Copac Record Match and Deduplication Procedure: Summary August 2017

The following provides a brief summary of the record match and deduplication procedure used to create the Copac database. The description is followed by a workflow model giving a visual overview of the match process.

Overview

Incoming data goes through an initial check to identify potential duplicates. Each incoming record is checked against the existing Copac data, being matched against the individual contributed records, including those already merged to form a consolidated Copac record.

After the initial match, potentially duplicate pairs of records then go through a detailed match process to confirm whether they are duplicates. Incoming records may form match pairs with multiple records, each match pair is tested in turn.

- If a pair of potentially duplicate records *fail* the Match test the new incoming record is added to Copac as a single, unconsolidated, record.
- If a pair of potentially duplicate records *pass* the Match tests the records are merged to form a consolidated record. If an incoming record has multiple match pairs that succeed in passing the match *all* the records will be brought together in a single consolidated record; a record will never appear in more than one consolidation.
- The incoming record may match with an existing Copac record that is itself already part of a larger set of records, so the new record will be merged into that larger consolidation. It is not necessary that each record in a consolidation matches every other record in that consolidation.

To create a consolidated record we use the largest record, to which we add content from the other matched records where appropriate, eg. spelling variations in a title will be retained for indexing only, whilst additional subject terms will be included for both indexing and display. In addition, within the consolidation we retain each of the original records so that a consolidated record can be expanded to view all the records as originally supplied, as exemplified in the following screenshot of a Copac record display (fig 1).

The data deduplication is a fluid process. For example, an incoming record that fails to find a match is added to Copac as a single record, but this may later form the basis of a new consolidation with the addition of further new records for the same document from other contributors. By contrast, a new incoming record may match with multiple existing single or consolidated records, pulling these together into a new consolidation. Similarly, where one record is the key to bringing together several records into a consolidation, if that key record is then deleted the remaining records in the consolidation may no longer match and a consolidation can split into two or more records. So the database evolves over time as new records are supplied and libraries delete or update their records, changing the record matches. This ensures, as far as possible, the ongoing accuracy of the record matches and associated data consolidation.

The match process itself evolves over time as the data changes and new problems emerge. But care is required to ensure that improving our ability to match records with specific problem features doesn't result in mistaken matches and consolidation errors that bring other records together incorrectly.

