EDIT - task #8792

Issues to have in mind for ERMS import

12/26/2019 12:06 PM - Andreas Müller

Status:	New	**Start date:**	
Priority:	New	**Due date:**	
Assignee:	Andreas Müller	**% Done:**	0%
Category:	cdmadapter	**Estimated time:**	0:00 hour
Target version:	PESI 2022		
Severity:	normal		

Description

The functioning of the ERMS import (and PESI export) is explained at: Erms_mapping

General information

Old databases are available at BGBM-PESISQL\SQLExpress.

Many issues have been handled in #1444. However as the ERMS-CDM is complex in often difficult there are many issues to take care for during upcoming imports.

Comparison with SQL see #8792#note-30 below

Data check before import

- Check that the rule still is true that synonyms are directly linked to their accepted taxon and not via recursion (at least for field tu_accfinal (which has only less NULL values then tu_accfinal but never differs from tu_acctaxon in value)

```
SELECT  *
FROM [tu] INNER JOIN tu acc ON tu.tu_acctaxon = acc.id
WHERE tu.tu_accfinal <> tu.id AND acc.id <> acc.tu_accfinal
```

- tu_acctaxon should not have relevant information

```
SELECT *
FROM tu
WHERE tu_acctaxon is NOT NULL AND tu_accfinal IS NULL OR tu_acctaxon <>  tu_accfinal
```

- Ranks

kingdom is identified by first number in tu_sp.
Only the kingdoms 2-7 should are supported, others should not exist and the code may not work

```
  SELECT *
  FROM tu
  WHERE tu.tu_sp not like '#2#%' AND tu.tu_sp not like '#3#%' AND tu.tu_sp not like '#4#%' AND tu.
tu_sp not like '#5#%' AND tu.tu_sp not like '#6#%' AND tu.tu_sp not like '#7#%'
```

- Nomen nudum should not be accepted taxon of another taxon

```
SELECT acc.tu_status, acc.tu_displayname acc, syn.tu_accfinal, syn.id , syn.tu_displayname syn
FROM    tu acc INNER JOIN
              tu AS syn ON acc.id = syn.tu_accfinal
WHERE  (acc.tu_status = 3)  AND syn.tu_accfinal <> syn.id
```

- accepted taxa should not have an unacceptreason

```
SELECT *
FROM tu
WHERE tu_status = 1 AND tu_unacceptreason IS NOT NULL
```

with some rare exceptions (in 2019 6 cases, not yet discussed if necessary)

- tu_status

There should be no synonyms with tu_status == 1

```
SELECT  *
FROM [tu] INNER JOIN tu acc ON tu.tu_accfinal = acc.id
WHERE tu.tu_accfinal <> tu.id AND tu.tu_status = 1
```

Note: see below for reverse condition

- Run ErmsImportActivator in CHECK_ONLY mode (and move above queries to validation)

open issues

- Misapplications

Currently misapplications are identified by a UNION statement in ErmsTaxonImport.getAcceptedTaxaKeys(). It searches for misidentification patterns in tu.tu_unacceptreason . We also tried to evaluate the tu_authority field but searching for "sensu", "non", "auct". But realized that all these patterns are not used specific enough in ERMS. "Auct" is often simply used for misspellings, "sensu" is used for unassessed taxa, genus transfer, sensu in the meaning of sec and others or simply unclear. "Non" is generally to unspecific as it can be used for homonyms, too.
So only the above query is possible but discussion with VLIZ is needed on data cleaning or standardization. Maybe even an own tu_status for misapplications is possible.
Also the status of the following taxa should be checked again:

```
SELECT *
from  tu
where tu.tu_unacceptreason like 'misidentification%' AND not (trim(tu.tu_unacceptreason) like 
'misidentification' OR trim(tu.tu_unacceptreason) like 'misidentifications')
```

However, the PESI DW does not have a status "misapplied" but only synonym and therefore recognizing misapplications is not (yet) so important. But it might be useful for cleaning the authorship data and maybe for future DW versions.

