Deduplication

DSpace Items

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 functionalities 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 (see deduplication alert)
- An administrative dashboard is available to the administrator to check for existent duplicates and merge group of items (see The administrative UI)

CRIS Objects

Since DSpace-CRIS 5.10 basic deduplication features have been implemented also for CRIS objects to identify and merge potential duplicates among researcher profiles, projects, organisations, etc.

the detection mechanism for CRIS Objects is the same illustrated above for DSpace items. Out of box is possible to configure which metadata are used to identify duplicate among each object types. Custom signature algorithm can be implemented and activeted via Spring bean in the same exact way than for publications (dspace items)

Manage potential CRIS duplicate

A batch script is provided to manage potential duplicates among CRIS Objects.

```
usage: org.dspace.app.cris.batch.ScriptListAndRejectDedupObjects

 -c,--compare      compare two objects
 -h,--help         help
 -i,--id <arg>     object id
 -n,--note <arg>   reject note
 -r,--reject       reject two objects
 -t,--type <arg>   object type


USAGE:
 List duplicates: -t <object type> [-i <object id>]
 Compare two objects: -c -t <object type> -i <first object id> <second object id>
 Reject two duplicate objects: -r -t <object type> -i <first object id> <second object id> [-n <reject note>]
```

So to list all the groups of potential duplicates for researcher profiles you need to execute

```
./dspace dsrun org.dspace.app.cris.batch.ScriptListAndRejectDedupObjects -t 9
```

using -t 10 you will get the list of potential duplicates among projects and with -t 11 among organisations. It is also possible to list potential duplicates of additional dynamic entities like journals, awards, etc. once the the dynamic object type is known (i.e. 1001, 1002, ...)

Info

Please note that the script show only potential duplicates with status "active" (i.e. CRIS entity MUST be not in withdrawn state)

To flag a potential duplicate as a fake detection you need to run the script specifying the type of the objects (9 for researcher profiles, etc.) and the id of the two objects.

Please note that, contrary to what happen for rejection of duplicate suggestion among dspace items, the rejection is only stored in the deduplication solr core. So if you rebuild the deduplication core using the org.dspace.app.cris.batch.DedupClient script you can potentially loss such information.

The org.dspace.app.cris.batch.DedupClient script has been extended to support the -t parameter as well so to allow reindexing of specific object types

Merge Script

A batch script is provided to merge different instances of a cris object in a single one. The script works on any kind of entity (researcher profiles, organisation units, projects, etc.) with the following rules

- any items linked to the merged cris object will be linked to the target cris object
- any cris objects linked to the merged cris object will be linked to the target cris object
- properties and nested object present only in the merged cris object are copied to the target. The parameters allow fine grain control about which properties copy and override.

```
usage: ScriptMergeCrisObject

 -d,--delete                  delete merged objects, the default (without
                              the -d option) is to disable them
 -h,--help                    help
 -m,--merge <arg>             CRIS ID(s) to merge into the target (use
                              multuple m if needed - merge occurs
                              respecting the order from left to right)
 -p,--replace_notempty <arg>  properties to override in the target with
                              the values from the merged objects IF NOT
                              EMPTY
 -r,--replace <arg>           properties to override in the target with
                              the values from the merged objects
 -s,--skip                    properties to ignore during the merge
 -t,--target <arg>            CRIS ID to retain (merge target)
 -x,--exclude                 Don't merge properties, only move link from
                              the merged object to the target


USAGE:
 ScriptMergeCrisObjects -t <crisID> -m <toMergeCRIS-ID1> m <toMergeCRIS-ID2> .. m <toMergeCRIS-IDn>  [-r propR1 -
r propR2... -r propRN] [-p prop1 -p prop2... -p propN] [-s]
```