# General Overview

INN-Reach’s Enhanced Match Point is an expansion of the algorithm used to match incoming contributed records with existing records in the MOBIUS catalog. The previous algorithm sought a match first on local record number and site code, then OCLC number. Enhanced Match Point expands this algorithm to a list of potentially matching fields, including ISBN, ISSN, and Standard Number.

When Enhanced Match Point finds a potentially matching field, INN-Reach performs several validation checks to determine whether the incoming and existing records are a true match. If the records pass Validation, INN-Reach then determines whether the incoming record will be attached as an institutional record to the existing record or whether the incoming record will become the new master record.

If the incoming and existing records fail the validation step, INN-Reach then checks the next field(s) in the hierarchy. If no fields identify a potential match or if the validation step fails for all potential matches, then incoming record is treated as a new master record.

Please note that validation step described in Step 2 is performed after each potential match. If validation fails on a potentially matched field, INN-Reach proceeds to the next step in the matching hierarchy

# Step 1 – Identify a Potential Match

## Match on Record Number and Site Code

INN-Reach compares the local record number and site code[[1]](#footnote-1) of the incoming record with the data stored in the ‘z’ index to determine whether the record has been previously contributed.

If a match is found, INN-Reach compares the OCLC # and Title of the incoming and existing records to confirm they are the same. The Title comparison uses the same normalization algorithm as described in Step 2 - Record Validation. If the OCLC # and Title in both records match, then INN-Reach moves to Step 3 – Determination of the Master Record.

If the OCLC # and Title do not match, INN-Reach splits the incoming record from the record against it was previously matched. INN-Reach then attempts to match on the Primary Match Key.

If a match on local record number and site code is not found, INN-Reach then attempts to match on the Primary Match Key.

## Match on Primary Match Key (OCLC #)

INN-Reach compares the OCLC number of the incoming record with the OCLC Number index, looking for a potential match. If a matching OCLC # is found, INN-Reach proceeds to Step 2 – Record Validation. If validation passes, INN-Reach then moves to Step 3 – Determination of the Master record.

If the validation step fails, INN-Reach then attempts to match on the Secondary Match Keys.

## Match on Secondary Match Keys

INN-Reach compares, in succession, the fields of the incoming bib records against the corresponding indexes. If a match is found, INN-Reach proceeds to Step 2 – Record Validation. If validation passes, INN-Reach then moves to Step 3 – Determination of the Master record.

If a match on a given Secondary Match Point is not found, or if a match is found but the validation step fails, INN-Reach attempts to find a match on the next Secondary Match Key in the hierarchy.

* 020 ISBN subfield a (all occurrences of the subfield in the incoming record)
* 022 ISSN subfield a (all occurrences of the subfield in the incoming record)
* 024 STANDARD subfield a (all occurrences of the subfield in the incoming record)

# Step 2 – Record Validation

INN-Reach performs the following checks when a potential match has been identified on Record Number and Site Code, Primary Match Key, or a Secondary Match Key. If the potential match passes all validation checks, INN-Reach then proceeds to determine which of the incoming or existing records should be the master record. If any of the validation checks fail, INN-Reach continues to try to identify a potential match.

## Imprint Validation

### Step 1: Presence of 260/264 field

INN-Reach examines both incoming and existing records to determine whether both records include a 260 or 264 field. If both records do not contain a 260/264 field, **Imprint Validation FAILS**, and therefore, **Validation FAILS**.

If one record contains a 260/264 field and the other does not, **Imprint Validation PASSES**. INN-Reach then proceeds to Title Validation.

If both records contain 260/264 fields, INN-Reach proceeds to the next step.

Note: INN-Reach gives precedence to the 260 field over the 264 field. If both a 260 and a 264 appear in a given record, INN-Reach uses the 260 field. If no 260 field is present, INN-Reach uses the 264 field.

