EDIT - task #7976

[PESI][ERMS] Erms taxon import

01/10/2019 02:44 PM - Andreas Müller

Status:	Closed		**Start date:**	
Priority:	Highest		**Due date:**	
Assignee:	Andreas Müller		**% Done:**	100%
Category:	cdmadapter		**Estimated time:**	0:00 hour
Target version:	PESI 2019			
Severity:	normal			

Description

Subticket for #1444:

The taxon import in ERMS is the most complex one therefore we open a new ticket for it.

Kingdoms:

1: - (Super)domain
2: Animalia
3: Plantae
4: Fungi
5: Protozoa
6: Bacteria
7: Chromista ??

8-14 are not relevant in ERMS (yet) - they do not appear in any tu.tu_sp tree as root

The Taxon table tu:

id: the id
tsn: ? (currently exported as extension)
tu_name: the epithet or uninomial
tu_displayname: full name without author (html?)
tu_authority: author and year (without et. al.?)
tu_parent: links to the parent, BUT for synonyms the parent is considered the name giving parent, e.g. for a species it is usually the genus given in the name and not the parent of the accepted name, for subgenus it is usually the genus; however, it is NOT always the case that the parent is name giving, there are also parent with no obvious relationship and therefore it is not clear what the parent relationship really means; tu_parent is always the last id in tu.tu_sp and therefore it is essential for creating the tree
tu_sp: the higher classification as tree index, BUT it is not the correct taxonomic classification as synonyms link to their nomenclatural parent and not to their accepted taxon

Most important issue:

- Decide on taxon relationships (parent-child, synonym-relationships
  - tu_parent is primarily nomenclatural and should only be used where a taxon is considered to be THE accepted taxon

Related issues:

Precedes EDIT - task #8792: Issues to have in mind for ERMS import	**New**

Associated revisions

Revision cd79d676 - 02/14/2019 02:28 PM - Andreas Müller

ref #1444 ref #7976 some changes to ERMS import

Revision 58968197 - 12/26/2019 07:21 PM - Andreas Müller

ref #1444, ref #7976 improve misapplication handling in ERMS import

Revision 8f402cc0 - 12/27/2019 09:39 AM - Andreas Müller

ref #1444, ref #2871, ref #7976 improve authorship handling for MAN

Revision c12d2903 - 12/27/2019 11:09 PM - Andreas Müller

ref #1444, ref #7976 better logging for unhandled unacceptReason

Revision c0411992 - 12/27/2019 11:11 PM - Andreas Müller

ref #1444, ref #7976 fix erronous subgenus formatting in ERMS

History

#2 - 01/10/2019 02:52 PM - Andreas Müller

- Description updated

#3 - 01/11/2019 09:25 AM - Andreas Müller

- Description updated

Parents are generally nomenclatural parents not taxonomical parents. There are different parent types (this is not really important as we will use parents only for nomenclatural reasons).

Subgenus with acc is same name (296)

```
SELECT acc.id, acc.tu_rank, acc.tu_displayname ,acc.tu_sp, acc.tu_parent, acc.tu_status,syn.id,
       syn.tu_name synName, syn.tu_rank synRank, syn.tu_displayname synDisplay, syn.tu_unacceptreason, syn.
tu_sp synTree, syn.tu_parent synParent, syn.tu_status synStatus
FROM tu syn INNER JOIN tu acc ON acc.id = syn.tu_acctaxon
WHERE syn.tu_status = 2 AND syn.tu_sp <> acc.tu_sp
  AND syn.tu_rank < 220 and syn.tu_displayname like '%(%' AND syn.tu_name = acc.tu_name
ORDER BY acc.tu_sp
```

Subgenus with accepted is not same genus

```
syn.tu_status = 2 AND syn.tu_sp <> acc.tu_sp AND syn.tu_rank < 220 and syn.tu_displayname like '%(%' AND syn.t
u_name <> acc.tu_name
```

etc.

#4 - 01/11/2019 09:27 AM - Andreas Müller

Errors in Data:

- Name inheritance:

Some (63) children differ in first name part from parent. Sometimes this is simply a wrong whitepace, but more often it is a misspelling:

```
  SELECT syn.id, synParent.tu_name parentEpi, synParent.tu_displayname parentDisplay, trim(replace(replace(
replace(left (syn.tu_displayname, len(syn.tu_displayname)- len(syn.tu_name)), ' var.', ''), ' f.', ''),
' subsp.', '')) shorted,
      syn.tu_name synName, syn.tu_rank synRank, syn.tu_displayname synDisplay, syn.tu_unacceptreason, right(
syn.tu_sp, 30) synTree, syn.tu_parent synParent,    syn.tu_status synStatus
  FROM tu syn
    INNER JOIN tu synParent ON synParent.id = syn.tu_parent
  WHERE syn.tu_rank >= 220 AND synParent.tu_displayname <> trim(replace(replace(replace(left (syn.
tu_displayname, len(syn.tu_displayname)- len(syn.tu_name)), ' var.', ''), ' f.', ''), ' subsp.', ''))
```

Subgenus seems not to have this problem and for genus and above it is not relevant anyway:

```
SELECT syn.id, synParent.tu_name parentEpi, synParent.tu_displayname parentDisplay, trim(left (replace(replace
(syn.tu_displayname, ' (', ''), ')', ''), len(syn.tu_displayname)- len(syn.tu_name)-3)) shorted,
      syn.tu_name childName, syn.tu_rank childRank, syn.tu_displayname childDisplay, syn.tu_unacceptreason,
right(syn.tu_sp, 30) synTree, syn.tu_parent childParent, syn.tu_status childStatus
FROM tu syn
  INNER JOIN tu synParent ON synParent.id = syn.tu_parent
WHERE syn.tu_rank <220 AND syn.tu_rank >180 AND synParent.tu_displayname <> trim(left (replace(replace(syn.
tu_displayname, ' (', ''), ')', ''), len(syn.tu_displayname)- len(syn.tu_name)-3))
```

