Edit - task #7798

[AM, ERS]: Import (freetext) nameInSource correctly

09/29/2018 11:40 AM - Andreas Müller

Status:	Resolved	**Start date:**	09/29/2018
Priority:	Highest	**Due date:**	
Assignee:	Andreas Müller	**% Done:**	100%
Category:	cdmadapter	**Estimated time:**	1.00 hour
Target version:	Euro+Med Migration		
Severity:	normal		

Description

For occurrence sources and maybe also for common names and facts.

Occurrences:

1. case: freetext exists parallel to link => if nameCache or fullnameCache are exactly equal then neglect freetext, otherwise store it parallel to name link; a list has been sent to ERS to clean up these cases (see #7798#note-4)

2. freetext is only name

   a. name can not be found for any Name.nameCache => we could create new TaxonName for this, but will not as it polutes the DB, we better try to clean up theses cases over time (see #7798#note-6)
   b. >1 name is mentioned, separated by '/'. We better clone these sources (now or later)
   c. a name exists in DB, we try to find the best matching name by first searching in the synonymy and, if not exists use the single matching name, if >1 matching names exists, log the case and use the first matching name (prefer name without author)

Most important for import: try to reduce the number of names that only come via occurrence import. These are

- "Orphaned name: A similar name" : names that do not belong to any taxon but a matching name was found in "synonymy" * Decide for invalid designations (why do they show up?) and orthographic variants
- "Orphaned name: A similar misapplied name" : like above but for misapplications. * Decide on all
- TBC

open issues:

- test code
- test for facts and common names

Related issues:

Related to Edit - bug #7406: [CHECK] Check if all original source names of oc...	**Resolved**	**05/09/2018**

Associated revisions

Revision c4591f7a - 09/29/2018 06:17 PM - Andreas Müller

ref #7798 import freetext nameInSource for occurrences best way and log all problems

Revision 14295b4b - 10/10/2018 12:51 PM - Andreas Müller

ref #7798 implement and improve occurrence.nameUsedInSource import for freetext and referenced names

History

#2 - 09/29/2018 11:42 AM - Andreas Müller

- Target version changed from Unassigned CDM tickets to Euro+Med Migration

#3 - 09/29/2018 11:57 AM - Andreas Müller

- Subject changed from Import freetext nameInSource correctly to AM: Import freetext nameInSource correctly

#4 - 09/29/2018 03:39 PM - Andreas Müller

im Berlin Model gibt es ja 2 Felder in die man den Namen, der in der Quelle für eine Verbreitungsangabe verwendet wurde, eintragen. Einmal als Freitext, das andere Mal als Link zu einem existierenden Namen.

In einigen Fällen sind beide Felder gefüllt, weichen aber (leicht) voneinander ab. Siehe Anhang.

Kannst du die bereinigen bzw. eine Regel angeben, wie sie behandelt werden sollen?

```
 SELECT
 os.OccurrenceSourceId, pt.PTRefFk,  n.fullNameCache, os.OldName, os.OldNameFk,oldN.nameCache, oldN.FullNameCa
che, n.nameCache, ar.Unit area, pt.RIdentifier, sumcat.emOccurSumCatId, sumcat.Short
FROM emOccurrence occ
 INNER JOIN PTaxon pt ON occ.PTNameFk = pt.PTNameFk AND occ.PTRefFk = pt.PTRefFk
 INNER JOIN Name n ON n.NameId = pt.PTNameFk
 INNER JOIN emOccurrenceSource os ON os.OccurrenceFk = occ.OccurrenceId
 LEFT OUTER JOIN Name oldN ON oldN.NameId = os.OldNameFk
  INNER JOIN emArea ar ON occ.AreaFk = ar.AreaId
 LEFT OUTER JOIN emOccurSumCat sumcat ON occ.SummaryStatus = sumcat.emOccurSumCatId
WHERE occ.occurrenceId IN ( SELECT occurrenceId FROM v_cdm_exp_occurrenceAll )
AND ( OldName IS NOT NULL AND oldNameFk IS NULL OR OldName IS NOT NULL AND oldN.NameCache <> OldName )
AND oldN.nameCache IS NOT NULL
ORDER BY pt.PTRefFk, oldN.nameCache
```

#5 - 09/29/2018 04:34 PM - Andreas Müller

- Description updated

#6 - 09/29/2018 04:35 PM - Andreas Müller

In einer nicht ganz kleinen Anzahl Fälle, für die der Name NUR als Freitext vorliegt, nicht als Link, wurde ein entsprechender Name in der Datenbank nicht gefunden.

