Eli Agbayani <eagbayani@eol.org>

task: summary data resources

41 messages

Eli Agbayani <eagbayani@eol.org> Mon, Sep 3, 2018 at 12:14 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
Help please. Using the spreadsheet, and the carnivora dataset.

I'm working on for example a single predicate = "http://eol.org/schema/terms/Present", which is under 'summary process' = "basal values".
Attached is a sample working file (sample.txt.zip) for this single predicate.

It has 3 sections:
1. Similar terms from [terms relationship files]:
2. Taxa (with ancestry) having data for predicate in question and similar terms:
3. Records from traits.csv having data for predicate in question and similar terms: (now with just a few columns)

Now looking at your instructions in worksheet "basal value". May I ask can you please manually generate the report you want, with maybe actual fields.
Or para-phrase your instructions (under 'prep' and 'steps') now with the sample data I've provided.

Please tell me if you still need other raw information/report.

The exercise is doable, I'm just still grasping the steps at the moment.
Thanks,
Eli

sample.txt.zip
143K

Jen Hammock <jen.hammock@gmail.com> Mon, Sep 3, 2018 at 2:07 PM
To: Eli Agbayani <eagbayani@eol.org>

Hang in there, Eli- I think this is the most complex of the methods, but it'll be important for /present and /habitat data, where there are the most records, so it'll be the most helpful method of the bunch. But yes, it's convoluted. I've attempted to work with a sample (just one taxon for /Present) to clarify somewhat. Some items to note:

- You'll want to keep the record identifier (eol_pk) because some of this task is identifying existing records, to apply flags to them
- You don't need the taxon ancestry, or the predicate ancestry, for the basal value method. The terms relationships files are only needed for the values ancestry

- You can discard literal values
- The trickiest part, I think, will be constructing the shared values ancestry tree. Each value may have several parents (to their left, in the parent-child and preferred-synonym files) and those parents can have parents of their own, creating multiple lines of ancestry for each value we started with. However, we are only interested in ancestors that connect one of the values we started with to others from the same set: to make the simplest possible tree that connects all five, or thirty, values that we have for one taxon, for one predicate. This is the tree that is used in steps 1-4.
- All that being said, I think I'll need to make you a curated relationships file for /present and another for /habitat. I think I'll want to pick and choose some terms. I'll point you to the new files when they're ready, but feel free to carry on with the terms relationships files in the meantime.

Let me know if this helps with getting to the "selected values". I'll bet I still have some explaining to do about flagging or creating records for those values once you have them.

Jen
[Quoted text hidden]

---

sample.txt
5K

Eli Agbayani <eagbayani@eol.org> Tue, Sep 4, 2018 at 11:02 AM
To: Jen Hammock <jen.hammock@gmail.com>, "Hammock, Jennifer" <HammockJ@si.edu>

Hi Jen, thanks for the sample.txt and further explanations.

Please bear with me. Question please:
In your 4th bullet point you mentioned: "**However, we are only interested in ancestors that connect one of the values we started with to others from the same set**".
For example in your sample.txt under section: **Shared Values Ancestry from [terms relationship files]:**
In the first row:
http://www.geonames.org/6255151 http://www.marineregions.org/gazetteer.php?p=details&id=australia

How was **http://www.marineregions.org/gazetteer.php?p=details&id=australia** got a parent that is **http://www.geonames.org/6255151** ?

I made some computations and got these:

term in question: [http://www.marineregions.org/gazetteer.php?p=details&id=australia]:
There are 2 preferred term(s) :
[0] => http://www.geonames.org/2077456
[1] => http://www.marineregions.org/mrgid/australia

parent(s) of http://www.geonames.org/2077456:
[0] => https://www.wikidata.org/entity/Q186198
[1] => https://www.wikidata.org/entity/Q41228
[2] => http://www.geonames.org/6255151
[3] => http://eol.org/schema/terms/Australasia

parent(s) of http://www.marineregions.org/mrgid/australia: -- NO parent

What is the criteria to pick **http://www.geonames.org/6255151** among the 4 parents of

http://www.geonames.org/2077456 ?

Thanks,
Eli

[Quoted text hidden]

Hammock, Jennifer <HammockJ@si.edu> Tue, Sep 4, 2018 at 11:26 AM
To: Eli Agbayani <eagbayani@eol.org>

Good question!

I'm afraid the answer may be computationally expensive. The criterion for choosing http://www.geonames.org/6255151 is that it is also an ancestor of other values in the original list, specifically

http://www.marineregions.org/gazetteer.php?p=details&id=4366
http://www.marineregions.org/gazetteer.php?p=details&id=4364
http://www.geonames.org/2186224
http://www.marineregions.org/gazetteer.php?p=details&id=4276
http://www.marineregions.org/gazetteer.php?p=details&id=4365

If this helps: I recorded no ancestors for http://www.marineregions.org/gazetteer.php?p=details&id=1904 because it didn't share any ancestors with the other values in the list. All of this with the caveat that I did this by hand and might have missed something...

Does that help?

Jen

From: Eli Agbayani [eagbayani@eol.org]
Sent: Tuesday, September 04, 2018 11:02 AM
To: Jen Hammock; Hammock, Jennifer
Subject: Re: task: summary data resources

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Tue, Sep 4, 2018 at 11:55 AM
To: "Hammock, Jennifer" <HammockJ@si.edu>

Hi Jen,
I see, makes sense.
I will put that criteria on script and see if I get the same parents as you have.
Will share when done.

Yes, I must get the rules/criteria right because this is the first step in the process, and like you said the trickiest.
Thanks,
Eli

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Tue, Sep 4, 2018 at 2:26 PM
To: "Hammock, Jennifer" <HammockJ@si.edu>

Hi Jen, I got same parents 9 out of 12. I put asterisk ** those three where we have different parents.

CHOSEN PARENT: http://www.geonames.org/6255151 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=australia]:
CHOSEN PARENT: http://www.geonames.org/6255151 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=4366]:
CHOSEN PARENT: http://www.geonames.org/6255151 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=4364]:
CHOSEN PARENT~: http://www.geonames.org/6255151 ---> [http://www.geonames.org/2186224]:
**CHOSEN PARENT~: http://www.marineregions.org/mrgid/1902 ---> [http://www.geonames.org/3370751]:
**CHOSEN PARENT: https://www.wikidata.org/entity/Q41228 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=1914]:
**CHOSEN PARENT: https://www.wikidata.org/entity/Q186198 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=1904]:
CHOSEN PARENT: https://www.wikidata.org/entity/Q41228 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=1910]:
CHOSEN PARENT: http://www.geonames.org/6255151 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=4276]:
CHOSEN PARENT: http://www.geonames.org/6255151 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=4365]:
CHOSEN PARENT~: https://www.wikidata.org/entity/Q41228 ---> [http://www.geonames.org/953987]:
CHOSEN PARENT~: https://www.wikidata.org/entity/Q41228 ---> [http://www.marineregions.org/mrgid/1914]:

I also attached (choosing_parent.txt) for calculation for each of the 12 terms.
This is also the ranking I computed to choose the parent:
[http://www.geonames.org/6255151] => 6
[https://www.wikidata.org/entity/Q41228] => 5
[http://www.marineregions.org/mrgid/14289] => 4
[https://www.wikidata.org/entity/Q186198] => 4
[http://www.marineregions.org/mrgid/1903] => 3
[http://www.marineregions.org/mrgid/1902] => 2
[http://www.marineregions.org/mrgid/1910] => 2
[http://eol.org/schema/terms/Australasia] => 2
[http://eol.org/schema/terms/Afrotropical] => 1
[http://www.geonames.org/6255146] => 1
[http://www.geonames.org/3358844] => 1

[http://www.marineregions.org/mrgid/1914] => 1
[http://www.marineregions.org/mrgid/1904] => 1

*Please advise on how are we going to proceed with the 3 discrepancy.

*Another question please: In your sample.txt, in section: Shared Values Ancestry from [terms relationship files]
You've added a 12th row, was that intentional?
12th row: https://www.wikidata.org/entity/Q41228 http://www.marineregions.org/mrgid/1914

Thanks that's it for now.
Eli

[Quoted text hidden]

---

choosing_parent.txt
8K

Hammock, Jennifer <HammockJ@si.edu> Tue, Sep 4, 2018 at 3:25 PM
To: Eli Agbayani <eagbayani@eol.org>

Cool!

Okay, let's see. I think this will be incomplete, because I'm not sure how to construct your whole tree from the information below.

For our first and discrepancy,

http://www.marineregions.org/mrgid/1902 is an ancestor of http://www.geonames.org/3370751, but so is http://www.marineregions.org/mrgid/1914, and the latter is a shared parent with http://www.marineregions.org/gazetteer.php?p=details&id=1914. The latter relationship comes from the preferred-synonym file.

This is related to the second discrepancy. You chose https://www.wikidata.org/entity/Q41228 as the parent of http://www.marineregions.org/gazetteer.php?p=details&id=1914 instead of making it a grandparent, via http://www.marineregions.org/mrgid/1914, as I did.

In the third discrepancy, I think you found:

http://www.marineregions.org/gazetteer.php?p=details&id=1904 shares an ancestor with several other values: https://www.wikidata.org/entity/Q186198

which I think is also perfectly correct, and I missed it.

And the added row at the bottom: yes, my process was to add each new shared parent node I discovered (eg: http://www.marineregions.org/mrgid/1914) as a child looking for a parent. I only ended up with that one additional row, because all the other parent nodes turned out to be roots, that could not be connected to each other through a shared parent.

Is your selection among multiple possible parents based on a ranking of how many descendants could be

matched to each? I was not taking that into account. My goal was to connect all possible pairs of nodes. I think this may produce children with multiple lines of ancestry, which I think is ok.

Let me know what you think,

Jen

From: Eli Agbayani [eagbayani@eol.org]
Sent: Tuesday, September 04, 2018 2:26 PM
To: Hammock, Jennifer
[Quoted text hidden]

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Wed, Sep 5, 2018 at 11:47 AM
To: "Hammock, Jennifer" <HammockJ@si.edu>

Hi Jen,
I may need a fresh start tomorrow but here is what I have now.

