EDIT - task #7420

Import for higher taxon graph for phycobank

05/15/2018 04:09 PM - Andreas Kohlbecker

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

Description

The higher taxon graph as discussed in #6173 will be imported via spreadsheets whereas we will use the normal implicit import format. The details of the format which will be defined by issue #7419

Taxa will be related as taxon relationships as well as classifications. The later is only for inner handling, not for showing or searching taxa in data portal via classifications.

Secundum for taxa should be nom. ref. => nominal taxa

The resulting taxonGraph is visualized at: http://api.cybertaxonomy.org/taxonGraph/

Related issues:

Related to PhycoBank - task #6173: Concept for a useful algae registry taxon ...	**Closed**
Related to EDIT - task #6137: Urgent imports	**In Progress**
Related to EDIT - task #7948: Editor for Classification Fragments	**Closed**
Related to EDIT - bug #10272: Import WoRMS (and other) higher classifications...	**New**
Blocked by EDIT - feature request #7419: Provide example NormalImplicit impor...	**Closed**
Copied to PhycoBank - task #7808: Further name duplicates	**New**
Copied to EDIT - task #7811: Import higher classifications	**Feedback**

Associated revisions

Revision cc9428e1 - 08/10/2018 02:05 PM - Andreas Müller

ref #7420 first version of phycobank higher classification import

Revision 2b1e32ea - 09/09/2018 10:05 PM - Andreas Müller

ref #7420 add config parameter to getWorksheetname for Excel import

Revision f36c41f1 - 09/09/2018 10:07 PM - Andreas Müller

ref #7420 adding phycobank as source

Revision 9eca8b4a - 09/09/2018 10:09 PM - Andreas Müller

ref #7420 define WorksheetName as config parameter (needed for Phycobank import)

Revision 781b44c5 - 09/27/2018 03:27 PM - Andreas Müller

ref #7420 updates to phycobank import according to requirements

Revision ae79d763 - 10/16/2018 05:45 PM - Andreas Müller

ref #7420 latest changes to PhycobankActivator

Revision 21e704da - 02/19/2019 04:01 PM - Andreas Müller

ref #7420 last changes to Phycobank higher classification import

History

#1 - 05/15/2018 04:10 PM - Andreas Kohlbecker

- Blocked by feature request #7419: Provide example NormalImplicit import file for phycobank higher taxon graph imports added

#2 - 05/15/2018 04:10 PM - Andreas Kohlbecker

- Related to task #6173: Concept for a useful algae registry taxon classification added

#3 - 05/18/2018 02:54 PM - Andreas Müller

- Subject changed from Adapt NormalImplecit import to import higher taxon graph phycobank to Import for higher taxon graph for phycobank

- Target version changed from Unassigned CDM tickets to Release 5.1

#4 - 05/18/2018 02:55 PM - Andreas Müller

- Related to task #6137: Urgent imports added

#5 - 05/18/2018 02:58 PM - Andreas Müller

- Description updated

#6 - 06/27/2018 02:28 PM - Andreas Müller

- Target version changed from Release 5.1 to Release 5.2

#7 - 07/31/2018 01:27 PM - Andreas Kohlbecker

- File Algen_Syllabus_NormalImplied_Test.xls added

Hi, here comes the spread sheet with the first data for the import: attachment:Algen_Syllabus_NormalImplied_Test.xls

das Excel-File „Algen_Syllabus_NormalImplied_Test“ enthält das alte Blatt „NormalImplied.txt“ mit Gegenüberstellung von Syllabus und WoRMS.

Auf Basis der bisherigen Besprechungen habe ich für die weitere Diskussion ein neues Blatt für den Syllabus angelegt „HigherRanksEnfwurfNeu“

Andreas M. bat darum, alle höheren Ränge, die bekannt sind explizit zu benennen. Fertig, alle gelb unterlegten Rangstufen sind nicht vorhanden oder werden im Syllabus nicht genutzt.
Gelb unterlegte Ränge, die leer sind, ersetzten vielfach eine Bezeichnung wie „incertis sedis“.

Fragen zu den Gattungsnamen:

1. Einige haben wir schon als Namen im System, einige als registrierte Gattungen, andere haben wir noch nicht im System, benötigen sie aber als Namen ggf. als registrierte Namen, Vorgehen?
2. Andreas M. fragte nach nomenklatorischen Autoren, wegen eventueller Homonyme
3. Brauchen wir eine Info zum Status (valid, invalid, illeg.)? für die Gattungen? (ich würde nur sagen, wenn sie registriert werden.)

#8 - 08/09/2018 10:44 AM - Andreas Müller

Please save future versions of the excel file in the current .xslx format, .xls is outdated.

#9 - 08/09/2018 10:45 AM - Andreas Müller

- File Algen_Syllabus_NormalImplied_Test.xlsx added

#10 - 08/09/2018 10:45 AM - Andreas Müller

- File deleted (Algen_Syllabus_NormalImplied_Test.xls)

#11 - 08/10/2018 02:07 PM - Andreas Müller

- Status changed from New to In Progress

- Priority changed from New to Highest

- Target version changed from Release 5.2 to Release 5.3

- % Done changed from 0 to 30

- Severity changed from normal to major

First version of import is ready and tested with Frey and WoRMS data running into empty local database.

Next step: test with running into phycobank database.

#12 - 08/10/2018 02:16 PM - Andreas Müller

- File Algen_Syllabus_NormalImplied_Worms_Test.xlsx added

#13 - 08/10/2018 02:23 PM - Andreas Müller

What should be done with existing IAPT data. Should they be adapted to the new data model?