### Step 2: Compare 260/264 $c (Date of Publication)

1. INN-Reach examines the Leader of both records.
	1. If the Leader byte 7 value is ‘s’ (serial), the system skips to Step 3: Compare 260/264 $a (Place of publication)
	2. If the Leader byte 7 value is something other than ‘s’, the system proceeds to the next step of this comparison
2. INN-Reach determines whether the 260/264 subfield c is present in both the incoming and existing records
	1. If the 260/264 subfield c is absent in one or both records, the system skips to Step 3: Compare 260/264 $a (Place of publication)
	2. If 260/264 subfield c is present in both records, the system proceeds to the next step of this comparison
3. INN-Reach normalizes the data in the first instance of the 260/264 $c in the incoming and existing records, as described below. Normalization may result in an empty string.
	1. If the data from one or both records normalizes to an empty string, the system skips to Step 3: Compare 260/264 $a (Place of publication)
	2. If the data from both records normalize to non-empty strings, the system continues to the next step of this comparison
4. INN-Reach compares the normalized data strings
	1. If the strings match, the system proceeds to Step 3: Compare 260/264 $a (Place of publication)
	2. If the strings do not match, **Imprint Validation FAILS (Validation Step FAILS)**.

#### Normalization Rules for 260/264 $c (Date of Publication)

1. Converts all characters to lower case
2. Strips punctuation
3. Strips leading English articles
4. Contract multiple “space” characters to single “space” characters
5. Strips any instance of the character ‘c’ that is followed by a digit (“c1960” becomes “1960”; “|cc1960” becomes “|c1960”)
6. Strips data within square brackets (and the brackets, themselves) (“|c[1964], 1960” becomes “|c 1960”
7. Extracts the first contiguous sequence of numbers in year format
	* Numeric sequence must contain 4 characters
	* Sequence must begin with a ‘1’ or ‘2’
	* If the sequence begins with ‘1’, the second character must be ‘6’, ‘7’, ‘8’, or ‘9’
	* If the sequence begins with ‘2’, the second character must be ‘0’
8. If the system cannot identify a sequence in year format, the string normalizes as an empty string

### Step 3: Compare 260/264 $a (Place of Publication)

1. INN-Reach determines whether the incoming and existing records contain a 260/264 subfield a.
	1. If the subfield is present on both records, the system continues to the next step in this comparison
	2. If one or both records is missing a 260/264 subfield a, the absence is treated as a match and **Imprint Validation** **PASSES**. INN-Reach then proceeds to Title Validation.
2. INN-Reach normalizes the data in the first instance of the 260/264 $a in the incoming and existing records, as described below. Normalization may result in an empty string.
	1. If the data from both records normalize to a non-empty string, the system continues to the next step in this comparison
	2. If the data from one or both records normalizes to an empty string, the string is treated as a match and **Imprint Validation** **PASSES**. INN-Reach then proceeds to Title Validation.
3. INN-Reach compares the strings
	1. If the stings match, **Imprint Validation** **passes**. INN-Reach then proceeds to Title Validation.
	2. If the strings do not match, the system continues to Step 4: Compare 264 $b (Name of Publisher)

#### Normalization Rules for 260/264 $a (Place of Publication) and $b (Name of Publisher)

1. Converts all characters to lower case
2. Strips punctuation
3. Strips leading English articles
4. Converts multiple “space” characters to sing “space” characters
5. Strips data within square brackets (as well as the brackets themselves). If multiple subfields are contained within a single set of brackets, the system strips the data as though there were brackets around the individual subfields
6. Extracts and concatenates the first continuous sequence of 4 non-space characters
7. Strips the entire string if it normalizes to “sl” or “sn”

Examples

|  |  |
| --- | --- |
| Original Subfield Data | Normalized Subfield Data |
| |aMaplewood, N.J. | mapl |
| |a[Maplewood, N.J.] New York | newy |
| |a[Maplewood, N.J.] | *empty string* |
| |a[Maplewood, N.J.|aNew York, N.Y.|bHarper Collins] | *empty string* |
| |asn | *empty string* |