If I DON'T fix the 2nd discrepancy (row 6 highlighted). It will be the only discrepancy we'll have.
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.marineregions.org/mrgid/1914
https://www.wikidata.org/entity/Q41228
https://www.wikidata.org/entity/Q186198
https://www.wikidata.org/entity/Q41228
http://www.geonames.org/6255151
http://www.geonames.org/6255151
https://www.wikidata.org/entity/Q41228
https://www.wikidata.org/entity/Q41228

If I fix the 2nd discrepancy, it will actually mess the other choices we previously gotten right (rows 7 & 8 highlighted):
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.marineregions.org/mrgid/1914
http://www.marineregions.org/mrgid/1914 ---> (discrepancy fixed; previously https://www.wikidata.org/entity/Q41228)
http://www.marineregions.org/mrgid/1904 ---> (previously https://www.wikidata.org/entity/Q186198)
http://www.marineregions.org/mrgid/1910 ---> (previously https://www.wikidata.org/entity/Q41228)
http://www.geonames.org/6255151
http://www.geonames.org/6255151
https://www.wikidata.org/entity/Q41228

https://www.wikidata.org/entity/Q41228

So there seem to be a contradiction in rules between cases. What do you think?
Thanks,
Eli

[Quoted text hidden]

Hammock, Jennifer <HammockJ@si.edu> Wed, Sep 5, 2018 at 11:57 AM
To: Eli Agbayani <eagbayani@eol.org>

Let's see... The second method looks right to me, with the proviso that the newly added nodes

http://www.geonames.org/6255151
http://www.marineregions.org/mrgid/1914
http://www.marineregions.org/mrgid/1904
http://www.marineregions.org/mrgid/1910
https://www.wikidata.org/entity/Q41228

be added to the collection of nodes to be checked for shared parents. This should give you https://www.wikidata.org/entity/Q41228 (again) as a parent for /1910 and /1914 (as well as already being a parent for two of the original nodes.)

Does that help?

Jen

From: Eli Agbayani [eagbayani@eol.org]
Sent: Wednesday, September 05, 2018 11:47 AM
[Quoted text hidden]

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Wed, Sep 5, 2018 at 12:17 PM
To: "Hammock, Jennifer" <HammockJ@si.edu>

Yes Jen, that makes sense, that's correct Glad you chose option 2.
Will now proceed.
Thanks,
Eli

[Quoted text hidden]

Hammock, Jennifer <HammockJ@si.edu> Wed, Sep 5, 2018 at 12:18 PM
To: Eli Agbayani <eagbayani@eol.org>

Cool, thanks! If any other complications arise, we can iterate more. I'm not sure I've explored all the possibilities...

Jen

From: Eli Agbayani [eagbayani@eol.org]
Sent: Wednesday, September 05, 2018 12:17 PM
[Quoted text hidden]

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Wed, Sep 5, 2018 at 2:03 PM
To: "Hammock, Jennifer" <HammockJ@si.edu>

Shared values ancestry tree computed:

http://www.geonames.org/6255151 http://www.marineregions.org/gazetteer.php?p=details&id=australia
http://www.geonames.org/6255151 http://www.marineregions.org/gazetteer.php?p=details&id=4366
http://www.geonames.org/6255151 http://www.marineregions.org/gazetteer.php?p=details&id=4364
http://www.geonames.org/6255151 http://www.geonames.org/2186224
http://www.marineregions.org/mrgid/1914 http://www.geonames.org/3370751
http://www.marineregions.org/mrgid/1914 http://www.marineregions.org/gazetteer.php?p=details&id=1914
http://www.marineregions.org/mrgid/1904 http://www.marineregions.org/gazetteer.php?p=details&id=1904
http://www.marineregions.org/mrgid/1910 http://www.marineregions.org/gazetteer.php?p=details&id=1910
http://www.geonames.org/6255151 http://www.marineregions.org/gazetteer.php?p=details&id=4276
http://www.geonames.org/6255151 http://www.marineregions.org/gazetteer.php?p=details&id=4365
https://www.wikidata.org/entity/Q41228 http://www.geonames.org/953987

new nodes to be added:
http://www.marineregions.org/mrgid/1902 http://www.marineregions.org/mrgid/1914
https://www.wikidata.org/entity/Q41228 http://www.marineregions.org/mrgid/1914
https://www.wikidata.org/entity/Q186198 http://www.marineregions.org/mrgid/1904
http://www.marineregions.org/mrgid/1903 http://www.marineregions.org/mrgid/1910
https://www.wikidata.org/entity/Q41228 http://www.marineregions.org/mrgid/1910

Thanks,
Eli

[Quoted text hidden]

Hammock, Jennifer <HammockJ@si.edu> Wed, Sep 5, 2018 at 2:04 PM
To: Eli Agbayani <eagbayani@eol.org>

Looks good!

Thanks for your patience,

Jen

From: Eli Agbayani [eagbayani@eol.org]
Sent: Wednesday, September 05, 2018 2:04 PM
[Quoted text hidden]

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Thu, Sep 6, 2018 at 6:08 AM
To: "Hammock, Jennifer" <HammockJ@si.edu>

Hi Jen,
For review please: files in attached archive.zip
page_id: 7662 | predicate: [http://eol.org/schema/terms/Habitat]
page_id: 46559197 | predicate: [http://eol.org/schema/terms/Present]
page_id: 46559217 | predicate: [http://eol.org/schema/terms/Present]

Hopefully I got step 1 right. I may have done the opposite though.
If you can please proceed with the steps for each.

Starting at Step 1, at this point, I'm now working on the initial shared values ancestry tree PLUS the added new nodes = COMBINED_ROWS. Is that right?
Sorry, more naive questions below.

Step 1
"xxx tips is >5, so find set 1 (parents except where they are roots)"
Question:
1. when you say 'parents', do you mean the one on the left?
If yes, so here I will pick from COMBINED_ROWS those rows where left side is not a root.
2. when you say roots, meaning it doesn't have a parent anymore?

Step 2
"yyy nodes is >4, find set 2"
"all parents are roots now, so steps 2 and 3 return the same set"
Question: so here I will pick from COMBINED_ROWS rows where left side are roots?

Step 4
"still >4 nodes, so select all roots"
Question: does this mean select all left side of COMBINED_ROWS where it is a root?

Thanks for the patience Jen.
Regards,
Eli

[Quoted text hidden]

Archive.zip
29K

Hammock, Jennifer <HammockJ@si.edu> Thu, Sep 6, 2018 at 9:55 AM

To: Eli Agbayani <eagbayani@eol.org>

Bother. The Smithsonian stripped your attachment. Could you send to jen.hammock@gmail.com, please?

Thanks!

From: Eli Agbayani [eagbayani@eol.org]
Sent: Thursday, September 06, 2018 6:08 AM
To: Hammock, Jennifer
Subject: Re: task: summary data resources

Attachment removed due to policy violation: archive.zip

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Thu, Sep 6, 2018 at 10:24 AM
To: Jen Hammock <jen.hammock@gmail.com>

Here it is Jen. Thanks.
[Quoted text hidden]

Archive.zip
29K

Eli Agbayani <eagbayani@eol.org> Thu, Sep 6, 2018 at 10:24 AM
To: "Hammock, Jennifer" <HammockJ@si.edu>

Ok, sending now. Thanks.
[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Thu, Sep 6, 2018 at 11:28 AM
To: Eli Agbayani <eagbayani@eol.org>

Hang in there, Eli- I'm still trying to make sure I understand your terminology, but I think we're getting pretty close.

I'm looking at 7662_result. It's a good example because it has a lot of values in the original list.

I think the construction of the tree is close, but a few cases are not as expected.

in the new nodes section I see a few nodes

http://purl.obolibrary.org/obo/ENVO_00000062
http://purl.obolibrary.org/obo/ENVO_00000002
http://purl.obolibrary.org/obo/ENVO_00000109

that I think are ancestors of only one of the original nodes. That should mean they can be discarded, so for instance, http://purl.obolibrary.org/obo/ENVO_00000856 will have no ancestors in the tree, and be "orphaned" so to speak.

There are also a couple of other parent nodes in the initial shared values ancestry tree

http://purl.obolibrary.org/obo/ENVO_01000155
http://eol.org/schema/terms/tropicalOrSubtropical

that are listed in the root nodes section. These also, being the parent of only one original node, can be discarded. So their children,

http://purl.obolibrary.org/obo/ENVO_00002033
http://purl.obolibrary.org/obo/ENVO_01000204

can also be orphaned. All three orphans should be included as roots, as well as the other orphans in the initial shared values ancestry tree,

http://purl.obolibrary.org/obo/ENVO_01000206
http://purl.obolibrary.org/obo/ENVO_00000463
http://purl.obolibrary.org/obo/ENVO_00002009
http://purl.obolibrary.org/obo/ENVO_00000446
http://purl.obolibrary.org/obo/ENVO_00000144

You have two other root nodes listed, both of which I *did* expect

http://purl.obolibrary.org/obo/ENVO_00002030
http://purl.obolibrary.org/obo/ENVO_00000446

I would also expect one more root node,

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

because it is both a parent of http://purl.obolibrary.org/obo/ENVO_00000033, and a grandparent of several other original values via http://purl.obolibrary.org/obo/ENVO_01000253

I'll go on to look at step 1 now.

We're getting there!

Jen
[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Thu, Sep 6, 2018 at 11:36 AM
To: Eli Agbayani <eagbayani@eol.org>

still on the tree-making step, I think the attached is the "complete tree" to use beginning in step 1, from 7662_result
[Quoted text hidden]

completeTree.csv
5K

Jen Hammock <jen.hammock@gmail.com> Thu, Sep 6, 2018 at 12:01 PM

To: Eli Agbayani <eagbayani@eol.org>

OK, take your time on the tree assembling, but here are some clarifications for the later steps. I'll try to put together the rest of the sample process for 7662 next.

:)

Jen

Step 1
"xxx tips is >5, so find set 1 (parents except where they are roots)"
Question:
1. when you say 'parents', do you mean the one on the left?
If yes, so here I will pick from COMBINED_ROWS those rows where left side is not a root.