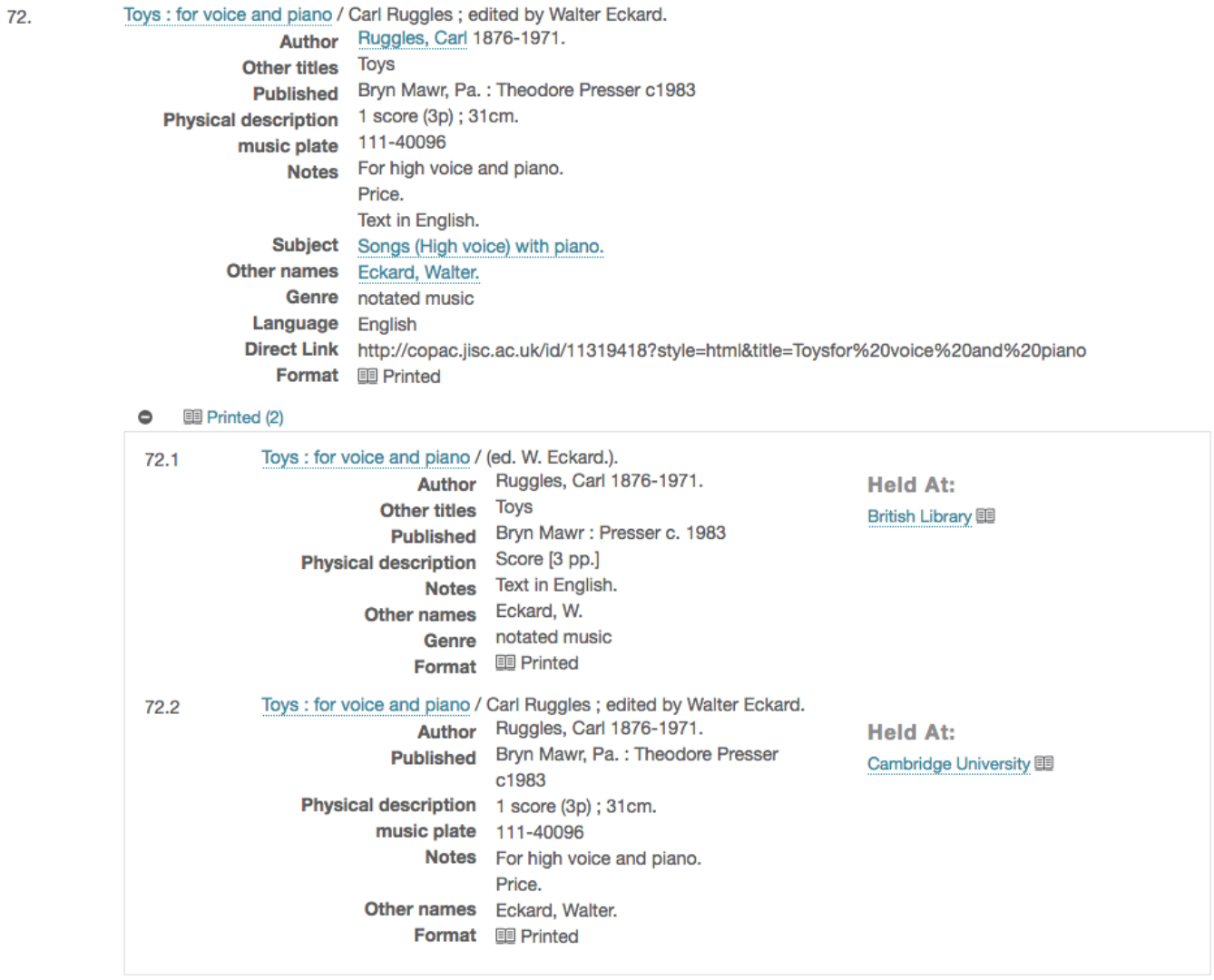


Fig 1. Copac screenshot: a Full record expanded to show the individual contributed records that have been brought together to create the consolidation – record 72 in this result set.

Part 1. Identifying potential duplicates

An incoming record goes through an initial check for potential duplicates by looking at where that incoming record would fit into the existing Title index. This identifies:

- Any exactly matching titles in the existing Copac records,
- Plus a defined number of titles before and after the position at which the incoming title would fit in the Title index. This number will vary with update size and workload, usually between 4-10 titles.

Any existing Copac record identified as a potential duplicate is formed into a pair with the incoming record and goes into the Match process that is used to confirm whether the records are duplicates.

2. Match process

The Match process confirms or rejects the record pairs formed by the potential duplicates identification process [Part 1]. Which route the records take through the Match process depends on an initial standard number match and/or the nature of the material described in the record.

2.1 Standard Number match 1

Record pairs containing Standard Number (SN) elements generally go through a Quick Match. Other records go through the Detailed Match process.

If any of the following Standard Number (SN) match 1 checks are true the record pair goes through the Quick match process, this speeds the matching process and also avoids having to match on some of the less consistent elements such as publisher. Otherwise the record pair goes through the Detailed Match process.

- Two periodical records with at least one matching ISSN
- OR All ISBN's match
- OR All ISMN's match
- OR All ESTC numbers match.

2.2 Quick Match process

For records that have passed SN match 1, the Quick Match checks the record pair for matching title and edition. If this match succeeds the duplicate record pair is confirmed and the records become part of a consolidation. If the match fails the incoming record is added to the database as a single, unconsolidated, record.

2.3 Detailed Match process

Records that fail SN match 1 go through the Detailed Match process. If the record pair fails ANY test the match process ceases. If the record pair passes ALL the match tests they are confirmed as duplicates and the records become part of a consolidation.

A second Standard Number (SN) check, SN match 2, is used to identify the route the record pair takes though the Detailed Match tests. This time the SN match only requires one SN in common between the records. This check assigns a flag to the record pair that either lets it though just the basic match tests, or forces it through the additional tests required where there is no SN in common between the records.

