Eli Agbayani <eagbayani@eol.org>

parents method

63 messages

Eli Agbayani <eagbayani@eol.org> Mon, Sep 17, 2018 at 12:16 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
The next task is the 'parents' method, right.

With the 2 recently done examples for 'taxon summary'
page_id = 46559118; predicate = "http://purl.obolibrary.org/obo/RO_0002439"; //preys on
page_id = 328607; predicate = "http://purl.obolibrary.org/obo/RO_0002470"; //eats

And one in our 'basal values':
page_id = 7662; predicate = "http://eol.org/schema/terms/Habitat" //habitat includes

Can you give me an example how to proceed with the 'parents' method.

e.g.
http://eol.org/schema/terms/Habitat habitat includes basal value
http://purl.obolibrary.org/obo/RO_0002439 preys on taxon summary
http://purl.obolibrary.org/obo/RO_0002470 eats taxon summary

Please tell me if you need other working files to achieve this. Seems like it.

Thanks,
Eli

3 attachments

7662_habitat_includes.txt
10K

328607_eats.txt
16K

46559118_preys on.txt
5K

Jen Hammock <jen.hammock@gmail.com> Mon, Sep 17, 2018 at 12:33 PM
To: Eli Agbayani <eagbayani@eol.org>

Right, let's see...

What I think I need for each is data for two species with the same predicate. I've got one already. Could you please send me...

taxon summary output for http://purl.obolibrary.org/obo/RO_0002470 for taxon ID 46559162

basal value output for http://eol.org/schema/terms/Habitat for taxon ID 328109 and 328607

?

Thanks!
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Mon, Sep 17, 2018 at 7:50 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, here it is.
By the way, page_id = 328109 is not found in Carnivora dataset (traits.csv).
I replaced it with 328682.
Thanks,
Eli
[Quoted text hidden]

3 attachments

328682_habitat_includes.txt
8K

46559162_eats.txt
13K

328607_habitat_includes.txt
12K

Jen Hammock <jen.hammock@gmail.com> Mon, Sep 17, 2018 at 8:11 PM
To: Eli Agbayani <eagbayani@eol.org>

Sorry! I keep forgetting you're relying on that Carnivora file. And I just realized we'll want more REC records among the species in order to make the example informative. Could I have both predicates also for taxon ids 46559217, 328609, and 328598?

Thanks; I really think I can give you a sample with those...

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Mon, Sep 17, 2018 at 9:07 PM
To: Jen Hammock <jen.hammock@gmail.com>

No problem Jen. Here it is.
Thanks,
Eli

[Quoted text hidden]

6 attachments

328598_eats.txt
20K

328609_eats.txt
28K

46559217_eats.txt
4K

328598_habitat_includes.txt
16K

328609_habitat_includes.txt
13K

46559217_habitat_includes.txt
6K

Jen Hammock <jen.hammock@gmail.com> Tue, Sep 18, 2018 at 4:41 PM
To: Eli Agbayani <eagbayani@eol.org>

Okay, let me know if you can make head or tail of this :) It should be your example for the parent process for taxon summaries. It kinda resembles the regular taxon summary. In fact, a couple of the new steps are things I'd like to apply to the regular taxon summary, but let's see if this works, first.

I'll get working on the basal value example.

Working order isn't sensitive. Feel free to carry on with media curation actions if you're on a roll.

Thanks!

Jen
[Quoted text hidden]

7662 parent taxon summary eats.csv
4K

Eli Agbayani <eagbayani@eol.org> Wed, Sep 19, 2018 at 1:44 AM
To: Jen Hammock <jen.hammock@gmail.com>

Thanks Jen.
Just quick correction. You mentioned:
"pretending it only has 5 children, 328598, 3288609, 46559217, 328682, and 328607"

- just a type, not *3288609* but *328609*
- we don't have predicate 'eats' for page_id = *328682*
- and you've forgot to add 46559162, which you actually used in the example.

Will continue.
Regards,
Eli

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Wed, Sep 19, 2018 at 4:55 AM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, 3 clarifications please.

1. Before the step to reduce hierarchies, you have a NEW STEP: remove 2913056.
Is the rule here: if the root of all the 'Hierarchies of taxon values' is the same, then remove it.
If not, don't do anything.
Is that correct?

2. In my 'Reduced hierarchies'. I have one extra ancestor added (2908256) for all eight hierarchies.
I added this because it meets our 2 criteria (in taxon summary):
first: it exists in other hierarchies
second: its child is not the same throughout all hierarchies
Is there an additional rule that excludes it?

3. In your 'Reduced hierarchies' you have this for 207661:
207661,
--- 1905 {1},
--- 2774383 {1},
--- 1 {1},

In mine, I don't have 1905. Was it included because it exists as a tip in one of the hierarchies?

Thanks. That's all for now.
Regards,
Eli

[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Wed, Sep 19, 2018 at 9:00 AM
To: Eli Agbayani <eagbayani@eol.org>

Oops, thanks on all those corrections, Eli. The note will be helpful when I'm comparing our results, I expect.