So, first you must identify the tips- any values that don't appear in the left column. The parents, for step one, will be the values to the left of the tip values. This is because of this i
2. when you say roots, meaning it doesn't have a parent anymore?
Yes, roots are any value that doesn't have a parent

Step 2
"yyy nodes is >4, find set 2"
"all parents are roots now, so steps 2 and 3 return the same set"
Question: so here I will pick from COMBINED_ROWS rows where left side are roots?
that bit of narrative, "all parents are roots now" only applies to the dataset example I was working on. If ever you have a set 1, or a set 2 or 3, where all parents of the values in the set are roots, the next step will return the same set as this step. This is because of this bit in those steps:

"... unless the parent is a root ancestor, in which case keep ..."

Step 4
"still >4 nodes, so select all roots"
Question: does this mean select all left side of COMBINED_ROWS where it is a root?
Yes, that sounds right. So in the 7662 example, I believe that would be something like this:

http://purl.obolibrary.org/obo/ENVO_01000206
http://purl.obolibrary.org/obo/ENVO_00000463
http://purl.obolibrary.org/obo/ENVO_00002009
http://purl.obolibrary.org/obo/ENVO_00000446
http://purl.obolibrary.org/obo/ENVO_00000144
http://purl.obolibrary.org/obo/ENVO_00002033
http://purl.obolibrary.org/obo/ENVO_01000204
http://purl.obolibrary.org/obo/ENVO_00000856
http://purl.obolibrary.org/obo/ENVO_00002030
http://purl.obolibrary.org/obo/ENVO_00000446
http://purl.obolibrary.org/obo/ENVO_00000873

[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Thu, Sep 6, 2018 at 12:38 PM
To: Eli Agbayani <eagbayani@eol.org>

OK, this is the last of what I have for now. I think the attached is close to the expected behavior for taxon 7662, for /habitat

I also realized my description so far doesn't account for orphan nodes, as I thought it did.

I think what is needed, in steps 1-4, is that where I said *unless the parent is a root ancestor, in which case...* that should be *unless the node or its parent is a root ancestor, in which case...*

I'll stop now until I hear from you :)

Jen
[Quoted text hidden]

7662_steps.csv
9K

Eli Agbayani <eagbayani@eol.org> Sat, Sep 8, 2018 at 11:22 AM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, quick question:
You mentioned:
There are also a couple of other parent nodes in the initial shared values ancestry tree

http://purl.obolibrary.org/obo/ENVO_01000155
http://eol.org/schema/terms/tropicalOrSubtropical

that are listed in the root nodes section. These also, being the parent of only one original node, can be discarded. So their children,

http://purl.obolibrary.org/obo/ENVO_00002033
http://purl.obolibrary.org/obo/ENVO_01000204

can also be orphaned.

Another case of this is this parent node: http://purl.obolibrary.org/obo/ENVO_00000300.
Which is a parent of only one original node: http://purl.obolibrary.org/obo/ENVO_00000301
Will the parent be discarded and child orphaned?

Or NOT, since the parent node is also a child of two different parents in the tree:

http://purl.obolibrary.org/obo/ENVO_00000446 http://purl.obolibrary.org/obo/ENVO_00000300
http://purl.obolibrary.org/obo/ENVO_01001305 http://purl.obolibrary.org/obo/ENVO_00000300

Thanks,
Eli

[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Sat, Sep 8, 2018 at 11:35 AM
To: Eli Agbayani <eagbayani@eol.org>

Good question! This is one of the downstream fiddly bits.

http://purl.obolibrary.org/obo/ENVO_00000301 should not be orphaned, because it has those ancestors, http://purl.obolibrary.org/obo/ENVO_00000446 and http://purl.obolibrary.org/obo/ENVO_01001305 in common with other original nodes. But if http://purl.obolibrary.org/obo/ENVO_00000300 doesn't provide another bifurcation, you can "cut out the middle man", removing http://purl.obolibrary.org/obo/ENVO_00000300 and making http://purl.obolibrary.org/obo/ENVO_00000301 a direct child of http://purl.obolibrary.org/obo/ENVO_00000446 and http://purl.obolibrary.org/obo/ENVO_01001305 .

This is the bit about making it the simplest possible tree that makes all the connections. if A->B->C and B performs no other function, B can be removed, leaving A->C

Does that make sense?

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Sat, Sep 8, 2018 at 4:20 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, before I proceed to Step 1. Attached is your and my 'complete tree'.
By the way I sorted them so easier to compare.

I only have 2 rows more than you, the rest is identical.

If you'll ask why I have an extra orphaned http://purl.obolibrary.org/obo/ENVO_00000300.
It is because in one of the 'new nodes' section I have an entry like so:
http://purl.obolibrary.org/obo/ENVO_00000109 http://purl.obolibrary.org/obo/ENVO_00000300

Where ENVO_00000109 is a parent of only one original node.
So the parent was discarded and ENVO_00000300 was orphaned.

Does that make sense?

Thanks,
Eli

[Quoted text hidden]

---

2 attachments

eli_complete_tree.txt
4K

jen_complete_tree.txt
4K

Eli Agbayani <eagbayani@eol.org> Sat, Sep 8, 2018 at 4:27 PM
To: Eli Agbayani <eagbayani@eol.org>

---------- Forwarded message ----------
From: **Jen Hammock** <jen.hammock@gmail.com>
Date: Thu, Sep 6, 2018 at 12:01 PM
Subject: Re: task: summary data resources
To: Eli Agbayani <eagbayani@eol.org>

OK, take your time on the tree assembling, but here are some clarifications for the later steps. I'll try to put together the rest of the sample process for 7662 next.

:)

Jen

Step 1
"xxx tips is >5, so find set 1 (parents except where they are roots)"
Question:
1. when you say 'parents', do you mean the one on the left?
If yes, so here I will pick from COMBINED_ROWS those rows where left side is not a root.

So, first you must identify the tips- any values that don't appear in the left column. The parents, for step one, will be the values to the left of the tip values. This is because of this i
2. when you say roots, meaning it doesn't have a parent anymore?
Yes, roots are any value that doesn't have a parent

Step 2
"yyy nodes is >4, find set 2"
"all parents are roots now, so steps 2 and 3 return the same set"
Question: so here I will pick from COMBINED_ROWS rows where left side are roots?
that bit of narrative, "all parents are roots now" only applies to the dataset example I was working on. If ever you have a set 1, or a set 2 or 3, where all parents of the values in the set are roots, the next step will return the same set as this step. This is because of this bit in those steps:

"... unless the parent is a root ancestor, in which case keep ..."
Correction made by Jen: ***I think what is needed, in steps 1-4, is that where I said*** ***unless the parent is a root ancestor, in which case... that should be*** ***unless the node or***

its parent is a root ancestor, in which case...

Step 4
"still >4 nodes, so select all roots"
Question: does this mean select all left side of COMBINED_ROWS where it is a root?
Yes, that sounds right. So in the 7662 example, I believe that would be something like this:

http://purl.obolibrary.org/obo/ENVO_01000206
http://purl.obolibrary.org/obo/ENVO_00000463
http://purl.obolibrary.org/obo/ENVO_00002009
http://purl.obolibrary.org/obo/ENVO_00000446
http://purl.obolibrary.org/obo/ENVO_00000144
http://purl.obolibrary.org/obo/ENVO_00002033
http://purl.obolibrary.org/obo/ENVO_01000204
http://purl.obolibrary.org/obo/ENVO_00000856
http://purl.obolibrary.org/obo/ENVO_00002030
http://purl.obolibrary.org/obo/ENVO_00000446
http://purl.obolibrary.org/obo/ENVO_00000873

Jen Hammock <jen.hammock@gmail.com> Sat, Sep 8, 2018 at 8:54 PM
To: Eli Agbayani <eagbayani@eol.org>

Let me see...

So it looks like 446 appears as both an orphan and an ancestor in your tree? The way I was thinking of documenting, it wouldn't need to be listed as an orphan if it also appears in any relationship pair.

As for 300... Ah, it's an original value, isn't it? In that case (my description above about cutting out middlemen notwithstanding) it should stay; an original node is not a middleman. But I don't think it should be an orphan. 109 is only a parent of one thing, but *it's still in the tree*, because it was an original node. So how should I have described that requirement...

"Ancestors can be removed if they are parents of only one node and are not original nodes" perhaps?

So there should be another record of 300, as a child of 109, (which I missed in my tree). And again, no node need be listed among the orphans if it appears elsewhere (and 300 appears with 3 different parents).

I have a feeling this process may be more complex than we need for the summary records task, but we'll be able to use these trees for a bunch of other things later, so I'm pretty sure it'll be worth the trouble :)

And, before I forget, I finally finished those new relationship files I threatened you with, one for /habitat and one for /present:

https://opendata.eol.org/dataset/terms-relationships/resource/c5ff5c62-a2ef-44be-9f59-88cd99bc8af2
https://opendata.eol.org/dataset/terms-relationships/resource/e1dcb51b-9a03-4069-b5bf-e18b6bc15798

Thanks!!

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Sun, Sep 9, 2018 at 3:27 AM
To: Jen Hammock <jen.hammock@gmail.com>

Okay Jen, so I will no longer use this parent - child, I used before:
https://opendata.eol.org/dataset/terms-relationships/resource/f8036c30-f4ab-4796-8705-f3ccd20eb7e9

And will just use these 2:

https://opendata.eol.org/dataset/terms-relationships/resource/c5ff5c62-a2ef-44be-9f59-88cd99bc8af2
https://opendata.eol.org/dataset/terms-relationships/resource/e1dcb51b-9a03-4069-b5bf-e18b6bc15798

specifically for /habitat and /present respectively.

Is that correct?
Thanks.

[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Sun, Sep 9, 2018 at 10:03 AM
To: Eli Agbayani <eagbayani@eol.org>

Yes, that's it. Thanks!

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Sun, Sep 9, 2018 at 5:11 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
Option 1: If we do the new rule: **node wouldn't need to be listed as an orphan if it also appears in any relationship pair.**
And only this rule, I will arrive at an identical completed tree as you have. Exactly the same as yours (jen_complete_tree.txt).

Option 2: If we do rule in Option 1 and the other rule:
"**Ancestors can be removed if they are parents of only one node BUT that node must NOT be an original node**"

If we do this, please see attached new completed tree (new_completed_tree.txt).

What do you think?
Thanks,
Eli

[Quoted text hidden]

2 attachments

new_complete_tree.txt
4K

jen_complete_tree.txt
4K

Jen Hammock <jen.hammock@gmail.com> Sun, Sep 9, 2018 at 5:34 PM
To: Eli Agbayani <eagbayani@eol.org>

Let's see... I've missed something. Oh!

