., 65X) can also be merged in this step.

Once one record is deduped into another record, the original record's fields are lost unless they are retained in the matched record.

If you do choose to merge your original fields into the matched record, we can merge either all of your fields or only the unique fields.

** Please note that you are free to add as many fields as you like to this step.*

ALL FIELDS — STEP 3

If you choose All, then our system will make sure to retain all selected original fields into the matched record. This includes the possibility of merging fields that may duplicate another field already present in the matched record. This option is useful for customers that wish to retain similar subject headings that only differ in indicator values.

UNIQUE FIELDS ONLY — STEP 3

If you choose Unique, then our system will normalize the headings using NACO standards and run a comparison check. Unique indicators are ignored during the comparison check.

The default is to not merge any fields, unless otherwise instructed.

NOTES

NOTES

Step 4: Basic Match Criteria

BASE RECORD

STEP 4

Keeps the first record in a file (or group of files). Any subsequent matching records will be merged into the original record. When a record is read into the deduping process, if it does not match any previously read records, it becomes the base record into which all future matches will be merged.

TAGS RECORD

STEP 4

Compares the original record with the potential match. The best match is determined to be the one with the most fields in the record, which may not necessarily represent the largest record.

LATEST RECORD

STEP 4

Keeps the record with the latest 005 field. If either record lacks an 005 field, it keeps the record with the latest 008 field. If either record lacks an 008 field (and 005 field), it keeps the base record.

LARGEST RECORD

STEP 4

Compares the original record with the potential match. The best match is determined to be the one that has the most data, which may not necessarily represent the one with the most fields.

SMALLEST RECORD

STEP 4

Compares the original record with the potential match. The best match is determined to be the one with the least amount of data, which may also include least number of fields.

INCOMING RECORD

STEP 4

Keeps the last record in the file (or group of files). All matching records will be merged into the last matched record. When a record is read into the deduping process and a matching base record already exists, the base record is merged into this incoming record.

The default is to keep the base record.

Step 5: Advanced Match Criteria

RECORD CONSTRAINT PREFERENCE — STEP 5

This step is optional and takes precedence over Step 4 if filled out. If both the original record and the potential match are equal after this step, then the choice in Step 4 takes precedence.

It is also important to note that the order in which Step 5 is filled out determines the order the system will take the desired match.

The match information in this step can either be a straight string search (e.g., "OCoLC") or it can be a regular expression match (e.g., "[0-9]$"). Both scenarios are case-insensitive.

For example, you may instruct us to include these two constraints:

1. 035 $a "OCoLC"
2. 949 $a "[0-9]$"

Our system would initially check to see if an 035 $a contains "OCoLC". If this is found, then this particular record will be preferred over any others.

If that is not found, our system will then check to see if the record has a 949 $a that *ends with any number.* Again, if found, that record takes precedence in the matching process.

If none of the terms filled out in Step 5 is present in the potential match, the choice in Step 4 takes precedence.

This step is useful for when potential matches contain actual information that indicates it is a good match or a preferred location code.

The default is to determine match based on Step 4, unless Step 5 is filled out.

NOTES

6

NOTES

Step 6: Hit Criteria

HIT CRITERIA — STEP 6

The Hit criteria determines the initial pool of matches during the deduplication process. These are the fields within your records that are most likely to generate a set of possible matches.

It is worth noting that these fields do not need to exist in every single one of your records. Rather, these are fields that, if they do exist, we can apply Verify criteria to further refine the best possible match.

Any records that do exist without any of these fields, however, will not be included as potential matches and will be output as nonmatches after processing. This ensures that brief records that lack sufficient information are either not used as matches or deduped.

Of course, Hit criteria can be set up to be as lax or as strict as your institution deems necessary. Making the criteria too strict will return fewer matches; making it too lax will increase the likelihood of false-positives.

** An example of a false-positive match is one where, although both titles match (hit criteria) and other fields match as well (verify criteria), the main difference between the two records is that one is a large-print and the other is a paper-back. In order to reduce this potential scenario, it may make sense to either eliminate a hit criteria or add more verify criteria.*

The default is to include 010, 020, 022, and 245 fields as Hit Criteria.

NUMBER OF VERIFIES — STEP 6

As part of this step, it is also necessary to inform us of how many Verify criteria should be required for each Hit criteria.

Since some Hit criteria will be more trusted over others (e.g., ISBN vs Title), customers can instruct our system to take a potential match only if *X* number of Verify criteria is also met.

The default is to include all Verify criteria for each Hit criteria.

Step 7: Verify Criteria

VERIFY CRITERIA — STEP 7

The Verify criteria is used to reduce the number of initial matches found using the Hit criteria. This step allows you to further define what constitutes a good match based on the existence of other fields within your records.