- Name status

The name status is only guessed by the value in tu_unacceptreason. As this field is not standardized this is not exact. The current patterns are mostly taken from the SQL script + some obvious further mappings.
A further discussion with VLIZ about the semantics of some values is necessary for further mapping (especially on ISZN related issues). Start with values with high counts!
See also Y:\BDI\PESI\sh\DBs2PESI\vCDM\erms300_Match_Relation&Status.sql

- Multiple open issues to be hanled by or discussed with VLIZ, mentioned in email to VLIZ 2019-12-23

- the status of taxa with status "alternate representation" should be discussed with VLIZ (and mayb ERS)

- interim published taxa

```
SELECT *
FROM  tu
WHERE tu_status = 9
```

None, or only very few should exist. Open question: should they be exported at all?

- type designations without type

```
SELECT *
FROM tu WHERE tu_typedesignation IS NOT NULL
  AND tu_typetaxon IS NULL
ORDER BY tu_rank, tu_typedesignation
```

These have a typedesignation (by original designation, by subsequent designation, ...) but no name. All are of rank subgenus or higher (and therfore should have a name type, not a specimen type).
Current count = 3177.

Possible explanations: name types are not European or missing data (not discussed with VLIZ yet)

Protozoa and Chromista

- which nomenclatural code to use? (currently both ICNafp, see ErmsTransformer.kingdomId2NomCode)
- what is correct formatting for infrageneric taxa (only Chormista), ERMS currently formats like ICZN

```
SELECT id, tu_displayname, tu_sp
FROM tu
WHERE ( (tu_sp LIKE '#7#%') OR (tu_sp LIKE '#5#%')) AND (tu_rank > 180) AND (tu_rank < 220)
```

Distributions

- take **occurrence_id** into account (currently not yet used, information about distribution status, needs mapping)
- the mapping of areas to countries and the "European" is not clear (handled in above mail)
- distributions attached to synonyms

```
SELECT  v.*, t.tu_status, t.tu_displayname, t.id, t.tu_accfinal
FROM dr v
INNER JOIN tu t ON v.tu_id = t.id
WHERE  t.tu_accfinal <> t.id
ORDER BY  tu_status
```

very few distributions are attached to synonyms (n=23), this was 0 before. Check with VLIZ if this is dirty data.

see also #1523 for further information

- There are exact duplicates: (#1444#note-37, handled in mail)

- Some distributions have a year, is this relevant?

```
SELECT tu_id FROM dr WHERE endyear IS NOT NULL
```

Does it mean the taxa in not observed there anymore is it really absent?
Do we generally include extinct taxa in PESI or should we neglect them?

```
SELECT count(*) as n , tu_id, gu_id, source_id, unacceptsource_id, unacceptreason, tu_valid,
valid_flag, certain_flag, map_flag, endemic_flag, exotic_flag, typelocality_flag, specimen_flag,
vagrant_flag, occurrence_id, origin_id, invasiveness_id, lat, long, depthshallow, depthdeep,
beginyear, beginmonth, beginday, endyear, endmonth, endday, min_abundance, max_abundance,
qualitystatus_id, noteSortable
FROM dr
GROUP BY   tu_id, gu_id, source_id, unacceptsource_id, unacceptreason, tu_valid, valid_flag,
certain_flag, map_flag, endemic_flag, exotic_flag, typelocality_flag, specimen_flag, vagrant_flag,
 occurrence_id, origin_id, invasiveness_id, lat, long, depthshallow, depthdeep, beginyear,
beginmonth, beginday, endyear, endmonth, endday, min_abundance, max_abundance, qualitystatus_id,
noteSortable
HAVING count(*)  > 1
ORDER BY n DESC
```

a few more without using qualitystatus_id in group by/select

(I added column noteSortable)

Languages

- very few (3?) really do not have a 639_3 code, others should be updated, handled in mail

```
SELECT        LanID, LanName, Status, Partner_Agency, [639_3], [639_2], B_Code, bt_equiv, [639_1],
 Element_Scope, Language_Type, Documentation
FROM            languages
WHERE         (LanID IN
                             (SELECT        lan_id
                               FROM            vernaculars)) AND (LanID <> [639_3])
ORDER BY LanName
```