The additional rule should not be

"Ancestors can be removed if they are parents of only one node BUT that node must NOT be an original node"

but

"Ancestors can be removed if they are parents of only one node BUT *the ancestor* must NOT be an original node"

Does that help?

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Mon, Sep 10, 2018 at 1:32 AM
To: Jen Hammock <jen.hammock@gmail.com>

Yesss, that certainly helped. Did the trick, thanks!

I now have identical tree as you have.
Except for just one additional row which we expected:
http://purl.obolibrary.org/obo/ENVO_00000109 http://purl.obolibrary.org/obo/ENVO_00000300
Attached your original tree and our latest tree with both rules.

I'm now moving along to the different steps. Will keep you posted.
Thanks,
Eli

[Quoted text hidden]

2 attachments

latest_complete_tree.txt
4K

jen_complete_tree.txt
4K

Eli Agbayani <eagbayani@eol.org> Mon, Sep 10, 2018 at 3:41 AM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, your instruction here for Step 1 seems cut.
You were saying...

"This is because of this i..."

[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Mon, Sep 10, 2018 at 4:54 AM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen, if you agree with our latest_complete_tree.txt.
May I request, can you please generate the updated set 1,2,3 & 4.
So I can compare when I get your complete instructions for Step 1.

It would be nice if format will be:
- complete tree
- all roots
- tips
- set 1
- set 2
- set 3
- set 4 OR all roots

Thanks,
Eli

[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Mon, Sep 10, 2018 at 9:26 AM
To: Eli Agbayani <eagbayani@eol.org>

Oops... I can't remember what I was trying to explain there. Looking at it now, "identify the tips and select their parents" seems to cover it. Checking your next message now...
[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Mon, Sep 10, 2018 at 10:45 AM
To: Eli Agbayani <eagbayani@eol.org>

OK, I'm pretty sure this is it- but if something doesn't match, it could still be Jen Error :)

[Quoted text hidden]

complete_tree_ steps_ 1-4.txt
12K

Eli Agbayani <eagbayani@eol.org> Mon, Sep 10, 2018 at 12:23 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
Thanks for the complete report.
I'm now able to generate correct lists for:

- complete tree
- roots
- tips

But I can't seem to get the correct list for Set 1.
Sorry, when I got the correct complete tree and roots and tips; I though the Set lists will be a breeze, but I can't seem to pass Set 1.
Can you please once again, state the step by step how did you arrive at your list for Set 1.

Thanks,
Eli

[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Mon, Sep 10, 2018 at 12:44 PM
To: Eli Agbayani <eagbayani@eol.org>

Sure- I think I've changed step 1 out from under you once or twice, so that's not too surprising. My current understanding, with which I constructed that set 1, is this:

- find all tips
- find all nodes that are parents of tips
- in each case, check whether either the tip or the parent is a root
    - if either the tip or the parent is a root, put the tip in set 1
    - if neither the tip nor the parent is a root, put the parent in set 1
- (deduplicate set 1)

hope that helps!
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Mon, Sep 10, 2018 at 2:47 PM
To: Jen Hammock <jen.hammock@gmail.com>

Perfect Jen. Thanks for the detailed instruction.
I now got all the steps correct :-)

Just take note that in your 2nd and 3rd sets (which are identical) have double entry for:
http://purl.obolibrary.org/obo/ENVO_00000873

Anyway, we can now move on to the next.
Please suggest what to do next based on our doc

Thanks,
Eli

[Quoted text hidden]

Jen Hammock <jen.hammock@gmail.com> Mon, Sep 10, 2018 at 2:50 PM
To: Eli Agbayani <eagbayani@eol.org>

Sweet, thanks!

The next most important category is a lot simpler (I think. Famous last words...) Please have a look at lifestage+statmeth next.

Yay!

Jen
[Quoted text hidden]

Eli Agbayani <eagbayani@eol.org> Mon, Sep 10, 2018 at 3:27 PM
To: Jen Hammock <jen.hammock@gmail.com>

Hi Jen,
So in our case: page_id: 7662 | predicate: [http://eol.org/schema/terms/Habitat]
I will be creating new rocords based on 'ROOT_ANCESTORS'.

```
if tips <= 5 SELECT ALL TIPS
else
   GET SET_1
   if SET_1 <= 4 SELECT SET_1
   else
      GET SET_2
      if SET_2 <= 4 SELECT SET_2
      else
         GET SET_3
         if SET_2 == SET_3
            if SET_3 <= 4 SELECT SET_3
            else SELECT ROOT_ANCESTORS
         else CONTINUE PROCESS UNTIL all parents of the values in the set are roots (or until current and
previous sets are identical), THEN IF <= 4 SELECT THAT SET else SELECT ROOT_ANCESTORS.
```

if(WHATEVER IS SELECTED == 1) label as: "PRM and REP"
elseif(WHATEVER IS SELECTED > 1) label as: "REP"

Thanks,
Eli

On Mon, Sep 10, 2018 at 2:50 PM, Jen Hammock <jen.hammock@gmail.com> wrote:

> Sweet, thanks!
>
> The next most important category is a lot simpler (I think. Famous last words...) Please have a look at lifestage+statmeth next.
>
> Yay!
>
> Jen
>
> On Mon, Sep 10, 2018 at 2:47 PM Eli Agbayani <eagbayani@eol.org> wrote:
>
>> Perfect Jen. Thanks for the detailed instruction.
>> I now got all the steps correct :-)
>>
>> Just take note that in your 2nd and 3rd sets (which are identical) have double entry for:
>> http://purl.obolibrary.org/obo/ENVO_00000873
>>
>> Anyway, we can now move on to the next.
>> Please suggest what to do next based on our doc
>>
>> Thanks,
>> Eli
>>
>> On Mon, Sep 10, 2018 at 12:44 PM, Jen Hammock <jen.hammock@gmail.com> wrote:
>>
>>> Sure- I think I've changed step 1 out from under you once or twice, so that's not too surprising. My current understanding, with which I constructed that set 1, is this:
>>>
>>> - find all tips
>>> - find all nodes that are parents of tips
>>> - in each case, check whether either the tip or the parent is a root
>>> - if either the tip or the parent is a root, put the tip in set 1
>>> - if neither the tip nor the parent is a root, put the parent in set 1
>>> - (deduplicate set 1)
>>>
>>> hope that helps!
>>>
>>> On Mon, Sep 10, 2018 at 12:23 PM Eli Agbayani <eagbayani@eol.org> wrote:
>>>
>>>> Hi Jen,

Thanks for the complete report.
I'm now able to generate correct lists for:

- complete tree
- roots
- tips

But I can't seem to get the correct list for Set 1.
Sorry, when I got the correct complete tree and roots and tips; I though the Set lists will be a breeze, but I can't seem to pass Set 1.
Can you please once again, state the step by step how did you arrive at your list for Set 1.

Thanks,
Eli

On Mon, Sep 10, 2018 at 10:45 AM, Jen Hammock <jen.hammock@gmail.com> wrote:
OK, I'm pretty sure this is it- but if something doesn't match, it could still be Jen Error :)

On Mon, Sep 10, 2018 at 4:54 AM Eli Agbayani <eagbayani@eol.org> wrote:
Hi Jen, if you agree with our latest_complete_tree.txt.
May I request, can you please generate the updated set 1,2,3 & 4.
So I can compare when I get your complete instructions for Step 1.

It would be nice if format will be:
- complete tree
- all roots
- tips
- set 1
- set 2
- set 3
- set 4 OR all roots

Thanks,
Eli

On Mon, Sep 10, 2018 at 1:32 AM, Eli Agbayani <eagbayani@eol.org> wrote:
Yesss, that certainly helped. Did the trick, thanks!

I now have identical tree as you have.
Except for just one additional row which we expected:
http://purl.obolibrary.org/obo/ENVO_00000109 http://purl.obolibrary.org/obo/ENVO_00000300
Attached your original tree and our latest tree with both rules.

I'm now moving along to the different steps. Will keep you posted.
Thanks,

Eli

On Sun, Sep 9, 2018 at 5:34 PM, Jen Hammock <jen.hammock@gmail.com> wrote:

Let's see... I've missed something. Oh!

The additional rule should not be

"Ancestors can be removed if they are parents of only one node BUT that node must NOT be an original node"

but

"Ancestors can be removed if they are parents of only one node BUT *the ancestor* must NOT be an original node"

Does that help?

Jen

On Sun, Sep 9, 2018 at 5:11 PM Eli Agbayani <eagbayani@eol.org> wrote:

Hi Jen,
Option 1: If we do the new rule: **node wouldn't need to be listed as an orphan if it also appears in any relationship pair.**
And only this rule, I will arrive at an identical completed tree as you have. Exactly the same as yours (jen_complete_tree.txt).

Option 2: If we do rule in Option 1 and the other rule:
"**Ancestors can be removed if they are parents of only one node BUT that node must NOT be an original node**"

If we do this, please see attached new completed tree (new_completed_tree.txt).

What do you think?
Thanks,
Eli

On Sat, Sep 8, 2018 at 8:54 PM, Jen Hammock <jen.hammock@gmail.com> wrote:

Let me see...

So it looks like 446 appears as both an orphan and an ancestor in your tree? The way I was thinking of documenting, it wouldn't need to be listed as an orphan if it also appears in any relationship pair.

As for 300... Ah, it's an original value, isn't it? In that case (my description above about cutting out middlemen notwithstanding) it should stay; an original node is not a middleman. But I don't think it should be an orphan. 109 is only a parent of one thing, but *it's still in the tree*, because it was an original node. So how should I have described that requirement...

"Ancestors can be removed if they are parents of only one node and are not original nodes" perhaps?

So there should be another record of 300, as a child of 109, (which I missed in my tree). And again, no node need be listed among the orphans if it appears elsewhere (and 300 appears with 3 different parents).

I have a feeling this process may be more complex than we need for the summary records task, but we'll be able to use these trees for a bunch of other things later, so I'm pretty sure it'll be worth the trouble :)

And, before I forget, I finally finished those new relationship files I threatened you with, one for /habitat and one for /present:

https://opendata.eol.org/dataset/terms-relationships/resource/c5ff5c62-a2ef-44be-9f59-88cd99bc8af2
https://opendata.eol.org/dataset/terms-relationships/resource/e1dcb51b-9a03-4069-b5bf-e18b6bc15798

Thanks!!

