EDIT - feature request #7801

AM: Deduplicate references

09/29/2018 12:32 PM - Andreas Müller

Status:	Closed	**Start date:**	
Priority:	Highest	**Due date:**	
Assignee:	Andreas Müller	**% Done:**	100%
Category:	cdmadapter	**Estimated time:**	15:00 hours
Target version:	Euro+Med Migration		
Severity:	normal		

Description

Many references are duplicates. We could try to deduplicate them during import.

TODO:

- check for parameters like annotations and extensions
- RefDetails

Related issues:

Related to EDIT - feature request #7800: Parse preliminary RefDetails	**Closed**
Related to EDIT - feature request #7799: AM: Parse authorteams	**Resolved**

Associated revisions

Revision f6275b1f - 10/14/2018 02:49 PM - Andreas Müller

ref #7801 unify cache initialization in deduplicationHelper

Revision db652f5c - 10/14/2018 02:52 PM - Andreas Müller

ref #7801 and ref #3787 deduplicate reference.authorstring and reference itself

History

#1 - 09/29/2018 12:32 PM - Andreas Müller

- Related to feature request #7800: Parse preliminary RefDetails added

#2 - 09/29/2018 12:32 PM - Andreas Müller

- Related to feature request #7799: AM: Parse authorteams added

#4 - 10/14/2018 02:57 PM - Andreas Müller

Find deduplicated records in CDM

```
SELECT osb.id, ref.id, osb.idInSource, osb.idNamespace, ref.refType, ref.titleCache, ref.authorship_id, ref.
abbrevTitleCache, ref.protectedTitleCache, 
ref.protectedAbbrevTitleCache, ref.abbrevTitle, ref.title, ref.volume, ref.*
FROM Reference ref LEFT OUTER JOIN Reference_OriginalSourceBase MN ON MN.Reference_id = ref.id 
LEFT OUTER JOIN OriginalSourceBase osb ON osb.id = MN.sources_id
 INNER JOIN (SELECT Reference_id, count(*) as n FROM Reference_OriginalSourceBase MN INNER JOIN 
OriginalSourceBase osb ON osb.id = MN.sources_id WHERE idNamespace <> 'import to Berlin Model' GROUP BY MN.
Reference_id HAVING n > 1) as drvTab2 ON drvTab2.Reference_id = ref.id
WHERE  (1 = 1) 
-- AND idNamespace <> 'RefDetail'
-- AND titleCache like '%unde%'
-- AND ab.protectedTitleCache = false
-- AND ab.id NOT IN (SELECT Team_id FROM AgentBase_AgentBase MM WHERE MM.Team_id IS NOT NULL) -- AND idInSOurc
e = '1'
 AND idInSource like '7712094'
-- AND ref.id IN (SELECT Reference_id FROM (SELECT Reference_id, count(*) as n FROM Reference_OriginalSourceBa
se MN GROUP BY MN.Reference_id HAVING n > 1) as drvTab)

/* AND titleCache IN (SELECT titleCache FROM (
SELECT r2.titleCache, r2.abbrevTitleCache, r2.authorship_id, count(*) n
FROM Reference r2
GROUP BY r2.titleCache, r2.abbrevTitleCache, r2.authorship_id
HAVING n > 1) as drv )
*/

ORDER BY ref.titleCache, ref.id, length(idInSource), idInSource, ref.refType
```

#5 - 10/14/2018 02:59 PM - Andreas Müller

- Description updated

#6 - 10/15/2018 03:44 PM - Andreas Müller

- Status changed from New to In Progress

- Priority changed from Priority14 to Highest

#7 - 09/17/2021 02:58 PM - Andreas Müller

- Status changed from In Progress to Closed

- % Done changed from 0 to 100

The import did run long time ago. Deduplication was done as far as it was done. As we can't change much here other then trying to further find duplicates and deduplicate them I think we can close this ticket.