- in ERMS are still 2 records for Romanian (ron and run), formerly one was Moldovian (which does not exist or Moldavian which is the same as Romanian - according to SIL)

```
SELECT LanID, LanName, Status, Partner_Agency, [639_3], [639_2], B_Code, bt_equiv, [639_1],
Element_Scope, Language_Type, Documentation
FROM  languages
WHERE [639_1] = 'ro' OR [639_2] = 'ron' OR [639_2] = 'rum' OR LanID = 'ron' OR LanID = 'rum'
```

One of them should be removed by VLIZ, but are already mapped to the same PESI languages (id=33, id=29 was removed)

Data cleaning:

- tu_accfinal IS NULL

```
  SELECT *
  FROM tu
  WHERE tu_accfinal IS NULL
```

should be checked. VLIZ should fix these (n=479). Currently they are handled as accepted taxa.

- synonyms being the accepted taxon of there parent

```
SELECT *
FROM tu syn
LEFT JOIN tu AS accTaxon ON myTaxon.tu_accfinal = accTaxon.id
WHERE  syn.tu_accfinal IS NOT NULL AND myTaxon.id = accTaxon.tu_parent AND accTaxon.id <> myTaxon.
id
```

These are often alternate representations (children) or autonyms (parents). They are explicitly handled in ErmsTaxonImport.getAcceptedTaxaKeys()

Note: accTaxon.id <> myTaxon.id is to remove Biota from query

- tu_status

There should be no accepted taxa with status 2,3,5 (unaccepted, nomen nudem, alternate representation)

```
SELECT  *
FROM [tu] INNER JOIN tu acc ON tu.tu_accfinal = acc.id
WHERE tu.tu_accfinal = tu.id AND tu.tu_status NOT IN (1, 7, 10, 8, 9, 6)
ORDER BY tu.tu_status
```

other open issues

- is there any meaning in the tu.tsn attribute?
- trim on notes, links, and others (difficult as they are often ntext, but sortableNote can be used)
- standardize sources.source_year field (currently many are imported as freetext)

known incorrect name handling on ERMS portal (http://www.marbef.org/data/aphia.php, or WoRMS http://www.marinespecies.org/)

- 4 parted names with incorrect rank marker, e.g (752445: Cyclonassa neritea var. kamieschensis lactescens <=> neritea kamieschensis subvar. lactescens), these should be logged during import or validation (but maybe commented)
- check again special cases from #1444#note-38

Related issues:

Follows EDIT - task #7976: [PESI][ERMS] Erms taxon import	**Closed**

History

#1 - 12/26/2019 12:06 PM - Andreas Müller

- Due date set to 02/10/2010

- Start date changed from 12/26/2019 to 02/10/2010

#2 - 12/26/2019 12:08 PM - Andreas Müller

- Due date deleted (02/10/2010)

- Target version changed from PESI 2019 to PESI 2022

- Start date deleted (02/10/2010)