Jen

On Sat, Sep 8, 2018 at 4:20 PM Eli Agbayani <eagbayani@eol.org> wrote:

> Hi Jen, before I proceed to Step 1. Attached is your and my 'complete tree'.
> By the way I sorted them so easier to compare.
>
> I only have 2 rows more than you, the rest is identical.
>
> If you'll ask why I have an extra orphaned http://purl.obolibrary.org/obo/ENVO_00000300.
> It is because in one of the 'new nodes' section I have an entry like so:
> http://purl.obolibrary.org/obo/ENVO_00000109 http://purl.obolibrary.org/obo/ENVO_00000300
>
> Where ENVO_00000109 is a parent of only one original node.
> So the parent was discarded and ENVO_00000300 was orphaned.
>
> Does that make sense?
>
> Thanks,
> Eli
>
>
>
>
>
>
>
> On Thu, Sep 6, 2018 at 12:38 PM, Jen Hammock <jen.hammock@gmail.com> wrote:
>
>> OK, this is the last of what I have for now. I think the attached is close to the expected behavior for taxon 7662, for /habitat

I also realized my description so far doesn't account for orphan nodes, as I thought it did.

I think what is needed, in steps 1-4, is that where I said *unless the parent is a root ancestor, in which case...* that should be *unless the node or its parent is a root ancestor, in which case...*

I'll stop now until I hear from you :)

Jen

On Thu, Sep 6, 2018 at 12:01 PM Jen Hammock <jen.hammock@gmail.com> wrote:

> OK, take your time on the tree assembling, but here are some clarifications for the later steps. I'll try to put together the rest of the sample process for 7662 next.
>
> :)
>
> Jen
>
> Step 1
> "xxx tips is >5, so find set 1 (parents except where they are roots)"
> ***Question:***
> ***1. when you say 'parents', do you mean the one on the left?***
> ***If yes, so here I will pick from COMBINED_ROWS those rows where left side is not a root.***
> So, first you must identify the tips- any values that don't appear in the left column. The parents, for step one, will be the values to the left of the tip values. This is because of this i
> ***2. when you say roots, meaning it doesn't have a parent anymore?***
> Yes, roots are any value that doesn't have a parent
>
> Step 2
> "yyy nodes is >4, find set 2"
> "all parents are roots now, so steps 2 and 3 return the same set"
> ***Question: so here I will pick from COMBINED_ROWS rows where left side are roots?***
> that bit of narrative, "all parents are roots now" only applies to the dataset example I was working on. If ever you have a set 1, or a set 2 or 3, where all parents of the values in the set are roots, the next step will return the same set as this step. This is because of this bit in those steps:
>
> *"... unless the parent is a root ancestor, in which case keep ..."*

Step 4
"still >4 nodes, so select all roots"
Question: does this mean select all left side of COMBINED_ROWS where it is a root?
Yes, that sounds right. So in the 7662 example, I believe that would be something like this:

http://purl.obolibrary.org/obo/ENVO_01000206
http://purl.obolibrary.org/obo/ENVO_00000463
http://purl.obolibrary.org/obo/ENVO_00002009
http://purl.obolibrary.org/obo/ENVO_00000446
http://purl.obolibrary.org/obo/ENVO_00000144
http://purl.obolibrary.org/obo/ENVO_00002033
http://purl.obolibrary.org/obo/ENVO_01000204
http://purl.obolibrary.org/obo/ENVO_00000856
http://purl.obolibrary.org/obo/ENVO_00002030
http://purl.obolibrary.org/obo/ENVO_00000446
http://purl.obolibrary.org/obo/ENVO_00000873

On Thu, Sep 6, 2018 at 11:36 AM Jen Hammock <jen.hammock@gmail.com> wrote:

> still on the tree-making step, I think the attached is the "complete tree" to use beginning in step 1, from 7662_result
>
> On Thu, Sep 6, 2018 at 11:28 AM Jen Hammock <jen.hammock@gmail.com> wrote:
>
>> Hang in there, Eli- I'm still trying to make sure I understand your terminology, but I think we're getting pretty close.
>>
>> I'm looking at 7662_result. It's a good example because it has a lot of values in the original list.
>>
>> I think the construction of the tree is close, but a few cases are not as expected.
>>
>> in the new nodes section I see a few nodes
>>
>> http://purl.obolibrary.org/obo/ENVO_00000062
>> http://purl.obolibrary.org/obo/ENVO_00000002
>> http://purl.obolibrary.org/obo/ENVO_00000109
>>
>> that I think are ancestors of only one of the original nodes. That should mean they can be discarded, so for instance, http://purl.obolibrary.org/obo/ENVO_00000856 will have no ancestors in the tree, and be "orphaned" so to speak.
>>
>> There are also a couple of other parent nodes in the initial shared values ancestry tree
>>
>> http://purl.obolibrary.org/obo/ENVO_01000155

http://eol.org/schema/terms/tropicalOrSubtropical

that are listed in the root nodes section. These also, being the parent of only one original node, can be discarded. So their children,

http://purl.obolibrary.org/obo/ENVO_00002033
http://purl.obolibrary.org/obo/ENVO_01000204

can also be orphaned. All three orphans should be included as roots, as well as the other orphans in the initial shared values ancestry tree,

http://purl.obolibrary.org/obo/ENVO_01000206
http://purl.obolibrary.org/obo/ENVO_00000463
http://purl.obolibrary.org/obo/ENVO_00002009
http://purl.obolibrary.org/obo/ENVO_00000446
http://purl.obolibrary.org/obo/ENVO_00000144

You have two other root nodes listed, both of which I *did* expect

http://purl.obolibrary.org/obo/ENVO_00002030
http://purl.obolibrary.org/obo/ENVO_00000446

I would also expect one more root node,

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

because it is both a parent of http://purl.obolibrary.org/obo/ENVO_00000033, and a grandparent of several other original values via http://purl.obolibrary.org/obo/ENVO_01000253

I'll go on to look at step 1 now.

We're getting there!

Jen

On Thu, Sep 6, 2018 at 10:24 AM Eli Agbayani <eagbayani@eol.org> wrote:

> Here it is Jen. Thanks.
>
> On Thu, Sep 6, 2018 at 9:55 AM, Hammock, Jennifer <HammockJ@si.edu> wrote:
>
> > Bother. The Smithsonian stripped your attachment. Could you send to jen.hammock@gmail.com, please?
> >
> > Thanks!
> >
> > ---
> >
> > **From:** Eli Agbayani [eagbayani@eol.org]
> > **Sent:** Thursday, September 06, 2018 6:08 AM
> > **To:** Hammock, Jennifer
> > **Subject:** Re: task: summary data resources

Attachment removed due to policy violation: archive.zip

Hi Jen,
For review please: files in attached archive.zip
page_id: 7662 | predicate: [http://eol.org/schema/terms/Habitat]
page_id: 46559197 | predicate: [http://eol.org/schema/terms/Present]
page_id: 46559217 | predicate: [http://eol.org/schema/terms/Present]

Hopefully I got step 1 right. I may have done the opposite though.
If you can please proceed with the steps for each.

Starting at Step 1, at this point, I'm now working on the initial shared values ancestry tree PLUS the added new nodes = COMBINED_ROWS. Is that right?
Sorry, more naive questions below.

Step 1
"xxx tips is >5, so find set 1 (parents except where they are roots)"
Question:
1. when you say 'parents', do you mean the one on the left?
If yes, so here I will pick from COMBINED_ROWS those rows where left side is not a root.
2. when you say roots, meaning it doesn't have a parent anymore?

Step 2
"yyy nodes is >4, find set 2"
"all parents are roots now, so steps 2 and 3 return the same set"
Question: so here I will pick from COMBINED_ROWS rows where left side are roots?

Step 4
"still >4 nodes, so select all roots"
Question: does this mean select all left side of COMBINED_ROWS where it is a root?

Thanks for the patience Jen.
Regards,
Eli

On Wed, Sep 5, 2018 at 2:04 PM, Hammock, Jennifer <HammockJ@si.edu> wrote:

> Looks good!
>
> Thanks for your patience,
>
> Jen
>
> ---
>
> **From:** Eli Agbayani [eagbayani@eol.org]
> **Sent:** Wednesday, September 05, 2018 2:04 PM
>
> **To:** Hammock, Jennifer
> **Subject:** Re: task: summary data resources
>
> Shared values ancestry tree computed:
>
> http://www.geonames.org/6255151
> http://www.marineregions.org/gazetteer.php?p=details&id=australia
> http://www.geonames.org/6255151
> http://www.marineregions.org/gazetteer.php?p=details&id=4366
> http://www.geonames.org/6255151
> http://www.marineregions.org/gazetteer.php?p=details&id=4364
> http://www.geonames.org/6255151
> http://www.geonames.org/2186224
> http://www.marineregions.org/mrgid/1914
> http://www.geonames.org/3370751
> http://www.marineregions.org/mrgid/1914
> http://www.marineregions.org/gazetteer.php?p=details&id=1914
> http://www.marineregions.org/mrgid/1904
> http://www.marineregions.org/gazetteer.php?p=details&id=1904
> http://www.marineregions.org/mrgid/1910
> http://www.marineregions.org/gazetteer.php?p=details&id=1910
> http://www.geonames.org/6255151

http://www.marineregions.org/gazetteer.php?p=details&id=4276
http://www.geonames.org/6255151
http://www.marineregions.org/gazetteer.php?p=details&id=4365
https://www.wikidata.org/entity/Q41228
http://www.geonames.org/953987

new nodes to be added:
http://www.marineregions.org/mrgid/1902
http://www.marineregions.org/mrgid/1914
https://www.wikidata.org/entity/Q41228
http://www.marineregions.org/mrgid/1914
https://www.wikidata.org/entity/Q186198
http://www.marineregions.org/mrgid/1904
http://www.marineregions.org/mrgid/1903
http://www.marineregions.org/mrgid/1910
https://www.wikidata.org/entity/Q41228
http://www.marineregions.org/mrgid/1910