Ah, 2913056: no, sorry, I should have said: this node is Life, the root of the whole Dynamic Hierarchy. If the common root of the dataset is anything else, you can leave it. Only remove it if it is 2913056. (That'll mean we have a particularly broad spread of taxa.)

That being said, the 2908256 problem: Ah, I missed a bifurcation. I thought it was always followed by 2910700 and then by 1. You've got all the rules just fine. What this means is that I'll want to expand the removal process for 2913056 to include more of the root of the DH, but for now, you're fine. I'll know better

how much more to prune when I've seen your result.

1905: oops, you're right, that's why I left it in, but that's not consistent with what we were doing in the first taxon summary, is it? If you thought the whole 1905 record should be excluded and the 1905 node excluded from the 207661 record, I think you're right.

Thanks!

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Wed, Sep 19, 2018 at 10:28 AM
To: Jen Hammock <jen.hammock@gmail.com>

Thanks for the answers Jen.

- Yes the inclusion of 1905 is not consistent in the first taxon summary rules. So yes, we can exclude it in the reduced hierarchy for 207661.
But it is okay to be included in prior step for 207661, in 'Hierarchies of taxon values'.

I will now proceed and will ask as I go along.
Thanks,
Eli
[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Wed, Sep 19, 2018 at 10:57 AM
To: Eli Agbayani <eagbayani@eol.org>

Great, thanks!

Meanwhile, here is your basal values parent method example. :)

Jen
[Quoted text hidden]

habitat for 7662.txt
22K

Jen Hammock <jen.hammock@gmail.com> Wed, Sep 19, 2018 at 11:20 AM
To: Eli Agbayani <eagbayani@eol.org>

Annnnnd, harking back to the parents taxon summary, after puttering around a bit, I have a longer list for you of root nodes to remove:

46702381
2910700
6061725
2908256
2913056

You can remove them even if they're tips, I think. Let's see how that goes.

Thanks again!

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Wed, Sep 19, 2018 at 1:20 PM
To: Jen Hammock <jen.hammock@gmail.com>

Okay noted on these additional root nodes (five total) to remove. Instead of just 2913056.

The rule here is:

- If the root node is any of these five, and if it is common to all 'hierarchies_of_taxon_values', then I'll remove that root node from all hierarchies.
- If there are multiple root nodes, but all are included in the magic five -> what to do?
- if there are multiple root nodes, some are outside of the magic five -> what to do?

Thanks,
Jen

[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Wed, Sep 19, 2018 at 1:28 PM
To: Eli Agbayani <eagbayani@eol.org>

Ah, great questions. Let's see...

- If there are multiple root nodes, but all are included in the magic five -> remove all
- if there are multiple root nodes, some are outside of the magic five -> remove magic 5 roots, leave the others

and we'll see what happens...

Thanks!

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Wed, Sep 19, 2018 at 1:32 PM
To: Jen Hammock <jen.hammock@gmail.com>

Noted. Thanks Jen.

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Wed, Sep 19, 2018 at 1:59 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
Attached is the result for test case 7662. For review.

Anyway, you have another NEW step after generating the 'Reduced hierarchies':

"NEW STEP: IF there are multiple roots, discard those representing less than 15% of the original records"

Our test case had only 1 root in the 'Reduced hierarchies' so the new step was skipped.
But can you please explain further.

- When you say "...discard those...", meaning discard the entire hierarchy that meets the criteria.
- What do you mean by 'original records'.

Thanks,
Eli

[Quoted text hidden]

7662 parent taxon summary eats.txt
4K

Jen Hammock <jen.hammock@gmail.com> Wed, Sep 19, 2018 at 2:36 PM
To: Eli Agbayani <eagbayani@eol.org>

Yup, let me see... ooh, those are both quite ambiguous, aren't they?

discard: yes, *in this step* discard means that whole hierarchy

"original records" is a set just upstream of your second section in your result file: "combined values from the original records (all REC records of children), deduplicated:Array". The list I want is before deduplication, i.e. if 207661 was a value for more than one of the child taxa, it should count more than once in the 15% calculation.

I'll have a look at your file and let you know if anything else looks interesting.

thanks!

Jen
[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Wed, Sep 19, 2018 at 2:46 PM
To: Eli Agbayani <eagbayani@eol.org>

Hi again, Eli!

The results file looks good except that 2908256 was retained as a root. It should be on the magic deletable taxa list.

and in your final output section, if I read it right, you don't include the PRM record in the list of REC records. I'm not sure how we'll deal with this ultimately (I need to consult Jeremy) but I suspect in most cases the PRM record should also be included in the REC records (I think queries will be simpler that way. Anyway, you know where they all are and that's the important thing for now.

wheee...

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Wed, Sep 19, 2018 at 3:20 PM
To: Jen Hammock <jen.hammock@gmail.com>

Oh yes 2908256, need to adjust.
It happened because the original root node is 2913056. Seems 2908256 came next after 2913056. A certain loop has to check until no more deletable taxa in root position.