Dies kann ein Hinweis darauf sein, dass eine abweichende Rechtschreibung vorliegt oder dass er in der Synonymie nicht vorkommt.

```
 SELECT
 os.OccurrenceSourceId, pt.PTRefFk,  n.fullNameCache, os.OldName, os.OldNameFk,oldN.nameCache, oldN.FullNameCa
che, n.nameCache, ar.Unit area, pt.RIdentifier, sumcat.emOccurSumCatId, sumcat.Short
FROM emOccurrence occ
 INNER JOIN PTaxon pt ON occ.PTNameFk = pt.PTNameFk AND occ.PTRefFk = pt.PTRefFk
 INNER JOIN Name n ON n.NameId = pt.PTNameFk
 INNER JOIN emOccurrenceSource os ON os.OccurrenceFk = occ.OccurrenceId
 LEFT OUTER JOIN Name oldN ON oldN.NameId = os.OldNameFk
  INNER JOIN emArea ar ON occ.AreaFk = ar.AreaId
 LEFT OUTER JOIN emOccurSumCat sumcat ON occ.SummaryStatus = sumcat.emOccurSumCatId
WHERE occ.occurrenceId IN ( SELECT occurrenceId FROM v_cdm_exp_occurrenceAll )
AND ( OldName IS NOT NULL AND oldNameFk IS NULL)
AND OldName not like '%/%'
AND OldName NOT IN (SELECT DISTINCT NameCache FROM Name n1 WHERE nameCache IS NOT NULL)
AND occ2.SummaryStatus IS NOT NULL)
 ORDER BY pt.PTRefFk, oldN.nameCache, os.OccurrenceSourceId
```

#7 - 09/29/2018 04:41 PM - Andreas Müller

- Description updated

- % Done changed from 0 to 20

#8 - 09/29/2018 04:42 PM - Andreas Müller

- Description updated

#9 - 09/29/2018 04:43 PM - Andreas Müller

- Description updated

#10 - 09/29/2018 04:43 PM - Andreas Müller

- Description updated

#11 - 09/29/2018 04:44 PM - Andreas Müller

- Description updated

#12 - 09/29/2018 06:14 PM - Andreas Müller

- Subject changed from AM: Import freetext nameInSource correctly to [CHECK]: Import freetext nameInSource correctly

- Status changed from New to In Progress

#13 - 09/29/2018 06:21 PM - Andreas Müller

- Description updated

- % Done changed from 20 to 40

#14 - 09/29/2018 06:21 PM - Andreas Müller

- Estimated time changed from 2.00 h to 1.00 h

#16 - 09/29/2018 06:27 PM - Andreas Müller

- Related to bug #7406: [CHECK] Check if all original source names of occurrences are correctly imported added

#17 - 10/02/2018 05:45 PM - Andreas Müller

- Subject changed from [CHECK]: Import freetext nameInSource correctly to [CHECK]: Import (freetext) nameInSource correctly

#18 - 10/02/2018 06:05 PM - Andreas Müller

- Subject changed from [CHECK]: Import (freetext) nameInSource correctly to [AM, ERS]: Import (freetext) nameInSource correctly

- Description updated

All names imported only by occurrences: (n=1435)

```
SELECT   n.nameId, n.nameCache, n.FullNameCache, n.uuid, count(*) as n
FROM            dbo.v_cdm_exp_namesOccurrenceSource v LEFT OUTER JOIN Name n ON n.NameID = v.NameId
WHERE v.NameId NOT IN
(SELECT        NameId
FROM            Name
WHERE (NameId IN (SELECT PTNameFk AS NameId FROM dbo.v_cdm_exp_taxaAll)) OR
        (NameId IN
            (SELECT        NameId
            FROM            dbo.v_cdm_exp_namesRelatedTo)) OR
        (NameId IN
            (SELECT        NameId
            FROM            dbo.v_cdm_exp_namesRelatedFrom)) OR
        (NameId IN
            (SELECT        NameId
            FROM            dbo.v_cdm_exp_namesCommonNameSource)) OR
        (NameId IN (7502960, 7709217, 7502034, 28349))
)
```

```
GROUP BY n.nameId, n.nameCache, n.FullNameCache, n.uuid
ORDER BY n DESC
```

#19 - 03/27/2019 05:58 PM - Andreas Müller

ERS decided to use synonymie names whereever possible. Equalness of existence of authors is not so important.
With this about 900 names remained as names not existing in current import but linked by occSources. These we explicitly added to the import via id in a separate query.

#20 - 03/27/2019 05:59 PM - Andreas Müller

- % Done changed from 40 to 90

Eckhard noch fragen, was mit den Namen geschehen soll, bei denen LinkedName und FreeTextName nicht übereinstimmen.
Dann kann das Ticket wohl geschlossen werden.

#21 - 04/02/2019 01:41 PM - Andreas Müller

- Status changed from In Progress to Resolved

- % Done changed from 90 to 100

Andreas Müller wrote:

> Eckhard noch fragen, was mit den Namen geschehen soll, bei denen LinkedName und FreeTextName nicht übereinstimmen.
> Dann kann das Ticket wohl geschlossen werden.

Entscheidung: werden im CDM bearbeitet, grundsätzlich sollen im die wirklichen Originalschreibweisen gespeichert werden, daher müssen diese Fälle manuell aufgeräumt werden, was im CDM leichter ist (im BM wird in diesen Fällen nur die Freitext-Variante angezeigt).

Put this to resolved but do not close yet to remind to cleanup after import. Mail in post-migration folder exists. Also remember to decide on pure freetext names. Should they be created as names or should they be kept as freetext. Mail also exists in post-migration.