### Step 4: Compare 260/264 $b (Name of Publisher)

1. INN-Reach determines whether the incoming and existing records contain a 260/264 subfield b.
	1. If the subfield is present on both records, the system continues to the next step in this comparison
	2. If one or both records is missing a 260/264 subfield b, the absence is treated as a match and **Imprint Validation** **PASSES**. INN-Reach then proceeds to Title Validation.
2. INN-Reach normalizes the data in the first instance of the 260/264 $b in the incoming and existing records, as described above. Normalization may result in an empty string.
	1. If the data from both records normalize to a non-empty string, the system continues to the next step in this comparison.
	2. If the data from one or both records normalizes to an empty string, the string is treated as a match and **Imprint Validation** **PASSES**. INN-Reach then proceeds to Title Validation.
3. INN-Reach compares the strings
	1. If the stings match, **Imprint Validation** **PASSES**. INN-Reach then proceeds to Title Validation.
	2. If the strings do not match, **Imprint Validation** **fails (Validation Step FAILS)**.

## Title Validation

### Data Used

 245 $a – Title
 245 $b – Remainder of title
 245 $n – Number of part/section of a work
 245 $p – Name of part/section of a work

### Normalization Rules

1. Convert all capitalized characters to lower case
2. Strip punctuation
3. Strip initial articles, based on 245, 2nd indicator
4. Reduce multiple sequential “space” characters to single “space” characters
5. Replace space-slash-space character strings with a single “space” character (ex. “apples / oranges” becomes “apples oranges”)
6. Replace UTF-8 codes with Western ASCII equivalents
7. Strip data within square brackets
8. Extracts “words” from the remaining data, where a “word” is defined as a 4-character string of non-space characters (ex, “cat” becomes “cat”, but “catastrophic” becomes “cata”)
	1. Subfield a: First three “words” in the first instance of the subfield
	2. Subfield b: First three “words” in the first instance of the subfield
	3. Subfield n: All words in the first instance of the subfield
	4. Subfield p: All words in the subfield. If multiple instances of the subfield are present, INN-Reach compares the data in the first instance of the subfield. If a match is found, INN-Reach then compares the data in the second instance of the subfield’

### Comparison Algorithm

1. Compares concatenated data from subfields a and b (“strict title comparison”)
	1. If a match is found, jump to step 3 ($n)
	2. If a match is not found, jump to step 2 ($a)
2. Compares data from subfield a (“lenient title comparison”)
	1. If a match is found, jump to step 3 ($n)
	2. If a match is not found, **Title Validation FAILS (Validation Step FAILS)**.
3. Compares data from subfield n
	1. If a match is found, jump to step 4 ($p)
	2. If a match is not found, **Title Validation FAILS (Validation Step FAILS)**.
4. Compares data from subfield p
	1. If a match is found, **Title Validation PASSES**. INN-Reach then proceeds to Video Media Format Validation.
	2. If a match is not found, **Title Validation FAILS (Validation Step FAILS)**.

## Video Media Format Validation

INN-Reach compares the System Details Note fields (538 $a) in the incoming and matching existing bib records.

1. INN-Reach determines whether the incoming and existing records contain a 538 subfield a.
	1. If the subfield is present on both records, the system continues to the next step in this comparison.
	2. If one or both records is missing a 538 subfield a, the absence is treated as a match and **Video Media Format Validation** **PASSES.** INN-Reach then proceeds to Large Print Indicator Validation.
2. INN-Reach normalizes the data in the first instance of the 538 $a in the incoming and existing records, as described below. Normalization may result in an empty string.
	1. If the data from both records normalize to a non-empty string, the system continues to the next step in this comparison.
	2. If the data from one or both records normalizes to an empty string, the string is treated as a match and **Video Media Format Validation** **PASSES.** INN-Reach then proceeds to Large Print Indicator Validation.
3. INN-Reach determines whether the normalized 538 strings contain text that distinguishes the video format (specifically, “vhs”, “dvd”, or “blu”).
	1. If the video format text is present in the strings for both the incoming and existing records, the system proceeds to the next step in this comparison.
	2. If the video format is absent from the strings for one or both records, the absence is considered a match and **Video Media Format Validation** **PASSES.** INN-Reach then proceeds to Large Print Indicator Validation.
4. INN-Reach compares the normalized 538 strings.
	1. If the strings match, the **Video Media Format Validation** **PASSES.** INN-Reach then proceeds to Large Print Indicator Validation.
	2. If the strings do not match, the **Video Media Format Validation** **FAILS (Validation Step FAILS).**

### Normalization Rules for 538 subfield a

1. Converts all characters to lowercase
2. Replaces all punctuation with “space” characters
3. Replaces multiple “space” characters with single “space” characters
4. Extracts the first three characters of the remaining string

## Large Print Indicator Validation

INN-Reach checks to see whether or not the word “large” is contained in one of the following fields of both the incoming and matching records.