Thanks,
Eli

[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Wed, Sep 19, 2018 at 3:27 PM
To: Eli Agbayani <eagbayani@eol.org>

ah, right; there may even be grandparents in that little cluster of magic nodes.
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Wed, Sep 19, 2018 at 3:34 PM
To: Jen Hammock <jen.hammock@gmail.com>

Here is updated result.
Please holler if anything is still not in order.
Thanks,
Eli

[Quoted text hidden]

7662 parent taxon summary eats v2.txt
3K

Jen Hammock <jen.hammock@gmail.com> Wed, Sep 19, 2018 at 3:48 PM
To: Eli Agbayani <eagbayani@eol.org>

A diet of mostly vertebrates, with the occasional spider or protostome. Yummy!

This looks good, thanks :)

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Fri, Sep 21, 2018 at 2:03 AM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, I'm moving along with the method: (basal values) 'parents'. Looking good. Will give you feedback soon.

But I just wanted to go back to the method: (taxon summary) 'parents'.
Particularly to this new step: we already had some exchanges about it. See thread.
"NEW STEP: IF there are multiple roots, discard those representing less than 15% of the original records"

Anyway, can you modify this attached and make a fake data that will meet the criteria for this new step.
It is better I get it right now, before I move along to other tasks.

Sorry, thanks Jen.
Regards,
Eli

[Quoted text hidden]

7662 parent taxon summary eats v2.txt
3K

Jen Hammock <jen.hammock@gmail.com> Fri, Sep 21, 2018 at 9:34 AM
To: Eli Agbayani <eagbayani@eol.org>

Yes, let's make sure this bit is working. It'll save us some surprising records.

Interesting. I thought there was a record in this dataset that matched the criteria. It's in your hierarchy list, before reducing the hierarchies: 23770663. It disappeared in the reduced hierarchies list, so it rather looks in your file like you performed this step. I put it back in the reduced hierarchies list and made the step explicit. So now I guess we should also figure out which criterion removed it in your file.

:)

Jen
[Quoted text hidden]

7662 parent taxon summary eats v2 (1).txt
4K

Eli Agbayani <eagbayani@eol.org> Fri, Sep 21, 2018 at 11:20 AM
To: Jen Hammock <jen.hammock@gmail.com>

Thanks for the sample Jen!

Let me check, yes I noticed that we lost 23770663 somewhere along the way. I will check where.
Thanks,
Eli

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Fri, Sep 21, 2018 at 12:16 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,

If without the new step of 15% calculation, I think I know how we lost this hierarchy in 'Reduced hierarchies'.

```
[23770663] => Array
    (
        [0] => 42430800
    )
```

We lost the ancestor 42430800 because it didn't meet the first rule when reducing hierarchies.
first rule: it exists in other hierarchies
second rule: its child is not the same throughout all hierarchies

The ancestor 42430800 didn't exist in other hierarchies.
So eventually we were left with just one root node (taxon ID=1)
Does that make sense?

Thanks,
Eli
[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Fri, Sep 21, 2018 at 12:52 PM
To: Eli Agbayani <eagbayani@eol.org>

Oh, that's interesting. This might be an order of operations thing, let me see...

OK, I think this comes down to at what point in the process the magic nodes get chopped off:

If 42430800 was lost, 23770663, as a tip, would remain, now a direct child of 2913056. AFTER all the other ancestor nodes like 42430800 were removed, THEN 2913056 and the other magic nodes should be removed. In this case, that would strand 23770663 as a tiny branch by itself, both root and tip- and then it would be removed because it appears in only 1/10 records. All tips should survive until the "roots <15% removal"

Does that help? Though it yields the same result in this case, I think it could be important when we do this with a set of hundreds of child taxa.

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Fri, Sep 21, 2018 at 1:08 PM

To: Jen Hammock <jen.hammock@gmail.com>

I see, I got the order of operations.
I will put in notes on the output to differentiate the different steps/sections.
Will remove shortcuts for now so we see how the values evolve, also for clarity.
And yes, I agree -- All tips should survive until the "roots <15% removal"
Thanks,
Eli

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Sat, Sep 22, 2018 at 11:15 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
Attached is the revised order of operations. Now the step to remove the 5 deletable taxa comes after the 'Reduced hierarchies' section.
Before I had it before the 'Reduced hierarchies' section.
Result should be clearer now, added some text comment on final step as well.
Thanks,
Eli

[Quoted text hidden]

7662 parent taxon summary eats v3.txt
5K

Jen Hammock <jen.hammock@gmail.com> Sun, Sep 23, 2018 at 10:07 AM
To: Eli Agbayani <eagbayani@eol.org>

Ah, that seems nice and readable;

I see 23770663 made it into the reduced hierarchies, and survived the trimming of the 5 magic taxa, but then it wasn't detected as a root? Unless I'm misreading, it still sounds like it was removed in some unknown manner before the trimming of the <15% roots. My understanding is that after all the assorted trimming, 23770663 should count as both a root and a tip.

Thanks!

