


Duplicate detection and merge tool

Duplication detection may or may not be in DSpace 7

 Per discussion in our meeting on [Sept 28, 2017](#), this tool should initially be considered a DSpace 7 add-on. While we hope to include this in the final scope of the DSpace 7 Submission process, we have said we will keep our initial DSpace 7 release closer to the features of DSpace 6. As this is a new feature, we need to determine if its implementation will be problematic in the given timelines and/or whether it would simply be released as an optional "beta" (disabled by default) feature.

Nonetheless, we encourage feedback on the designs of this feature. We feel this is a useful DSpace feature for the future, whether it makes it into the DSpace 7 release or a later release.

The functionality will be based on the already existent implementation available in DSpace-CRIS see ([The administrative UI](#), [deduplication alert](#)).

The functionality is largely inspired by the [SOLR official de-duplication approach](#), for each item one or more signatures are computed using pluggable implementation.

A [signature](#) is a value that summarizes the information in the item using a [pluggable transformation](#) (case insensitive, ascii transcription, identifier normalisation, etc), out of box implementation based on a normalization of a single metadata (such as an identifier or the title) or a combination of metadata (such as title + year, etc.) are included.

Two items are flagged as potential matches if they share at least one signature.

Feedback on potential matches (reject or duplicate flag) are stored in the database table dedup_reject

Signatures and matched groups are computed when an item is updated and stored on a dedicated SOLR core this makes **extremely fast and lightweight to check for potential duplicate**. This SOLR core is maintained using DedupEventConsumer a script [DedupClient](#) is provided to rebuild the index or build it the first time if you are migrating from a previous version.

Two functionalities have two points of interaction with the users

- During the submission and the workflow, the potential duplicates are presented and feedback from the submitter and validator are collected
- An administrative dashboard is available to the administrator to check for existent duplicates and merge groups of items

Below the initial wireframes:

DSpace 7 - Deduplication

← → × ↗

Q

Deduplication

All

Reported for Merge

Compare

Title

Submitter check:

0

Administrator check:

1

Total

1

Identifiers

Submitter check:

3

Administrator check:

1

Total





4





[Edit form Deduplication functionalities \(editable mockups\)](#)

The page show the potential duplicate group by signature. You can select the items to compare or reject. The system records the action "No duplicate" so the match is not show in the future.

Item(s) found: 3

☐ No duplicate
 ☐ Compare

☒ Collection "Sample collection"
 Internal identifier: 1





☐ Collection "Sample collection"
 Internal identifier: 2





☒ Collection "Sample collection"
 Internal identifier: 3