Thanks,
Eli

On Wed, Sep 5, 2018 at 12:18 PM, Hammock, Jennifer <HammockJ@si.edu> wrote:

> Cool, thanks! If any other complications arise, we can iterate more. I'm not sure I've explored all the possibilities...
>
> Jen
>
> ---
>
> **From:** Eli Agbayani [eagbayani@eol.org]
> **Sent:** Wednesday, September 05, 2018 12:17 PM
>
> **To:** Hammock, Jennifer
> **Subject:** Re: task: summary data resources
>
> Yes Jen, that makes sense, that's correct Glad you chose option 2.
> Will now proceed.
> Thanks,
> Eli

On Wed, Sep 5, 2018 at 11:57 AM, Hammock, Jennifer <HammockJ@si.edu> wrote:

> Let's see... The second method looks right to me, with the proviso that the newly added nodes
>
> http://www.geonames.org/6255151
> http://www.marineregions.org/mrgid/1914
> http://www.marineregions.org/mrgid/1904
> http://www.marineregions.org/mrgid/1910
> https://www.wikidata.org/entity/Q41228
>
> be added to the collection of nodes to be checked for shared parents. This should give you https://www.wikidata.org/entity/Q41228 (again) as a parent for /1910 and /1914 (as well as already being a parent for two of the original nodes.)
>
> Does that help?
>
> Jen
>
> ---
>
> **From:** Eli Agbayani [eagbayani@eol.org]
> **Sent:** Wednesday, September 05, 2018 11:47 AM
>
> **To:** Hammock, Jennifer
> **Subject:** Re: task: summary data resources
>
> Hi Jen,
> I may need a fresh start tomorrow but here is what I have now.
>
> If I DON'T fix the 2nd discrepancy (row 6 highlighted). It will be the only discrepancy we'll have.
> http://www.geonames.org/6255151
> http://www.geonames.org/6255151
> http://www.geonames.org/6255151
> http://www.geonames.org/6255151
> http://www.marineregions.org/mrgid/1914
> **https://www.wikidata.org/entity/Q41228**
> https://www.wikidata.org/entity/Q186198
> https://www.wikidata.org/entity/Q41228
> http://www.geonames.org/6255151
> http://www.geonames.org/6255151
> https://www.wikidata.org/entity/Q41228
> https://www.wikidata.org/entity/Q41228

If I fix the 2nd discrepancy, it will actually mess the other choices we previously gotten right (rows 7 & 8 highlighted):
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.marineregions.org/mrgid/1914
http://www.marineregions.org/mrgid/1914 ---> (discrepancy fixed; previously https://www.wikidata.org/entity/Q41228)
http://www.marineregions.org/mrgid/1904 ---> (previously https://www.wikidata.org/entity/Q186198)
http://www.marineregions.org/mrgid/1910 ---> (previously https://www.wikidata.org/entity/Q41228)
http://www.geonames.org/6255151
http://www.geonames.org/6255151
https://www.wikidata.org/entity/Q41228
https://www.wikidata.org/entity/Q41228

So there seem to be a contradiction in rules between cases. What do you think?
Thanks,
Eli

On Tue, Sep 4, 2018 at 3:25 PM, Hammock, Jennifer <HammockJ@si.edu> wrote:

> Cool!
>
> Okay, let's see. I think this will be incomplete, because I'm not sure how to construct your whole tree from the information below.
>
> For our first and discrepancy,
>
> http://www.marineregions.org/mrgid/1902 is an ancestor of http://www.geonames.org/3370751, but so is http://www.marineregions.org/mrgid/1914, and the latter is a shared parent with http://www.marineregions.

org/gazetteer.php?p=details&id=1914. The latter relationship comes from the preferred-synonym file.

This is related to the second discrepancy. You chose https://www.wikidata.org/entity/Q41228 as the parent of http://www.marineregions.org/gazetteer.php?p=details&id=1914 instead of making it a grandparent, via http://www.marineregions.org/mrgid/1914, as I did.

In the third discrepancy, I think you found:

http://www.marineregions.org/gazetteer.php?p=details&id=1904 shares an ancestor with several other values: https://www.wikidata.org/entity/Q186198

which I think is also perfectly correct, and I missed it.

And the added row at the bottom: yes, my process was to add each new shared parent node I discovered (eg: http://www.marineregions.org/mrgid/1914) as a child looking for a parent. I only ended up with that one additional row, because all the other parent nodes turned out to be roots, that could not be connected to each other through a shared parent.

Is your selection among multiple possible parents based on a ranking of how many descendants could be matched to each? I was not taking that into account. My goal was to connect all possible pairs of nodes. I think this may produce children with multiple lines of ancestry, which I think is ok.

Let me know what you think,

Jen

From: Eli Agbayani [eagbayani@eol.org]
Sent: Tuesday, September 04, 2018 2:26 PM
To: Hammock, Jennifer

Subject: Re: task: summary data resources

Hi Jen, I got same parents 9 out of 12. I put asterisk ** those three where we have different parents.

CHOSEN PARENT: http://www.geonames.org/6255151 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=australia]:
CHOSEN PARENT: http://www.geonames.org/6255151 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=4366]:
CHOSEN PARENT: http://www.geonames.org/6255151 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=4364]:
CHOSEN PARENT~: http://www.geonames.org/6255151 ---> [http://www.geonames.org/2186224]:
**CHOSEN PARENT~: http://www.marineregions.org/mrgid/1902 ---> [http://www.geonames.org/3370751]:
**CHOSEN PARENT: https://www.wikidata.org/entity/Q41228 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=1914]:
**CHOSEN PARENT: https://www.wikidata.org/entity/Q186198 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=1904]:
CHOSEN PARENT: https://www.wikidata.org/entity/Q41228 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=1910]:
CHOSEN PARENT: http://www.geonames.org/6255151 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=4276]:
CHOSEN PARENT: http://www.geonames.org/6255151 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=4365]:
CHOSEN PARENT~: https://www.wikidata.org/entity/Q41228 ---> [http://www.geonames.org/953987]:
CHOSEN PARENT~: https://www.wikidata.org/entity/Q41228 ---> [http://www.marineregions.org/mrgid/1914]:

I also attached (choosing_parent.txt) for calculation for each of the 12 terms.
This is also the ranking I computed to choose the parent:

[http://www.geonames.org/6255151] => 6
[https://www.wikidata.org/entity/Q41228] => 5
[http://www.marineregions.org/mrgid/14289] => 4
[https://www.wikidata.org/entity/Q186198] => 4
[http://www.marineregions.org/mrgid/1903] => 3
[http://www.marineregions.org/mrgid/1902] => 2
[http://www.marineregions.org/mrgid/1910] => 2
[http://eol.org/schema/terms/Australasia] => 2
[http://eol.org/schema/terms/Afrotropical] => 1
[http://www.geonames.org/6255146] => 1
[http://www.geonames.org/3358844] => 1
[http://www.marineregions.org/mrgid/1914] => 1
[http://www.marineregions.org/mrgid/1904] => 1

*Please advise on how are we going to proceed with the 3 discrepancy.

*Another question please: In your sample.txt, in section: Shared Values Ancestry from [terms relationship files]
You've added a 12th row, was that intentional?
12th row: https://www.wikidata.org/entity/Q41228 http://www.marineregions.org/mrgid/1914

Thanks that's it for now.
Eli

On Tue, Sep 4, 2018 at 11:55 AM, Eli Agbayani <eagbayani@eol.org> wrote:

> Hi Jen,
> I see, makes sense.
> I will put that criteria on script and see if I get

the same parents as you have.
Will share when done.

Yes, I must get the rules/criteria right because this is the first step in the process, and like you said the trickiest.
Thanks,
Eli

On Tue, Sep 4, 2018 at 11:26 AM, Hammock, Jennifer <HammockJ@si.edu> wrote:

> Good question!
>
> I'm afraid the answer may be computationally expensive. The criterion for choosing http://www.geonames.org/6255151 is that it is also an ancestor of other values in the original list, specifically
>
> http://www.marineregions.org/gazetteer.php?p=details&id=4366
> http://www.marineregions.org/gazetteer.php?p=details&id=4364
> http://www.geonames.org/2186224
> http://www.marineregions.org/gazetteer.php?p=details&id=4276
> http://www.marineregions.org/gazetteer.php?p=details&id=4365
>
> If this helps: I recorded no ancestors for http://www.marineregions.org/gazetteer.php?p=details&id=1904 because it didn't share any ancestors with the other values in the list. All of this with the caveat that I did this by hand and might have missed something...
>
> Does that help?
>
> Jen
>
> ---
>
> **From:** Eli Agbayani [eagbayani@eol.org]
> **Sent:** Tuesday, September 04, 2018 11:02 AM
> **To:** Jen Hammock; Hammock, Jennifer
> **Subject:** Re: task: summary data resources

Hi Jen, thanks for the sample.txt and further explanations.

Please bear with me. Question please:
In your 4th bullet point you mentioned:
"**However, we are only interested in ancestors that connect one of the values we started with to others from the same set**".
For example in your sample.txt under section: **Shared Values Ancestry from [terms relationship files]:**
In the first row:
http://www.geonames.org/6255151
http://www.marineregions.org/gazetteer.php?p=details&id=australia

How was **http://www.marineregions.org/gazetteer.php?p=details&id=australia** got a parent that is **http://www.geonames.org/6255151** ?

I made some computations and got these:

term in question:
[http://www.marineregions.org/gazetteer.php?p=details&id=australia]:
There are 2 preferred term(s) :
[0] => http://www.geonames.org/2077456
[1] => http://www.marineregions.org/mrgid/australia

parent(s) of http://www.geonames.org/2077456:
[0] => https://www.wikidata.org/entity/Q186198
[1] => https://www.wikidata.org/entity/Q41228
[2] => http://www.geonames.org/6255151
[3] => http://eol.org/schema/terms/Australasia

parent(s) of
http://www.marineregions.org/mrgid/australia: -- NO parent

What is the criteria to pick
http://www.geonames.org/6255151
among the 4 parents of
http://www.geonames.org/2077456 ?