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Sun, Sep 23, 2018 at 11:23 AM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
So if I have something like this:

Hierarchies after removal of the 5 deletable taxa:Array
(

```
    [46557930] => Array
        (
            [0] => 2774383
            [1] => 1
        )
    [207661] => Array
        (
            [0] => 2774383
            [1] => 1
        )
    [10459935] => Array
        (
            [0] => 1
        )
    [1905] => Array
        (
            [0] => 2774383
            [1] => 1
        )
    [23770663] => Array
        (
        )
    [695] => Array
        (
            [0] => 46557930
            [1] => 2774383
            [2] => 1
        )
    [46559166] => Array
        (
            [0] => 46557930
            [1] => 2774383
            [2] => 1
        )
    [166] => Array
        (
            [0] => 1
        )
)
```

I actually have two roots (1 & 23770663). And not just one (1). Is that it?
So when the step for the "roots < 15% removal" comes I have two roots to contend with.
Is that correct?

[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Sun, Sep 23, 2018 at 11:36 AM
To: Eli Agbayani <eagbayani@eol.org>

Yes, that's how I see it, anyway. If I'm missing something about 23770663, let me know, but in general, an isolated node with no ancestry should, I think, be both a root and a tip.

Thanks!

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Sun, Sep 23, 2018 at 1:04 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
Here it is, updated result.
Thanks for your patience.
Regards,
Eli

[Quoted text hidden]

7662 parent taxon summary eats v4.txt
6K

Jen Hammock <jen.hammock@gmail.com> Sun, Sep 23, 2018 at 3:33 PM
To: Eli Agbayani <eagbayani@eol.org>

Yup, I think we're on the same page now.

ooops! Except for one thing, I think. I was so concerned about the roots I think I missed it. OK:

when you construct the hierarchies, only using Landmark taxa, relying on that annotated DH file I showed you, I think it leaves something out: all taxa with rank=Family are landmark taxa, but I think that resource doesn't say so. Sorry!

This may or may not affect our example, but for the parent level taxon summary process, when constructing the (landmark taxa only) initial hierarchies, please include all family level ancestors, as well as the landmark taxa from that resource file.

Another thing I think I haven't asked you for yet: please construct the hierarchies the same way for the "regular" taxon summary process also (in addition to the parent-level one). Landmark ancestors only, but including family level ancestors.

And then I think this process might be ready. Thanks for *your* patience,

;)

Jen

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Sun, Sep 23, 2018 at 8:20 PM

To: Jen Hammock <jen.hammock@gmail.com>

No problem Jen. Here is updated result.
We now have slightly more ancestries in "Hierarchies of taxon values" section.
As expected since many (or maybe all) taxa in DH file is with taxonRank = family and Landmark is blank.
These were excluded before.

Anyway in our test case, the final result didn't change.
Thanks,
Eli

[Quoted text hidden]

7662 parent taxon summary eats v5.txt
6K

Jen Hammock <jen.hammock@gmail.com> Sun, Sep 23, 2018 at 8:43 PM
To: Eli Agbayani <eagbayani@eol.org>

Looks good! Yes, I expect this may have more effect on taxa where we have fewer records available.

I've lost track, but if all the processes are ready for the methods we've talked about, could you please make some resource files from the Carnivora sample file? Let's see, what questions does that bring up...

Which taxa: for the initial (non-parent) processes, let's include just taxa with rank=species. And for the parent processes, just the landmark taxa (including families, again).

Which predicates: for the initial processes, please include all the predicates listed in the predicates sheet for each method. For the parent methods, just the predicates listed for each in the parents sheet.

What do the resources look like? For REC and PRM records that existed already, I guess you'll just need the record identifier (I think that's eol_pk in the traits table) and the REC and PRM flags. That might as well be a separate file from the newly created records. The newly created PRM and REC records can be in regular DwC-A form, with an extra column (or two?) in the MoF and Assoc files for the REC and PRM flags.

Metadata for the newly created records: we'll probably go back and forth a few times with this. There are instructions in the simple answers doc, but they might be cryptic. Please send me any questions- or if it's easier, we can do the reverse engineering thing again; you send me sample skeleton records and I send them back with metadata.

And once I've seen a larger sample of the records, I might ask you to tweak the rules again, but probably it'll be superficial things like adding more magic deletable nodes or something...

And we will eventually come back to those other methods we haven't tackled yet, but those are less urgent.

Thanks!!

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Mon, Sep 24, 2018 at 6:38 AM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, let us do the reverse engineer.
Before we do the bulk taxa, let me figure out first what to do for a single taxon.

Attached is for method = basal value; predicate = habitat includes; page_id = 46559217.
Please tell me if you need other raw files.
I have questions inside file: 46559217_resource.txt

Other test cases to follow.
Thanks,
Eli

[Quoted text hidden]

46559217.zip
18K

Eli Agbayani <eagbayani@eol.org> Mon, Sep 24, 2018 at 7:02 AM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, another test case. For reverse engineer.
Method = taxon summary; page_id = 328598; predicate = "http://purl.obolibrary.org/obo/RO_0002470"; //eats

Sorry can you please give me an example of a resource file for this type of method.
Please tell me if you need other raw files.