Fields should be selected on the basis of the match-rate you expect to see with the deduplication. Selecting fewer fields will result in more matches; selecting more fields will result in fewer, but better matches.

Verify criteria should also be selected in conjunction with Hit criteria. Based on how you answer in Step 6 for Number of Verifies, the criteria you select in Step 7 does not necessarily need to be present in all matches that are found.

** Hit criteria may also include MUST Verify points, which are specifically attached to your hit criteria. These MUST Verify points instruct the system that certain verify criteria is absolutely required in order for the match to be taken. Please talk with your Project Manager for more details.*

*** 008 date variance can also be set by client, where the default is off by 1 year (e.g., "1973" will match on 008 date of "1972"). The system will also attempt to match partial dates if desired (e.g., "197u" will find a match on "1973").*

The default is what the customer chooses for this step.

FULL, PARTIAL, WITHIN — STEP 7

Full compares the full verify string up to the verify length. This is considered the most "safe" match criteria and will return the best matches due to its inherent strictness.

Partial truncates the compare strings to the shortest string, then does a full compare. "Dogs" in one record and "Dog" on other record results in: both truncated to "Dog" and compared.

Within searches each compare string truncated at verify length against the full un-truncated string of the other field. "Cat" will find a match on "The cat in the hat".

** Verify length default is 100 characters, but can be set to any value (less or more). Clients may also specify number of words within verify length to check. For instance, 100 length and 5 words will truncate a search string to 100 characters, and choose the first 5 words to use as the verify point.*

The default is to do a Full compare at Length of 100 characters for each field.

NOTES

Backstage
LIBRARY WORKS

DEDUPLICATION PROFILE

DEDUPLICATION OPTIONS

1 Record Format (*check one*)

□ MARC-8 Records will be processed & returned in MARC-8 format

□ UTF-8 Records will be processed & returned in UTF-8 format

□ MARCXML Records will be processed & returned in MARCXML format

2 Record Stamp (*check one*)

□ Yes Add 040 $d UtOrBLW

□ All records □ Merged records only

□ No Do not add 040 $d UtOrBLW to records

3 Merge Fields (*check all that may apply*)

□ ________ Field □ All fields □ Unique fields only

□ ________ Field □ All fields □ Unique fields only

□ ________ Field □ All fields □ Unique fields only

□ ________ Field □ All fields □ Unique fields only

4 Basic Match Criteria (*check one – this step is required)*

□ BASE First record in a file, into which all future matches will be merged

□ TAGS Record with most number of tags (*not necessarily largest record*)

□ LATEST Record with latest date in 005 (*primary*) or 008 (*secondary*)

□ LARGEST Record with most data (*not necessarily most tags*)

□ SMALLEST Record with least data

□ INCOMING Last record in a file, where all previous matches will be merged

5 Advanced Match Criteria (*check one – this step is optional, takes precedence over Step 4 above*)

Order below also represents order in which Advance Match Criteria is applied

□ ________ Field ________ Sub String / RegExp: ________________________________

□ ________ Field ________ Sub String / RegExp: ________________________________

6 Hit Criteria (*check all that may apply*)

□ 010 a Number of Verifies for this field: ________________________________

□ 020 a Number of Verifies for this field: ________________________________

□ 022 a Number of Verifies for this field: ________________________________

□ 245 abhnp Number of Verifies for this field: ________________________________

□ ______ ______ Number of Verifies for this field: ________________________________

□ ______ ______ Number of Verifies for this field: ________________________________

7 Verify Criteria (*check all that may apply*)

□	LDR	05,06	—	□	Full	:	□	Partial	:	□	Within	- Length:	□ Must
□	008	Date	—	□	Full	:	□	Partial	:	□	Within	- Length:	□ Must
□	010	a	—	□	Full	:	□	Partial	:	□	Within	- Length:	□ Must
□	020	a	—	□	Full	:	□	Partial	:	□	Within	- Length:	□ Must
□	022	a	—	□	Full	:	□	Partial	:	□	Within	- Length:	□ Must
□	1XX	a	—	□	Full	:	□	Partial	:	□	Within	- Length:	□ Must
□	245	abhnp	—	□	Full	:	□	Partial	:	□	Within	- Length:	□ Must
□	250	a	—	□	Full	:	□	Partial	:	□	Within	- Length:	□ Must
□	260	b	—	□	Full	:	□	Partial	:	□	Within	- Length:	□ Must
□	________	________	—	□	Full	:	□	Partial	:	□	Within	- Length:	□ Must

8 Additional Information (*please supply any necessary additional information or comments*)