Thanks,
Eli

On Mon, Sep 3, 2018 at 2:07 PM, Jen Hammock <jen.hammock@gmail.com> wrote:

> Hang in there, Eli- I think this is the most complex of the methods, but it'll be important for /present and /habitat data, where there are the most records, so it'll be the most helpful method of the bunch. But yes, it's convoluted. I've attempted to work with a sample (just one taxon for /Present) to clarify somewhat. Some items to note:
>
> - You'll want to keep the record identifier (eol_pk) because some of this task is identifying existing records, to apply flags to them
> - You don't need the taxon ancestry, or the predicate ancestry, for the basal value method. The terms relationships files are only needed for the values ancestry
> - You can discard literal values

- The trickiest part, I think, will be constructing the shared values ancestry tree. Each value may have several parents (to their left, in the parent-child and preferred-synonym files) and those parents can have parents of their own, creating multiple lines of ancestry for each value we started with. However, we are only interested in ancestors that connect one of the values we started with to others from the same set: to make the simplest possible tree that connects all five, or thirty, values that we have for one taxon, for one predicate. This is the tree that is used in steps 1-4.
- All that being said, I think I'll need to make you a curated relationships file for /present and another for /habitat. I think I'll want to pick and choose some terms. I'll point you to the new files when they're ready, but feel free to carry on with the terms relationships files in the meantime.

Let me know if this helps with getting to the "selected values". I'll bet I still have some explaining to do about flagging or creating records for those values once you have them.

Jen

On Mon, Sep 3, 2018 at 12:14 PM Eli Agbayani <eagbayani@eol.org> wrote:

> Hi Jen,
> Help please. Using the spreadsheet, and the carnivora dataset.
>
> I'm working on for example a single predicate = "http://eol.org/schema/terms/Present", which is under 'summary process' = "basal values". Attached is a sample working file (sample.txt.zip) for this single predicate.
>
> It has 3 sections:
> 1. Similar terms from [terms relationship files]:

2. Taxa (with ancestry) having data for predicate in question and similar terms:
3. Records from traits.csv having data for predicate in question and similar terms: (now with just a few columns)

Now looking at your instructions in worksheet "basal value". May I ask can you please manually generate the report you want, with maybe actual fields.
Or para-phrase your instructions (under 'prep' and 'steps') now with the sample data I've provided.

Please tell me if you still need other raw information/report.

The exercise is doable, I'm just still grasping the steps at the moment.
Thanks,
Eli

Jen Hammock <jen.hammock@gmail.com> Mon, Sep 10, 2018 at 3:42 PM
To: Eli Agbayani <eagbayani@eol.org>

That looks right to me!

7662 is a good example for debugging, but probably not for publishing summary values, because it's a large

taxonomic group. We may tweak this later, but it may be a good starting point to filter this process by rank, aiming to get summary values only at species level.

(There's a separate process later in the doc for selecting summary values for higher rank taxa, which we'll get to a little later.)

On Mon, Sep 10, 2018 at 3:27 PM Eli Agbayani <eagbayani@eol.org> wrote:

> Hi Jen,
> So in our case: page_id: 7662 | predicate: [http://eol.org/schema/terms/Habitat]
> I will be creating new rocords based on 'ROOT_ANCESTORS'.
>
> if tips <= 5 SELECT ALL TIPS
> else
> GET SET_1
> if SET_1 <= 4 SELECT SET_1
> else
> GET SET_2
> if SET_2 <= 4 SELECT SET_2
> else
> GET SET_3
> if SET_2 == SET_3
> if SET_3 <= 4 SELECT SET_3
> else SELECT ROOT_ANCESTORS
> else CONTINUE PROCESS UNTIL all parents of the values in the set are roots (or until current and previous sets are identical), THEN IF <= 4 SELECT THAT SET else SELECT ROOT_ANCESTORS.
>
> if(WHATEVER IS SELECTED == 1) label as: "PRM and REP"
> elseif(WHATEVER IS SELECTED > 1) label as: "REP"
>
> Thanks,
> Eli
>
> On Mon, Sep 10, 2018 at 2:50 PM, Jen Hammock <jen.hammock@gmail.com> wrote:
>
>> Sweet, thanks!
>>
>> The next most important category is a lot simpler (I think. Famous last words...) Please have a look at lifestage+statmeth next.
>>
>> Yay!
>>
>> Jen
>>
>> On Mon, Sep 10, 2018 at 2:47 PM Eli Agbayani <eagbayani@eol.org> wrote:
>>
>>> Perfect Jen. Thanks for the detailed instruction.
>>> I now got all the steps correct :-)

Just take note that in your 2nd and 3rd sets (which are identical) have double entry for:
http://purl.obolibrary.org/obo/ENVO_00000873

Anyway, we can now move on to the next.
Please suggest what to do next based on our doc

Thanks,
Eli

On Mon, Sep 10, 2018 at 12:44 PM, Jen Hammock <jen.hammock@gmail.com> wrote:

Sure- I think I've changed step 1 out from under you once or twice, so that's not too surprising. My current understanding, with which I constructed that set 1, is this:

- find all tips
- find all nodes that are parents of tips
- in each case, check whether either the tip or the parent is a root
  - if either the tip or the parent is a root, put the tip in set 1
  - if neither the tip nor the parent is a root, put the parent in set 1
- (deduplicate set 1)

hope that helps!

On Mon, Sep 10, 2018 at 12:23 PM Eli Agbayani <eagbayani@eol.org> wrote:

Hi Jen,
Thanks for the complete report.
I'm now able to generate correct lists for:

- complete tree
- roots
- tips

But I can't seem to get the correct list for Set 1.
Sorry, when I got the correct complete tree and roots and tips; I though the Set lists will be a breeze, but I can't seem to pass Set 1.
Can you please once again, state the step by step how did you arrive at your list for Set 1.

Thanks,
Eli

On Mon, Sep 10, 2018 at 10:45 AM, Jen Hammock <jen.hammock@gmail.com> wrote:

OK, I'm pretty sure this is it- but if something doesn't match, it could still be Jen Error :)

On Mon, Sep 10, 2018 at 4:54 AM Eli Agbayani <eagbayani@eol.org> wrote:

Hi Jen, if you agree with our latest_complete_tree.txt.

May I request, can you please generate the updated set 1,2,3 & 4.
So I can compare when I get your complete instructions for Step 1.

It would be nice if format will be:
- complete tree
- all roots
- tips
- set 1
- set 2
- set 3
- set 4 OR all roots

Thanks,
Eli

On Mon, Sep 10, 2018 at 1:32 AM, Eli Agbayani <eagbayani@eol.org> wrote:

> Yesss, that certainly helped. Did the trick, thanks!
>
> I now have identical tree as you have.
> Except for just one additional row which we expected:
> http://purl.obolibrary.org/obo/ENVO_00000109 http://purl.obolibrary.org/obo/ENVO_00000300
> Attached your original tree and our latest tree with both rules.
>
> I'm now moving along to the different steps. Will keep you posted.
> Thanks,
> Eli
>
> On Sun, Sep 9, 2018 at 5:34 PM, Jen Hammock <jen.hammock@gmail.com> wrote:
>
>> Let's see... I've missed something. Oh!
>>
>> The additional rule should not be
>>
>> "Ancestors can be removed if they are parents of only one node BUT that node must NOT be an original node"
>>
>> but
>>
>> "Ancestors can be removed if they are parents of only one node BUT *the ancestor* must NOT be an original node"
>>
>> Does that help?
>>
>> Jen
>>
>> On Sun, Sep 9, 2018 at 5:11 PM Eli Agbayani <eagbayani@eol.org> wrote:
>>
>>> Hi Jen,
>>> Option 1: If we do the new rule: **node wouldn't need to be listed as an orphan if it also appears in any relationship pair.**
>>> And only this rule, I will arrive at an identical completed tree as you have. Exactly the

same as yours (jen_complete_tree.txt).

Option 2: If we do rule in Option 1 and the other rule:
"**Ancestors can be removed if they are parents of only one node BUT that node must NOT be an original node**"

If we do this, please see attached new completed tree (new_completed_tree.txt).

What do you think?
Thanks,
Eli

On Sat, Sep 8, 2018 at 8:54 PM, Jen Hammock <jen.hammock@gmail.com> wrote:

> Let me see...
>
> So it looks like 446 appears as both an orphan and an ancestor in your tree? The way I was thinking of documenting, it wouldn't need to be listed as an orphan if it also appears in any relationship pair.
>
> As for 300... Ah, it's an original value, isn't it? In that case (my description above about cutting out middlemen notwithstanding) it should stay; an original node is not a middleman. But I don't think it should be an orphan. 109 is only a parent of one thing, but *it's still in the tree*, because it was an original node. So how should I have described that requirement...
>
> "Ancestors can be removed if they are parents of only one node and are not original nodes" perhaps?
>
> So there should be another record of 300, as a child of 109, (which I missed in my tree). And again, no node need be listed among the orphans if it appears elsewhere (and 300 appears with 3 different parents).
>
> I have a feeling this process may be more complex than we need for the summary records task, but we'll be able to use these trees for a bunch of other things later, so I'm pretty sure it'll be worth the trouble :)
>
> And, before I forget, I finally finished those new relationship files I threatened you with, one for /habitat and one for /present:
>
> https://opendata.eol.org/dataset/terms-relationships/resource/c5ff5c62-a2ef-44be-9f59-88cd99bc8af2
> https://opendata.eol.org/dataset/terms-relationships/resource/e1dcb51b-9a03-4069-b5bf-e18b6bc15798
>
> Thanks!!
>
> Jen
>
> On Sat, Sep 8, 2018 at 4:20 PM Eli Agbayani <eagbayani@eol.org> wrote:

Hi Jen, before I proceed to Step 1. Attached is your and my 'complete tree'.
By the way I sorted them so easier to compare.

I only have 2 rows more than you, the rest is identical.

If you'll ask why I have an extra orphaned http://purl.obolibrary.org/obo/ENVO_00000300.
It is because in one of the 'new nodes' section I have an entry like so:
http://purl.obolibrary.org/obo/ENVO_00000109 http://purl.obolibrary.org/obo/ENVO_00000300

Where ENVO_00000109 is a parent of only one original node.
So the parent was discarded and ENVO_00000300 was orphaned.

Does that make sense?