Thanks,
Eli

[Quoted text hidden]

328598.zip
30K

Jen Hammock <jen.hammock@gmail.com> Mon, Sep 24, 2018 at 12:10 PM
To: Eli Agbayani <eagbayani@eol.org>

OK, here goes the Arctic fox example:

For selected values available in multiple records, let's do an order of precedence based on metadata, with an arbitrary tie-breaker (which you'll need in this case; sorry!). Please count the number of references attached to each candidate record, add 1 if there is a bibliographicCitation for the record, and choose the record with the highest number. In case of a tie, break it with any arbitrary method you like.

The new record for ENVO_01001125 is attached. Please excuse any format or field mapping errors. It's meant to be a valid DwC-A, except for one extra column in taxa and another extra column in

MeasurementOrFact and Associations. For the extra column in taxa, you might use whatever column label, URI, etc that you used when mapping EOL v2 taxa to specific taxon IDs in v3. I think a similar extra column appeared in the taxa file then? The extra column in MoF and Assoc is new, and I proposed a URI for it.

Anything else... Oh, the references- I found four records whose values were children of ENVO_01001125. They each had a reference, but two of them had the same reference string (full_reference in resource file. In the Carnivora data file these are found in the metadata table; the reference string is found in the literal column, wherever the predicate column contains "http://eol.org/schema/reference/referenceID".) Two of the four reference strings were identical, so after deduplicating, I ended up with three. If any of the records had also had a bibliographicCitation, I would have added the text of those as additional references.

Source: constructed from the taxonID and the predicate. It's not as specific as I'd like but it'll do. Right now it points to beta. Let's leave it like that for now, but try to remember that after launch, these resources should be rebuilt with the beta removed.

Occurrences file: this need never include anything but two columns of arbitrary identifiers to map the record to the taxonID in the taxa file.

How's that for a start?

:)

Jen
[Quoted text hidden]

sample_46559217_newrec.zip
6K

Eli Agbayani <eagbayani@eol.org> Mon, Sep 24, 2018 at 12:37 PM
To: Jen Hammock <jen.hammock@gmail.com>

Thanks Jen. Working on it now.
Just to make sure, this is the DwCA you sent me. I put it in a spreadsheet for easy eyeballing.
Will give you feedback soon.
Thanks,
Eli

[Quoted text hidden]

46559217.xlsx
488K

Jen Hammock <jen.hammock@gmail.com> Mon, Sep 24, 2018 at 12:47 PM
To: Eli Agbayani <eagbayani@eol.org>

Oh, you can do it backwards! Yes, that looks right. Here's the taxon summary for the Northern raccoon eating things. It works much the same, in the Assoc file rather than MoF, but since the data values are taxa, those are reflected in the taxa and occurrence files, along with that hungry raccoon.

I'll template from your spreadsheet next time :)

Jen
[Quoted text hidden]

sample_328598_newrec.zip
6K

Eli Agbayani <eagbayani@eol.org> Mon, Sep 24, 2018 at 1:19 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
Clarification please:

The selected record (ENVO_01001125) from 46559217_habitat.txt that is not found in traits.csv for this taxon, ended up in the MoF.
That is the only row in our MoF.

Will I also be adding the other value_uris in our selected records list in MoF? or somewhere?
e.g.
These six value_uris below are not represented in the DwCA. Except for the eol_pk used in getting the references for MoF.
Is that correct?

Page ID --- eol_pk --- Selected value_uri --- Label
46559217 --- R96-PK42940163 --- http://eol.org/schema/terms/temperate_grasslands_savannas_and_shrublands --- REP
46559217 --- R512-PK24322763 --- http://purl.obolibrary.org/obo/ENVO_00000078 --- REP
46559217 --- R512-PK24381251 --- http://purl.obolibrary.org/obo/ENVO_00000220 --- REP
46559217 --- R512-PK24428398 --- http://purl.obolibrary.org/obo/ENVO_00000446 --- REP
46559217 --- R512-PK24244192 --- http://purl.obolibrary.org/obo/ENVO_00000446 --- REP
46559217 --- R512-PK23617608 --- http://purl.obolibrary.org/obo/ENVO_00000572 --- REP
46559217 --- R512-PK24249316 --- http://purl.obolibrary.org/obo/ENVO_00002033 --- REP

Thanks,
Eli
[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Mon, Sep 24, 2018 at 1:47 PM
To: Eli Agbayani <eagbayani@eol.org>

Yeah, I thought about incorporating those into the resource file, but I'm not sure where they'd go. Some of this might be rejiggered once we've talked with Jeremy- he's going to need to ingest these into the database somehow. But for now, I'm thinking the existing records would be in separate files, something like the attached. I included page ID, though I don't think it's necessary. I'm pretty sure it could be just the record's EOL-pk and the label(s).

I'm not sure how many different resource files there should be. The existing records to flag for all predicates and all taxa, species and parent processes, might all go together in one file. The new records might go all together in one DwC archive. I think the only reason to split them apart would be if those files are too large to

be convenient to download. In which case you might do something like have a file for /present, a file for /habitat, and a file for everything else.

Does that help?

Jen
[Quoted text hidden]

existing records sample.csv
1K

Eli Agbayani <eagbayani@eol.org> Tue, Sep 25, 2018 at 1:24 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, favor please.
Can you please arrange the 'Associations' sheet. This is coming from your file (sample_328598_newrec.zip).
The Association extension has mis-aligned columns.
Thanks.