2.3.1 Basic Match tests

1. If both records have an ISSN or ISMN or ISBN or ESTC number, do they have one in common? [SN match2]
2. Are there more than 4 ISBN's? If so do they match?
   Merging records for single volumes of sets with multi-volume records is potentially problematic. But we want to be able to match records where one has, say, ISBN's for paperback and hardback whilst the other has only the paperback ISBN.
3. Do the dates match?
   This is *not* used where both records in a pair are periodicals.
   Uses 008, 260, 264.
4. If both records are *periodicals* do the hierarchical places match?
   Uses 752. This is primarily for matching some newspaper records.
5. Do the titles match?
   This checks 245 title as well as volume for multi-part works. It uses a fuzzy match allowing for minor variation, but preserving single letter 'words'. A smaller subset of subfields are used for matching periodicals. Includes checks for more complex title, edition and statement of responsibility details, including title truncation, in pre-1800 works and older records.
6. Do the editions match?
   Matches word and number variants.
7. Do the series volumes match?
   Uses the 440 if present, or 490.
8. Do the authors match?
   Corporate author stopwords are removed and there is a fuzzy match process that allows for some minor variation. The match uses 1XX and 7XX fields. If the usual author fields are not present it will check the 130, 730, 720, 245.

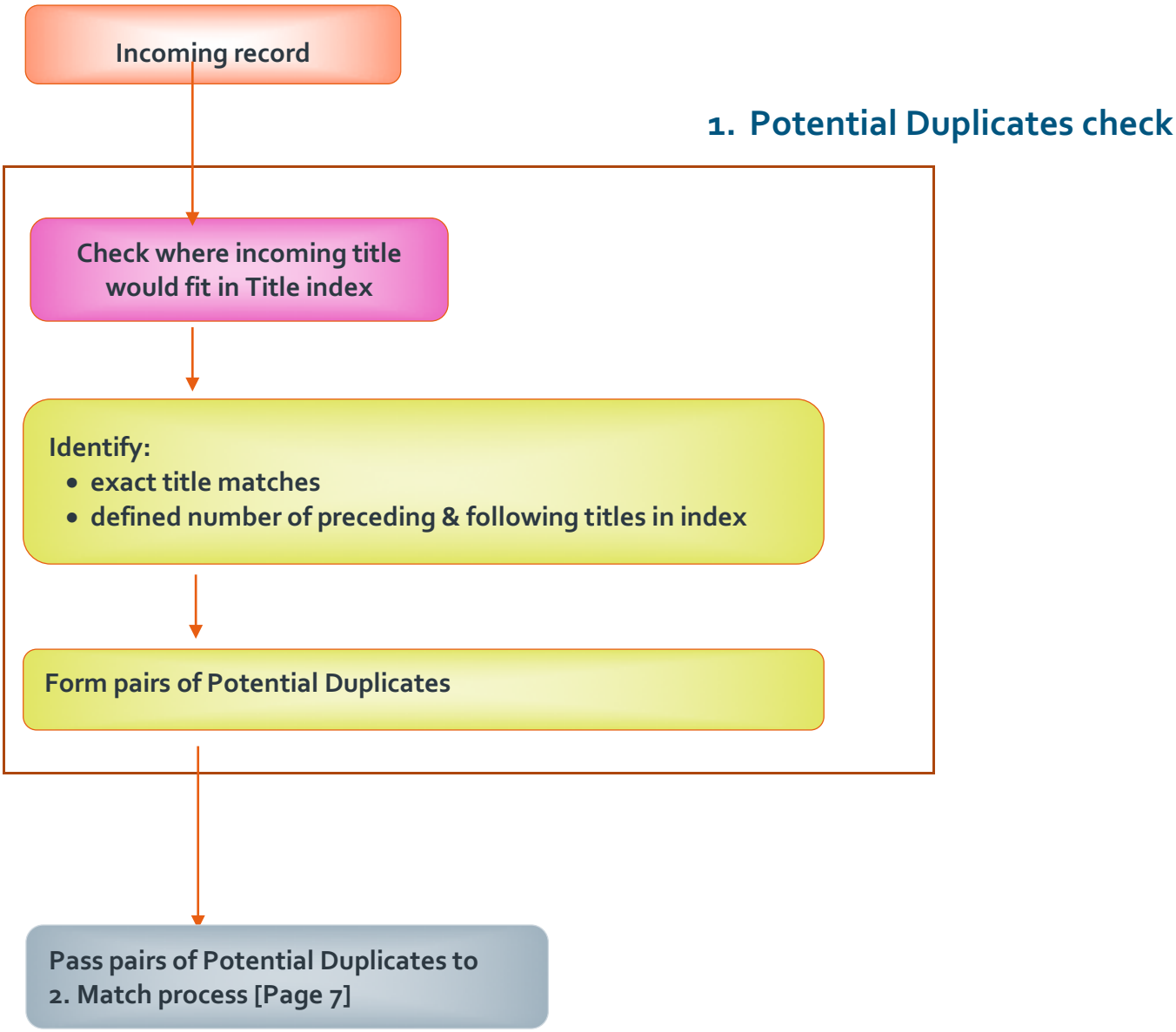
2.3.2 Additional Match tests

If the records failed SN match 2 then the following additional tests are used:

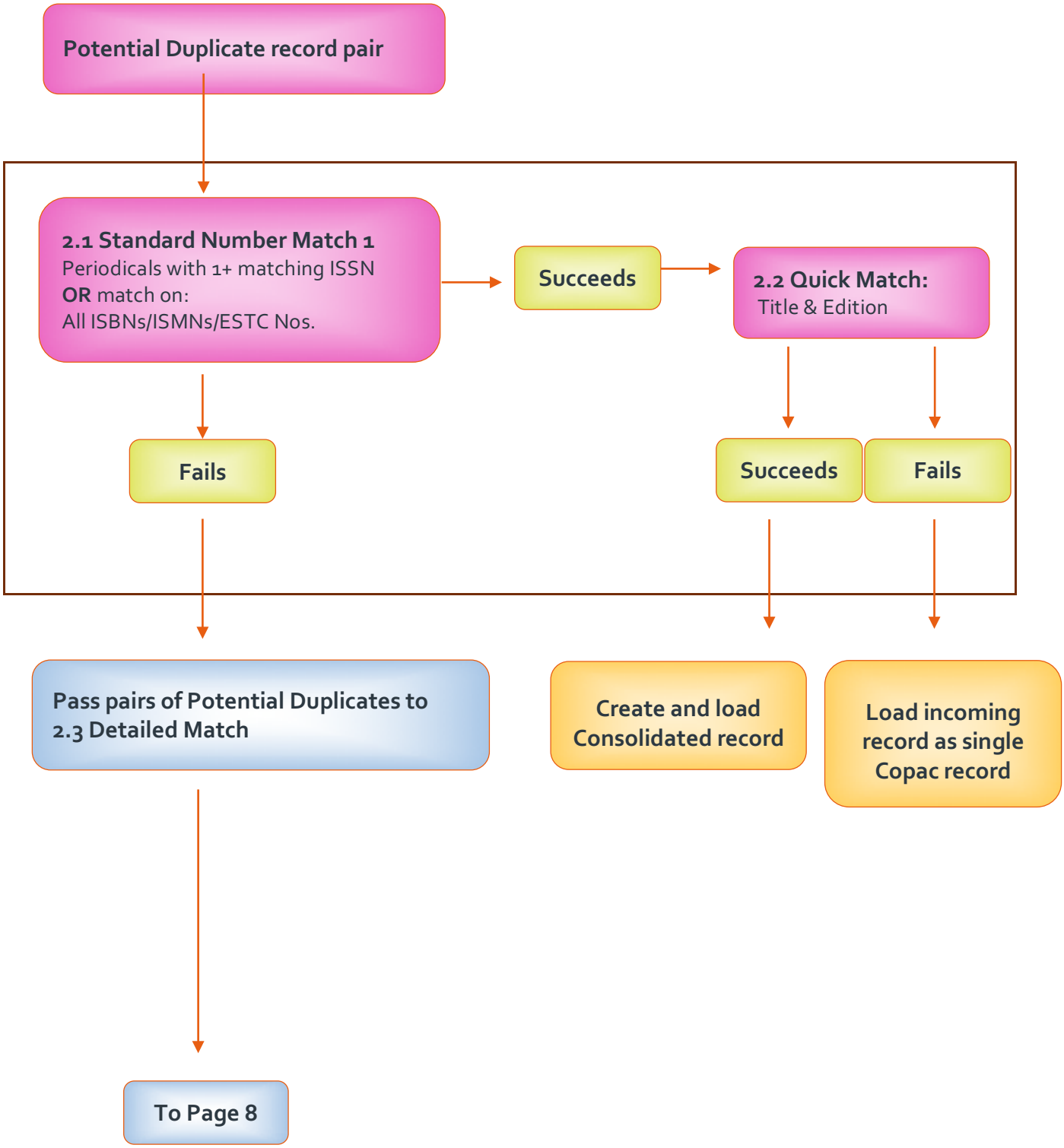
1. Do the pages match?
   Uses the 300. This is *not* used where both records in a pair are periodicals.
2. Do the publisher names match?
   Uses the 264, 260. Common stopwords are excluded and there is a partial match on publisher name and/or location depending on the information available.
3. Do the map scales match?
   Uses the 034.
4. Do the music score types match?
   Uses the 300, 240, 245. It checks for a range of score types eg. choral score.