#3 - 12/26/2019 12:16 PM - Andreas Müller

- Description updated

#4 - 12/26/2019 12:18 PM - Andreas Müller

- Description updated

#5 - 12/26/2019 12:18 PM - Andreas Müller

- Description updated

#6 - 12/26/2019 12:31 PM - Andreas Müller

- Description updated

#7 - 12/26/2019 12:44 PM - Andreas Müller

- Description updated

#8 - 12/26/2019 12:47 PM - Andreas Müller

- Description updated

#9 - 12/26/2019 12:49 PM - Andreas Müller

- Description updated

#10 - 12/26/2019 12:53 PM - Andreas Müller

- Description updated

#11 - 12/26/2019 01:00 PM - Andreas Müller

- Description updated

#12 - 12/26/2019 01:11 PM - Andreas Müller

- Description updated

#13 - 12/26/2019 01:22 PM - Andreas Müller

- Description updated

#14 - 12/26/2019 01:46 PM - Andreas Müller

- Description updated

#15 - 12/26/2019 02:11 PM - Andreas Müller

- Description updated

#16 - 12/26/2019 02:12 PM - Andreas Müller

- Description updated

#17 - 12/26/2019 02:14 PM - Andreas Müller

- Description updated

#18 - 12/26/2019 02:24 PM - Andreas Müller

- Description updated

#19 - 12/26/2019 02:46 PM - Andreas Müller

- Description updated

#20 - 12/26/2019 03:30 PM - Andreas Müller

- Description updated

#21 - 12/26/2019 03:36 PM - Andreas Müller

- Description updated

#22 - 12/26/2019 03:42 PM - Andreas Müller

- Description updated

#23 - 12/26/2019 03:53 PM - Andreas Müller

- Description updated

#24 - 12/26/2019 04:01 PM - Andreas Müller

- Description updated

#25 - 12/26/2019 04:03 PM - Andreas Müller

- Description updated

#26 - 12/26/2019 04:05 PM - Andreas Müller

- Description updated

#27 - 12/26/2019 04:52 PM - Andreas Müller

- Description updated

#28 - 12/26/2019 04:55 PM - Andreas Müller

- Description updated

#29 - 12/26/2019 04:57 PM - Andreas Müller

- Description updated

#30 - 12/26/2019 05:02 PM - Andreas Müller

- Description updated

Open issues from comparing SQL scripts (Y:\BDI\PESI\sh\DBs2PESI\vCDM):

!!!! Do check again because maybe I checked against Y:\BDI\PESI\sh\DBs2PESI\v12 instead vCDM !!!

erms130_language.sql:

- not yet checked (low priority)

erms230_Source.sql:

- Parsing RefYear and AuthorString (low priority)

erms300_Match_Relation&Status.sql: MEDIUM

- test synonym relations
- same attribute => to avoid duplicates?

erms400_Taxon.sql: HIGH

- not yet checked
- update Taxon set KingdomFk = 6 where GenusOrUninomial = 'Monera'

erms420_UpdateTaxon.sql: DONE

erms430_Kingdom.sql: DONE (updates Rank.Kingdom, probably not relevant)

erms450_RelTaxon.sql:

- Name relations => always to accepted taxon, where tu.tu_status > 1 and q.QualifierId < 100
- taxon relations => always to accepted taxon, where et.id != et.tu_accfinal and q.QualifierId > 100
- fill RelTaxon via tu=>tu_accfinal != tu=>id ; ! cave: do not duplicate entries via Matching table, only tu_unacceptreason = NULL
- by default use 'is synonym of' if unacceptreason does not exist or does not match any >100 RealtionType, where et.id != et.tu_accfinal
- 'is taxonomically included in', where et.tu_parent is not null and et.tu_status = 1 and et.id <> et.tu_parent

- put tu_unacceptreason into notes

erms470_NameStatus.sql: DONE

erms500_Vernacular.sql: DONE (source handling is new and differs from SQL)

erms520_Images.sql: see open issues in image export (handling of multiple images)

erms550_Note.sql:

- update Note set Note_1 = replace(Note_1, char(11), ' ') (low priority)
- tests for ecology facts

erms560_Link.sql: LOW

- final tests
- "encyclopedia of => Image"
- "fishbase ambigous => Discuss with VLIZ

erms600_Source_2.sql: DONE (test for sources still missing)

erms900_Occurrence.sql: DONE (see #1523 for comments on status handling)

#31 - 12/26/2019 05:13 PM - Andreas Müller

- Description updated

#32 - 12/26/2019 05:13 PM - Andreas Müller

- Description updated

#33 - 12/26/2019 05:45 PM - Andreas Müller

- Description updated

#34 - 12/26/2019 06:13 PM - Andreas Müller

- Description updated

#35 - 12/26/2019 06:32 PM - Andreas Müller

- Description updated

#36 - 12/26/2019 06:39 PM - Andreas Müller

- Description updated

#37 - 12/26/2019 06:41 PM - Andreas Müller

- Due date set to 01/11/2019

- Start date set to 01/11/2019

- Follows task #7976: [PESI][ERMS] Erms taxon import added

#38 - 12/26/2019 06:42 PM - Andreas Müller

- Due date deleted (01/11/2019)

- Start date deleted (01/11/2019)

#39 - 12/26/2019 10:53 PM - Andreas Müller

- Description updated

#40 - 12/26/2019 11:21 PM - Andreas Müller

- Description updated

#42 - 12/27/2019 12:18 AM - Andreas Müller

- Description updated

#43 - 12/27/2019 12:40 AM - Andreas Müller

- Description updated

#44 - 12/27/2019 12:43 AM - Andreas Müller

- Description updated