- there is one case, where a nomen nudum is the accepted taxon for an other taxon

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

- there are 3317 cases where the tu_acctaxon has not a status "accepted" (2213 are self-referencing, 1104 are really different records, none of the "accepted" link again to another record, though there seem to be such records in WORMS, e.g.

```
SELECT  acc.id,  acc.tu_status AS accStatus, acc.tu_displayname AS accName,
        syn.id, syn.tu_status AS synStatus, syn.tu_displayname AS synName
FROM    tu acc INNER JOIN
        tu AS syn ON acc.id = syn.tu_acctaxon
WHERE   (acc.tu_status <> 1) -- AND  syn.id = acc.id --  AND acc.id <> acc.tu_acctaxon
ORDER BY accStatus, synStatus
```

- 289 cases where parent taxon is not accepted => probably not critical assuming that the parent child relationship is not a taxonomical one

- 610 taxa being alternative names but accepted taxon has not status accepted. Especially critical for both having status 5 (alternative name)

```
SELECT  acc.id, acc.tu_status AS accStatus, acc.tu_displayname AS accName,
        syn.id, syn.tu_status AS synStatus, syn.tu_displayname AS synName
FROM    tu acc INNER JOIN
        tu AS syn ON acc.id = syn.tu_acctaxon
WHERE   (acc.tu_status <> 1) AND (syn.tu_status IN (1,5,9,10))  AND  syn.id <> acc.id
-- AND acc.id <> acc.tu_acctaxon
ORDER BY accStatus, synStatus
```

#5 - 01/11/2019 12:36 PM - Andreas Müller

- Description updated

#6 - 01/11/2019 03:17 PM - Andreas Müller

699 accepted taxa (acctaxon = id) do have a parent which is not accepted. 2/3 of the parents has status synonym, 1/3 has status alternate representation.

~~This is CRITICAL as these children are difficult to handle.~~ => not really true anymore with the current strategy to build tree only from tu_accfinal accepted taxa

```
SELECT  child.tu_status childStatus, child.tu_displayname child, substring(child.tu_sp, 2,1) kingdom, child.
tu_rank childRank, child.id childId, child.tu_acctaxon childAccId,
 par.tu_status parStatus, par.id, par.tu_name, par.tu_displayname, par.tu_acctaxon, par.tu_rank, accPar.
tu_displayname accParent
FROM  tu child INNER JOIN tu par ON par.id = child.tu_parent INNER JOIN tu accPar ON par.tu_acctaxon = accPar.
id
WHERE child.tu_acctaxon = child.id AND par.tu_acctaxon <> par.id -- AND par.tu_status = 2
ORDER BY  kingdom, child.tu_rank, par.tu_status
```

#7 - 09/01/2019 06:12 PM - Andreas Müller

- Priority changed from New to Highest

#8 - 10/02/2019 12:12 AM - Andreas Müller

- Description updated

#9 - 10/02/2019 08:34 AM - Andreas Müller

There are 82 tu_status=accepted taxa with unaccept reason: (mail sent)

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

#10 - 10/02/2019 09:12 AM - Andreas Müller

ACCEPTED TAXON:

336 taxa have NO accepted taxon (tu_acctaxon IS NULL)

57003 are their own accepted taxon (tu_acctaxon = id) (tu_status : 1=54790, 2=8, 3=10, 5=29, 6=410, 7=71, 8=385, 10=1300)

35706 have another accepted taxon (tu_acctaxon = id) (tu_status: 2=33804, 3=39, 5=1842, 6=21)

```
SELECT *  --  tu_status, count(*) as n
FROM tu
WHERE tu_acctaxon =/<> id/IS NULL
-- GROUP BY tu_status
ORDER BY tu_status
```

#11 - 10/02/2019 02:34 PM - Andreas Müller

- Description updated

#12 - 10/02/2019 02:37 PM - Andreas Müller

NO synonyms exist with an accepted taxon again being a synonym of another accepted taxon:

```
SELECT  acc.id accId, acc.tu_rank, acc.tu_displayname accName ,acc.tu_sp, acc.tu_parent, acc.tu_status,syn.id,
 syn.tu_name synName, syn.tu_rank
        synRank, syn.tu_displayname synDisplay, syn.tu_unacceptreason, syn.tu_sp synTree, syn.tu_parent, synPa
rent, syn.tu_status synStatus
FROM tu syn LEFT JOIN tu acc ON acc.id = syn.tu_acctaxon -- AND syn.id <> syn.tu_acctaxon
WHERE
 syn.id <> syn.tu_acctaxon
 AND acc.id <> acc.tu_acctaxon
```

#13 - 11/29/2019 04:09 PM - Andreas Müller

- Status changed from New to In Progress

#14 - 12/26/2019 06:41 PM - Andreas Müller

- Status changed from In Progress to Closed

- % Done changed from 0 to 100

#15 - 12/26/2019 06:41 PM - Andreas Müller

- Precedes task #8792: Issues to have in mind for ERMS import added

```
SELECT  acc.id accId, acc.tu_rank, acc.tu_displayname accName ,acc.tu_sp, acc.tu_parent, acc.tu_status,syn.id,
 syn.tu_name synName, syn.tu_rank
        synRank, syn.tu_displayname synDisplay, syn.tu_unacceptreason, syn.tu_sp synTree, syn.tu_parent, synPa
```