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 it's implementation will be problematic in the given timelines and/or whether it would simply be released as a 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 summarize 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 make **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 functionality have two point 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 group of items

Below the initial wireframes:

⇔ ⇒ × ☆	DSpace 7 - Submission form	
New item	Indexing Journal Info Proceeding FoundRef CC License CC License If your entry is really a duplicate Interrupt it ! You can ask the submitter about the "master" record (if it has not been published) or the system administrator to make changes Discard submission Summary 2014. Sample Article Blue, Earth have it high	is enabled the see the same backwork in the match even confirm that the match even con
		1

Edit form Deduplication functionalities (editable mockups)

	DSpace 7 - Submission form	_
New item		
Drop here the actual content or bibliogree Collection ? Sample collection Another sample collection Yet another Collection Enter the names of the authors of Author/s ?	Your request will be visible to users authorized to edit the item marked as a duplicate. If this is not a duplicate select 'Cancel'.	
Author, 3 Author, 3 Affiliation ? University of 1 Author, 1 X Author, 2 X Title* ?	Send Cancel	
Identifiers ? DOI -	fetch metadata + more Date ? / /	
Indexing Files & access condition	••	
CC Licence		
Discard	Save for later Deposit	

Deduplication All Reported for Merge Compare Title Submitter check: 0 Administrator check: 1 Total 1	
Identifiers Submitter check: 3 Administrator check: 1 Total 4	
	"

DSpace 7 - Deduplication	
	2
Deduplication - Records duplicates for "Title"	
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.	
Signature: 39hd8djfhnchfosjfu4830h493jf	
Item(s) found: 3	
No duplicate Compare	
Collection "Sample collection"	
and are and and and the second and t	
te mar an ane tanke ver second	
Collection "Sample collection"	
et ann an ann ann ann ann ann ann ann ann	
Collection "Sample collection"	
Internal identifier: 3	
aunorane and an anti-the anti-the anti-the	
	11

	DSpace 7 - Deduplication	
◇ ▷ × ☆		
Deduplication - compare for s	signature "39hd8djfhnchfosjfu4830h493jf"	
Collection "Sample collection" Internal identifier: 1 	Collection "Sample collection" Internal identifier:	
Bitstreams		
Show / Hide values Show / Hide final object	Merge	
Authors (DC.CONTRIBUTOR AUTHOR)		•
Surname one, Name one	1 Delete	
Surname two, Name two	1 Delete	
Surname three, Name three	The Delete	
Surname four, Name four	1 Delete	
Date of issue (DC.DATE.ISSUED)		-
Abstract (DC.DESCRIPTION.ABSTRACT)		-
Identifiers (DC.IDENTIFIER.URI)		•
Identifiers (DC.IDENTIFIER.URI)		•
		11