* MARC tag 245, subfield h (Medium)
* MARC tag 250 (Edition statement)
* MARC tag 300 (Physical Description)
1. INN-Reach determines whether the incoming record contains any of the above fields
	1. If the incoming record contains any of the above fields, the system proceeds to the next step in this comparison
	2. If the incoming record does not contain any of the above fields, then **Large Print Indicator Validation PASSES (Validation Step PASSES).** INN-Reach then proceeds to Step 3 – Determination of the Master Record.
2. INN-Reach determines whether the existing record contains any of the above fields
	1. If the existing record contains any of the above fields, the system proceeds to the next step in this comparison
	2. If the existing record does not contain any of the above fields, then **Large Print Indicator Validation PASSES (Validation Step PASSES).** INN-Reach then proceeds to Step 3 – Determination of the Master Record.
3. INN-Reach examines the incoming and existing records to determine whether the word “large” appears in any of the above fields in the records
	1. If the word “large” appears in the fields in both records, then **Large Print Indicator Validation PASSES (Validation Step PASSES).** INN-Reach then proceeds to Step 3 – Determination of the Master Record.
	2. If the word “large” is present in the fields of one record, but not the other, then **Large Print Indicator Validation FAILS (Validation Step FAILS).**
	3. If the word “large” does not appear in the fields of either record, then **Large Print Indicator Validation PASSES (Validation Step PASSES).** INN-Reach then proceeds to Step 3 – Determination of the Master Record.

# Step 3 – Determination of Master Record

If a matching record is found in Step 1, and all of the validation checks pass in Step 2, INN-Reach then makes a determination as to whether the contributing record should replace the current Master Record, or be added to the existing Master Record as an Institutional Record. INN-Reach makes the following comparisons between the contributing record and the current Master Record.

1. Do the records contain 008 (Fixed Length Data) fields?
	1. If both do or both do not, proceed to the next check;
	2. If one does and the other does not, the record that does becomes the new Master Record.
2. Do the records contain 505 (Formatted Contents Note) fields?
	1. If both do or both do not, proceed to the next check;
	2. If one does and the other does not, the record that does becomes the new Master Record.
3. Do the records contain 520 (Summary) fields?
	1. If both do or both do not, proceed to the next check;
	2. If one does and the other does not, the record that does becomes the new Master Record.
4. Do the records contain 655 (Genre/Form) fields?
	1. If both do or both do not, proceed to the next check;
	2. If one does and the other does not, the record that does becomes the new Master Record.
5. Do the records contain 007 (Physical Description Fixed Field) fields?
	1. If both do or both do not, proceed to the next check;
	2. If one does and the other does not, the record that does becomes the new Master Record.
6. Do the records contain 880 (Alternate Graphic Representation) fields?
	1. If both do or both do not, proceed to the next check;
	2. If one does and the other does not, the record that does becomes the new Master Record.
7. What is the Encoding Level (Leader byte 17) of the records?
	1. If both records have the same Encoding Level, proceed to the next check;
	2. If one record has a lower Encoding Level, the record becomes the new Master Record.
8. What is the contributing local system?
	1. If both records are from MERLIN, Saint Louis University, or Washington University, proceed to the next check;
	2. If both records are not from MERLIN, Saint Louis University, or Washington University, proceed to the next check;
	3. If only one record is from MERLIN, Saint Louis University, or Washington University, that record becomes the new Master Record.
9. Which record was contributed first? – The record that was contributed earliest becomes the Master Record
1. For cluster systems, this is the institutional code (e.g., 6arch), rather than the Agency code. In other words, INN-Reach is looking at the local system, not the individual library. [↑](#footnote-ref-1)