PS: shortly I'll submit the resource file: Basal values for all Carnivora dataset. Taxa with rank species.
[Quoted text hidden]

328598.xlsx
488K

Jen Hammock <jen.hammock@gmail.com> Tue, Sep 25, 2018 at 1:27 PM
To: Eli Agbayani <eagbayani@eol.org>

oops, how did that happen?
[Quoted text hidden]

328598.xlsx
496K

Eli Agbayani <eagbayani@eol.org> Tue, Sep 25, 2018 at 1:50 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
Attached is the resource file for all taxa with 'species' rank in Carnivora dataset (method = Basal values).
This is the DwCA, meaning new records.

I'm still working on the regular tab-delimited file for the 'existing records'. Forgot to implement the case for selected values available in multiple records:
- order of precedence based on metadata
- and if needed, an arbitrary tie-breaker as last solution.

Thanks,
Eli

[Quoted text hidden]

basal_values.tar.gz
38K

Jen Hammock <jen.hammock@gmail.com> Tue, Sep 25, 2018 at 4:36 PM
To: Eli Agbayani <eagbayani@eol.org>

Hi, Eli!

Thanks for the sample file! Now Katja and I have seen more output, we have a couple of tweaks to ask you for. (Surprise!)

This is a bit like the "delete magic taxa" in the taxon summary method. We'd like two delete steps for basal_values, each with its own hit list of magic deletable entities. Both can happen after the tree is assembled, before "step 1". I don't think the order of the two steps matters, but the order below should work.

DELETE ALONG WITH CHILDREN

- look for these nodes in the list of roots
- are there any other roots aside from the nodes in this list?
    - if not, do nothing
    - if so, delete all of these root nodes and all of their descendants- entire branches should be removed. I expect there may be weird cases, where a tip might be a child of one of these roots, and also of a root not on this list. So if this is cleaner: keep the nodes NOT on this list, and all their descendants. Discard everything else
- the list:
    - http://purl.obolibrary.org/obo/ENVO_00000094
    - http://purl.obolibrary.org/obo/ENVO_01000155
    - http://purl.obolibrary.org/obo/ENVO_00000002
    - http://purl.obolibrary.org/obo/ENVO_00000077

DELETE, BUT KEEP THE CHILDREN

- look for these nodes in the list of roots
- remove them. Their immediate children are now roots.
- the list:
    - http://purl.obolibrary.org/obo/ENVO_01001305
    - http://purl.obolibrary.org/obo/ENVO_00002030
    - http://purl.obolibrary.org/obo/ENVO_01000687

....and then proceed with step 1 as before :)

(it's OK if occasionally this leaves you with no records.)

Holler if anything is unclear. Thanks!

Jen

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Wed, Sep 26, 2018 at 5:44 AM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, attached in one of our test cases before.
Both old and new results for method: basal values
page_id: 7662 | predicate: [http://eol.org/schema/terms/Habitat]
The 'new' has the new delete steps. For review.

Please tell me if you need other test cases (old and new results) as guide.
*Attached test case didn't go to the 2nd deletion step.

Thanks,
Eli

[Quoted text hidden]

2 attachments

new_7662_basal_values_habitat.txt
11K

old_7662_basal_values_habitat.txt
11K

Jen Hammock <jen.hammock@gmail.com> Wed, Sep 26, 2018 at 9:40 AM
To: Eli Agbayani <eagbayani@eol.org>

Sorry, Eli- I'm not following this one. I can see a lot of tips were removed, but looking at your original list of roots, I don't think any of them were on the delete list, so I'm not sure anything should have changed in this case...

Ohh... I think you may have used my delete list for DELETE, BUT KEEP THE CHILDREN, for your DELETE ALONG WITH CHILDREN step.

But you may want to sanity check my deduction!

Jen
[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Wed, Sep 26, 2018 at 11:33 AM
To: Eli Agbayani <eagbayani@eol.org>

Ah, I think I might know what I said.

Trying those instructions again below, and sample file attached with just the tree construction and the new delete steps. Does that help?

Jen

> > > DELETE ALONG WITH CHILDREN
>
> - look for these nodes in the list of roots
> - are there any other roots aside from the nodes in this list?
>   - if not, do nothing
>   - if so, keep the root nodes that are NOT on this list, and all their descendants. Discard all other nodes
> - the list:
>   - http://purl.obolibrary.org/obo/ENVO_00000094
>   - http://purl.obolibrary.org/obo/ENVO_01000155
>   - http://purl.obolibrary.org/obo/ENVO_00000002
>   - http://purl.obolibrary.org/obo/ENVO_00000077
>
> DELETE, BUT KEEP THE CHILDREN
>
> - look for these nodes in the list of roots
> - remove them. Their immediate children are now roots.
> - the list:
>   - http://purl.obolibrary.org/obo/ENVO_01001305
>   - http://purl.obolibrary.org/obo/ENVO_00002030
>   - http://purl.obolibrary.org/obo/ENVO_01000687
>
> ....and then proceed with steps 1-4
>
> (it's OK if occasionally this leaves you with no records.)

new_7662_basal_values_habitat_delete_steps.txt
15K

Eli Agbayani <eagbayani@eol.org> Thu, Sep 27, 2018 at 1:54 AM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, thanks for the reverse engineer output. We're now getting the same results. Cool.