#14 - 08/10/2018 02:27 PM - Andreas Müller

Andreas Müller wrote:

> What should be done with existing IAPT data. Should they be adapted to the new data model?

Also we need to decide how to handle IAPT species. They are currently attached to IAPT genus. Is this still wanted in future?

Also IAPT genus do have authors. We need to decide if these genus names should be matched during import of new names which currently have no authors.

#15 - 09/09/2018 10:16 PM - Andreas Müller

- Status changed from In Progress to Resolved

- Assignee changed from Andreas Müller to Andreas Kohlbecker

- % Done changed from 30 to 50

please review results of import on test.cdm_phycobank

#16 - 09/17/2018 11:33 AM - Andreas Müller

- Status changed from Resolved to Feedback

- Assignee changed from Andreas Kohlbecker to Andreas Müller

- Target version changed from Release 5.3 to Release 5.4

Needs to be adapted

#17 - 09/26/2018 05:43 PM - Andreas Kohlbecker

Hi Andreas,

a small change of the strategy: All Taxa and TaxonRelations which belong to the classifications-graph should have the Phycobank as secReference or citation in case of the relations. By this the taxa and relations belonging to the graph can be identified.

Andreas

#18 - 09/27/2018 02:24 PM - Andreas Müller

Andreas Kohlbecker wrote:

> Hi Andreas,
>
> a small change of the strategy: All Taxa and TaxonRelations which belong to the classifications-graph should have the Phycobank as secReference or citation in case of the relations. By this the taxa and relations belonging to the graph can be identified.
>
> Andreas

I understand this for the taxa as we use only 1 taxon per name so we can't give sec references per concept used. This was already decided before.

I do not understand it for taxon relations. I thougt for them the references should be references of the classification used to not loose this information. Which garph do you need to identify this way. Are there any other "taxonomically included in" relationships expected in the database then those for phycobank? And for which use-case do you need to identify the graph?

Also I should mention that we think about having a "graph" link for each relationship in (near) future. This is an idea that comes from discussing different types of DefinedTerm relationship graphs (collections, lists, trees, directed graphs, undefined graphs). This way you can group graphs while using reference or soon sources.reference is not a good idea for holding graph data together for multiple reasons.

#19 - 09/27/2018 02:59 PM - Andreas Müller

- Status changed from Feedback to Resolved

- Assignee changed from Andreas Müller to Andreas Kohlbecker

I did run a new import to test.cdm_phycobank. Please check the results.

One issue I can see is that the existing IAPT data sometimes do have different ranks then the imported data. Therefore the names/taxa are not deduplicated. Example: Cryptophyceae is Division in IAPT but Phylum in Frey + Worms. This needs to be sorted out before the final import.

#20 - 09/27/2018 03:05 PM - Andreas Müller

You should check by

```
SELECT titleCache, count(*) as n
FROM TaxonName tn
GROUP BY tn.titleCache
Having n > 1
```

or

```
SELECT nameCache, count(*) as n
FROM TaxonName tn
GROUP BY tn.nameCache
Having n > 1
```

for multiple occurrences of the same name in phycobank. There are even names with preliminary flag (or nameCache == null). This should probably not happen in a registration database.

#21 - 10/08/2018 08:31 AM - Andreas Kohlbecker

Andreas Müller wrote:

> You should check by
>
> ```
> SELECT titleCache, count(*) as n
> FROM TaxonName tn
> GROUP BY tn.titleCache
> Having n > 1
> ```
>
> or
>
> ```
> SELECT nameCache, count(*) as n
> FROM TaxonName tn
> GROUP BY tn.nameCache
> Having n > 1
> ```
>
> for multiple occurrences of the same name in phycobank. There are even names with preliminary flag (or nameCache == null). This should probably not happen in a registration database.

Genus name duplicates are already handled in #7748

#22 - 10/08/2018 08:54 AM - Andreas Kohlbecker

all other data import and data cleaning issues are copied to #7808

#23 - 10/08/2018 08:55 AM - Andreas Kohlbecker

- Copied to task #7808: Further name duplicates added

#24 - 10/08/2018 10:45 AM - Andreas Kohlbecker

- Description updated

#25 - 10/08/2018 11:04 AM - Andreas Kohlbecker

- Status changed from Resolved to Feedback

- Assignee changed from Andreas Kohlbecker to Andreas Müller

- % Done changed from 50 to 90

I reviewed the imported taxon graph relations. The resulting graph exactly matches the expectations.
The implementation is ready, so we now need the complete higher classification data for the final imports of the various classifications.
I think we should close this ticket in favor of creating a new ticket for the actual import tasks.

#26 - 10/08/2018 11:21 AM - Andreas Müller

Ok, can you close the ticket and open a ticket for what ever you stil need?

#27 - 10/08/2018 02:59 PM - Andreas Kohlbecker

- Copied to task #7811: Import higher classifications added

#28 - 10/08/2018 03:00 PM - Andreas Kohlbecker

- Status changed from Feedback to Closed

- % Done changed from 90 to 100

new ticket for the actual import tasks created: #7811

#29 - 10/25/2018 10:45 AM - Andreas Müller

- Target version deleted (Release 5.4)

#30 - 02/22/2022 03:03 PM - Andreas Müller

- Related to task #7948: Editor for Classification Fragments added

#31 - 03/14/2023 04:33 PM - Andreas Müller

- Related to bug #10272: Import WoRMS (and other) higher classifications for Phycobank added

Files

Algen_Syllabus_NormalImplied_Test.xlsx	13.4 KB	08/09/2018	Andreas Müller
Algen_Syllabus_NormalImplied_Worms_Test.xlsx	15.2 KB	08/10/2018	Andreas Müller