Match and Deduplication Workflow model

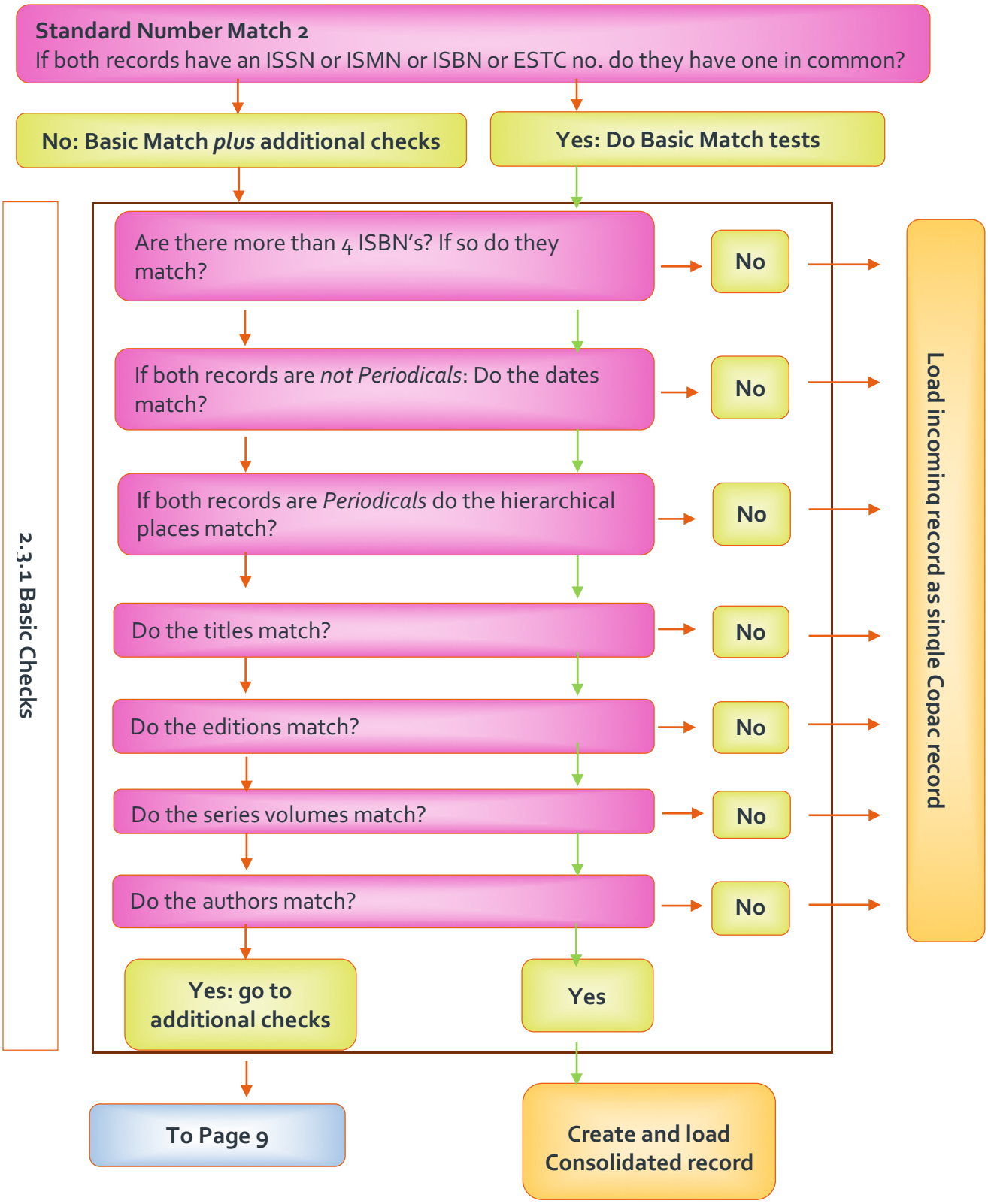
1. Identifying Potential Duplicates



2. Match Process



2.3.1 Detailed Match: Basic Match tests



2.3.2 Detailed Match: Additional Match tests