Attached is another test case where BOTH deletion steps exist.
method: Basal values
page_id: 328609 | predicate: [http://eol.org/schema/terms/Habitat]

If this is good/correct, then I can generate again the bulk Carnivora taxa.
Thanks,
Eli

[Quoted text hidden]

328609_basal_values_habitat.txt

Jen Hammock <jen.hammock@gmail.com> Thu, Sep 27, 2018 at 10:49 AM
To: Eli Agbayani <eagbayani@eol.org>

Hi, Eli! You've been busy :) OK, just starting on these.

I just looked at the newly added delete steps in the basal values method. In your example, I think everything is good except, in the delete-but-keep-children step, when those two roots are deleted, some other nodes should become roots; for instance, /ENVO_00000043, which I believe no longer has any parents in the hierarchy.

Everything else looks good!

Moving on to your next thread...
[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Thu, Sep 27, 2018 at 10:59 AM
To: Eli Agbayani <eagbayani@eol.org>

Yes, I made a mess of that, didn't I? Sorry, I'm not sure what happened. See if this doesn't make more sense...

Jen

On Thu, Sep 27, 2018 at 8:37 AM Eli Agbayani <eagbayani@eol.org> wrote:

> Hi Jen, I'm working on this one now. That is writing resource file for method: taxon summary.
> Can you please check if the Association sheet is correct?
> The 2nd row doesn't have:
> Occurrence ID Association Type Target Occurrence ID
>
> By the way this is the data behind this spreadsheet:
> Array
> (
> [page_id] => 328598
> [predicate] => http://purl.obolibrary.org/obo/RO_0002470
>
> [root] => 46557930
> [root label] => PRM
> [Selected] => Array
> (
> [0] => 46557930
> [1] => 207661
>)
>
> [Selected label] => REP
>)
>
> Thanks,

> Eli
> [Quoted text hidden]

Archive.zip
6K

Eli Agbayani <eagbayani@eol.org> Thu, Sep 27, 2018 at 10:12 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
So in step 2, I deleted these 6 pairs.

http://purl.obolibrary.org/obo/ENVO_00002030	http://purl.obolibrary.org/obo/ENVO_00000043
http://purl.obolibrary.org/obo/ENVO_00002030	http://purl.obolibrary.org/obo/ENVO_00000447
http://purl.obolibrary.org/obo/ENVO_01001305	http://purl.obolibrary.org/obo/ENVO_00000043
http://purl.obolibrary.org/obo/ENVO_01001305	http://purl.obolibrary.org/obo/ENVO_00000300
http://purl.obolibrary.org/obo/ENVO_01001305	http://purl.obolibrary.org/obo/ENVO_01000177
http://purl.obolibrary.org/obo/ENVO_01001305	http://purl.obolibrary.org/obo/ENVO_01000178

And supposedly all 6 children (5 deduplicated) will become roots.
But since all 5 also exist in the left side of the pairs I didn't make them roots.
Is my assumption wrong?
I kinda thought it was consistent with what you did with 7662 (see attached). Not all left-out children became roots.
Thanks,
Eli

> [Quoted text hidden]

new_7662_basal_values_habitat_delete_steps.txt
17K

Jen Hammock <jen.hammock@gmail.com> Thu, Sep 27, 2018 at 11:12 PM
To: Eli Agbayani <eagbayani@eol.org>

You're right; my 7662 and my 328609 aren't consistent. It is true that not all left out children become roots.
Please bear with me; I'll try to articulate the process I do want...

I *think* the trimmed hierarchy is correct in both examples. But what does the it mean?

- rows with just one node in them are only needed for orphans (nodes that don't appear anywhere else). These nodes are both tips and roots.
- tips are nodes that never appear on the left of a two node row
- roots are nodes that never appear on the right of a two node row
- nodes that appear both on the left of a two node row and on the right of a two node row are neither roots nor tips

So in 7662, when I removed 00002030 and 01001305, that left these children to account for (deduplicated):

http://purl.obolibrary.org/obo/ENVO_00000043
http://purl.obolibrary.org/obo/ENVO_00000447
http://purl.obolibrary.org/obo/ENVO_00000873
http://purl.obolibrary.org/obo/ENVO_00002010
http://purl.obolibrary.org/obo/ENVO_00002019
http://purl.obolibrary.org/obo/ENVO_01000020
http://purl.obolibrary.org/obo/ENVO_00000300

three got put into orphan/singlet rows, because they don't appear elsewhere in the hierarchy, so they're both roots and tips:

http://purl.obolibrary.org/obo/ENVO_00002010
http://purl.obolibrary.org/obo/ENVO_00002019
http://purl.obolibrary.org/obo/ENVO_01000020

http://purl.obolibrary.org/obo/ENVO_00000300 appears as a parent and a child elsewhere in the hierarchy, so it's neither a root nor a tip

The other three:

http://purl.obolibrary.org/obo/ENVO_00000043
http://purl.obolibrary.org/obo/ENVO_00000447
http://purl.obolibrary.org/obo/ENVO_00000873

appear elsewhere in the hierarchy, but never, I think, on the right of a two node row, so I *should* have flagged all three of them as roots. Ooops!

(But that's a good result. It means Carnivora occur in a wider range of habitats than we previously reported, including marine and freshwater, so we haven't left out the seals and otters this way.)

Sorry for my flaky reverse engineering. Does that help?

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Fri, Sep 28, 2018 at 2:21 PM
To: Jen Hammock <jen.hammock@gmail.com>

Thanks for the explanation Jen. Indeed that helps alot.
Please see attached for those 2 test cases we have.

I put a new section ----Diagnostics ----- just so to easily see where do nodes in question go.

Anyway please take note that I assumed all orphans/single rows also ends up in roots.

Thanks,
Eli

[Quoted text hidden]

---

2 attachments

7662_basal_values_habitat.txt
18K

328609_basal_values_habitat.txt
26K

Jen Hammock <jen.hammock@gmail.com> Sat, Sep 29, 2018 at 10:34 AM
To: Eli Agbayani <eagbayani@eol.org>

Thanks for the diagnostic section, Eli! Yes, those delete steps look good. Gosh, are we almost there? :)

Thanks for persisting!

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Sun, Sep 30, 2018 at 1:43 AM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, we now have the same DwCA for method: taxon summary. Thanks. See attached

Just one question: regarding label for the root in Associations. This was the instruction before after the tree is constructed for method: taxon summary:
- Select all immediate children of the root and label REP.
- Label the root PRM

But I see in your DwCA that the label for target root is "REP;PRM".
Which one to use?
Thanks,
Eli

PS: and yes, we're getting there. I will now continue on 'parent: basal values'. The one where I left off when we started writing resource files. Thanks.

[Quoted text hidden]

---

taxon_summary.tar.gz
2K

Jen Hammock <jen.hammock@gmail.com> Sun, Sep 30, 2018 at 10:22 AM
To: Eli Agbayani <eagbayani@eol.org>

Thanks for checking, Eli! I wasn't sure in general whether to have PRM records carry REP labels also, but for now let's say yes- in general, where I say to label something PRM, assume I also mean REP.

However, in this case (surprise!) one more tweak. Please make the PRM record, not the root, but the REP record that appears in the most hierarchies in the original list. I think we might end up doing that with all four applicable methods (basal values and taxon summary, regular and parents). Does that make sense?

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Mon, Oct 1, 2018 at 4:10 AM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
In the issue of selecting PRM (or PRM & REP) using the most no. of records it existed FROM THE ORIGINAL RECORDS.
Before we do it, just some clarifications.

Please see attached sample outputs for:
- method: basal values
- method: taxon summary
- method: parents - taxon summary

For method: basal values. At least in all our test cases, we get > 1 value. So no PRM there. We always get the REP values.
Question:
Are we now also going to compute for the PRM from among the REP values with the most no. of records FROM THE ORIGINAL RECORDS. Something we didn't do before.

For method: taxon summary (3 examples). Currently, we get as PRM the immediate children of root with the most no. of records it existed.
Anyway I've put in the result also just for comparison, the no. of records it existed FROM THE ORIGINAL RECORDS.
The same totals most of the time and for those slight difference, we still end up with the same PRM record.
Question:
Do we still want to shift to use the ORIGINAL RECORDS in getting the PRM?

For method: parents - taxon summary. All direct children of the remaining root are REP records, the one that appears in the most ancestries is the PRM.
Question:
I've not yet computed the totals if we're going to use the ORIGINAL RECORDS to pick the PRM.
If we do, it is doable but a bit of an organization since we are actually processing the children of the taxon in question.
I just want to confirm our decision before I do the computation.

Thanks,
Eli

[Quoted text hidden]

---

sample_outputs.txt
8K

Jen Hammock <jen.hammock@gmail.com> Mon, Oct 1, 2018 at 9:21 AM
To: Eli Agbayani <eagbayani@eol.org>

Thanks for checking, Eli!

For method: parents - taxon summary, keep doing what you're already doing, using the REP records of all the children.

For the taxon summary method- I'm sorry, I haven't been able to figure it out. I suspect the method you're using already is fine, but I can't tell which records are not included, eg: for /Habitat for 7662, which two records aren't included in the 73?

Whichever way we end up handling that method, I expect the same procedure should work on the basal values method.

Thanks for your patience!

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Mon, Oct 1, 2018 at 10:53 AM
To: Jen Hammock <jen.hammock@gmail.com>

Thanks for the answers Jen.
So status quo for now for:
- method: parents taxon summary
- method: taxon summary

My next to-do is to get the PRM for method: Basal values. Using the REP values, with the most no. of records from original records.
And we can just re-evaluate once we get more results, adjust if needed.

Thanks,
Eli

[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Mon, Oct 1, 2018 at 11:27 AM
To: Eli Agbayani <eagbayani@eol.org>

Sounds good!

Thanks,

Jen
[Quoted text hidden]