Thanks,
Eli

On Thu, Sep 6, 2018 at 12:38 PM, Jen Hammock <jen.hammock@gmail.com> wrote:

> OK, this is the last of what I have for now. I think the attached is close to the expected behavior for taxon 7662, for /habitat
>
> I also realized my description so far doesn't account for orphan nodes, as I thought it did.
>
> I think what is needed, in steps 1-4, is that where I said *unless the parent is a root ancestor, in which case...* that should be *unless the node or its parent is a root ancestor, in which case...*
>
> I'll stop now until I hear from you :)
>
> Jen
>
> On Thu, Sep 6, 2018 at 12:01 PM Jen Hammock <jen.hammock@gmail.com> wrote:
>
>> OK, take your time on the tree assembling, but here are some clarifications for the later steps. I'll try to put together the rest of the sample process for 7662 next.
>>
>> :)
>>
>> Jen
>>
>> Step 1
>> "xxx tips is >5, so find set 1 (parents except where they are roots)"

Question:
1. when you say 'parents', do you mean the one on the left?
If yes, so here I will pick from COMBINED_ROWS those rows where left side is not a root.
So, first you must identify the tips- any values that don't appear in the left column. The parents, for step one, will be the values to the left of the tip values. This is because of this i
2. when you say roots, meaning it doesn't have a parent anymore?
Yes, roots are any value that doesn't have a parent

Step 2
"yyy nodes is >4, find set 2"
"all parents are roots now, so steps 2 and 3 return the same set"

Question: so here I will pick from COMBINED_ROWS rows where left side are roots?
that bit of narrative, "all parents are roots now" only applies to the dataset example I was working on. If ever you have a set 1, or a set 2 or 3, where all parents of the values in the set are roots, the next step will return the same set as this step. This is because of this bit in those steps:

"... unless the parent is a root ancestor, in which case keep ..."

Step 4
"still >4 nodes, so select all roots"
Question: does this mean select all left side of COMBINED_ROWS where it is a root?
Yes, that sounds right. So in the 7662 example, I believe that would be something like this:

http://purl.obolibrary.org/obo/ENVO_01000206
http://purl.obolibrary.org/obo/ENVO_00000463
http://purl.obolibrary.org/obo/ENVO_00002009
http://purl.obolibrary.org/obo/ENVO_00000446
http://purl.obolibrary.org/obo/ENVO_00000144
http://purl.obolibrary.org/obo/ENVO_00002033
http://purl.obolibrary.org/obo/ENVO_01000204
http://purl.obolibrary.org/obo/ENVO_00000856
http://purl.obolibrary.org/obo/ENVO_00002030
http://purl.obolibrary.org/obo/ENVO_00000446
http://purl.obolibrary.org/obo/ENVO_00000873

On Thu, Sep 6, 2018 at 11:36 AM Jen Hammock
<jen.hammock@gmail.com> wrote:
> still on the tree-making step, I think the attached is the "complete tree" to

use beginning in step 1, from 7662_result

On Thu, Sep 6, 2018 at 11:28 AM Jen Hammock <jen.hammock@gmail.com> wrote:

> Hang in there, Eli- I'm still trying to make sure I understand your terminology, but I think we're getting pretty close.
>
> I'm looking at 7662_result. It's a good example because it has a lot of values in the original list.
>
> I think the construction of the tree is close, but a few cases are not as expected.
>
> in the new nodes section I see a few nodes
>
> http://purl.obolibrary.org/obo/ENVO_00000062
> http://purl.obolibrary.org/obo/ENVO_00000002
> http://purl.obolibrary.org/obo/ENVO_00000109
>
> that I think are ancestors of only one of the original nodes. That should mean they can be discarded, so for instance, http://purl.obolibrary.org/obo/ENVO_00000856 will have no ancestors in the tree, and be "orphaned" so to speak.
>
> There are also a couple of other parent nodes in the initial shared values ancestry tree
>
> http://purl.obolibrary.org/obo/ENVO_01000155
> http://eol.org/schema/terms/tropicalOrSubtropical
>
> that are listed in the root nodes section. These also, being the parent of only one original node, can be discarded. So their children,
>
> http://purl.obolibrary.org/obo/ENVO_00002033
> http://purl.obolibrary.org/obo/ENVO_01000204
>
> can also be orphaned. All three orphans should be included as roots, as well as the other orphans in the initial shared values ancestry tree,
>
> http://purl.obolibrary.org/obo/ENVO_01000206
> http://purl.obolibrary.org/obo/ENVO_00000463
> http://purl.obolibrary.org/obo/ENVO_00002009
> http://purl.obolibrary.org/obo/ENVO_00000446
> http://purl.obolibrary.org/obo/ENVO_00000144
>
> You have two other root nodes listed, both of which I *did* expect
>
> http://purl.obolibrary.org/obo/ENVO_00002030
> http://purl.obolibrary.org/obo/ENVO_00000446
>
> I would also expect one more root node,

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

because it is both a parent of http://purl.obolibrary.org/obo/ENVO_00000033, and a grandparent of several other original values via http://purl.obolibrary.org/obo/ENVO_01000253

I'll go on to look at step 1 now.

We're getting there!

Jen

On Thu, Sep 6, 2018 at 10:24 AM Eli Agbayani <eagbayani@eol.org> wrote:

> Here it is Jen. Thanks.
>
> On Thu, Sep 6, 2018 at 9:55 AM, Hammock, Jennifer <HammockJ@si.edu> wrote:
>
> > Bother. The Smithsonian stripped your attachment. Could you send to jen.hammock@gmail.com, please?
> >
> > Thanks!
> >
> > ---
> >
> > **From:** Eli Agbayani [eagbayani@eol.org]
> > **Sent:** Thursday, September 06, 2018 6:08 AM
> > **To:** Hammock, Jennifer
> > **Subject:** Re: task: summary data resources
> >
> > Attachment removed due to policy violation: archive.zip
> >
> > ---
> >
> > Hi Jen,
> > For review please: files in attached archive.zip
> > page_id: 7662 | predicate: [http://eol.org/schema/terms/Habitat]
> > page_id: 46559197 | predicate: [http://eol.org/schema/terms/Present]
> > page_id: 46559217 | predicate: [http://eol.org/schema/terms/Present]
> >
> > Hopefully I got step 1 right. I may have done the opposite though.
> > If you can please proceed with the steps for each.
> >
> > Starting at Step 1, at this point, I'm now working on the initial shared values ancestry tree PLUS the added new nodes = COMBINED_ROWS. Is that right?
> > Sorry, more naive questions below.

Step 1
"xxx tips is >5, so find set 1 (parents except where they are roots)"
Question:
1. when you say 'parents', do you mean the one on the left?
If yes, so here I will pick from COMBINED_ROWS those rows where left side is not a root.
2. when you say roots, meaning it doesn't have a parent anymore?

Step 2
"yyy nodes is >4, find set 2"
"all parents are roots now, so steps 2 and 3 return the same set"
Question: so here I will pick from COMBINED_ROWS rows where left side are roots?

Step 4
"still >4 nodes, so select all roots"
Question: does this mean select all left side of COMBINED_ROWS where it is a root?

Thanks for the patience Jen.
Regards,
Eli

On Wed, Sep 5, 2018 at 2:04 PM, Hammock, Jennifer <HammockJ@si.edu> wrote:

> Looks good!
>
> Thanks for your patience,
>
> Jen
>
> ---
>
> **From:** Eli Agbayani [eagbayani@eol.org]
> **Sent:** Wednesday, September 05, 2018 2:04 PM
>
> **To:** Hammock, Jennifer
> **Subject:** Re: task: summary data resources
>
> Shared values ancestry tree computed:

http://www.geonames.org/6255151
http://www.marineregions.org/gazetteer.php?p=details&id=australia
http://www.geonames.org/6255151
http://www.marineregions.org/gazetteer.php?p=details&id=4366
http://www.geonames.org/6255151
http://www.marineregions.org/gazetteer.php?p=details&id=4364
http://www.geonames.org/6255151
http://www.geonames.org/2186224
http://www.marineregions.org/mrgid/1914
http://www.geonames.org/3370751
http://www.marineregions.org/mrgid/1914
http://www.marineregions.org/gazetteer.php?p=details&id=1914
http://www.marineregions.org/mrgid/1904
http://www.marineregions.org/gazetteer.php?p=details&id=1904
http://www.marineregions.org/mrgid/1910
http://www.marineregions.org/gazetteer.php?p=details&id=1910
http://www.geonames.org/6255151
http://www.marineregions.org/gazetteer.php?p=details&id=4276
http://www.geonames.org/6255151
http://www.marineregions.org/gazetteer.php?p=details&id=4365
https://www.wikidata.org/entity/Q41228
http://www.geonames.org/953987

new nodes to be added:
http://www.marineregions.org/mrgid/1902
http://www.marineregions.org/mrgid/1914
https://www.wikidata.org/entity/Q41228
http://www.marineregions.org/mrgid/1914
https://www.wikidata.org/entity/Q186198
http://www.marineregions.org/mrgid/1904
http://www.marineregions.org/mrgid/1903
http://www.marineregions.org/mrgid/1910
https://www.wikidata.org/entity/Q41228
http://www.marineregions.org/mrgid/1910

Thanks,
Eli

On Wed, Sep 5, 2018 at 12:18 PM, Hammock, Jennifer <HammockJ@si.edu> wrote:

> Cool, thanks! If any other complications arise, we can iterate more. I'm not sure I've explored all the possibilities...
>
> Jen
>
> ---
>
> **From:** Eli Agbayani [eagbayani@eol.org]
> **Sent:** Wednesday, September 05, 2018 12:17 PM
>
> **To:** Hammock, Jennifer
> **Subject:** Re: task: summary data resources
>
> Yes Jen, that makes sense, that's correct Glad you chose option 2.
> Will now proceed.
> Thanks,
> Eli
>
> On Wed, Sep 5, 2018 at 11:57 AM, Hammock, Jennifer <HammockJ@si.edu> wrote:
>
>> Let's see... The second method looks right to me, with the proviso that the newly added nodes
>>
>> http://www.geonames.org/6255151
>> http://www.marineregions.org/mrgid/1914
>> http://www.marineregions.org/mrgid/1904
>> http://www.marineregions.org/mrgid/1910
>> https://www.wikidata.org/entity/Q41228
>>
>> be added to the collection of nodes to be checked for shared parents. This should give you https://www.wikidata.org/entity/Q41228 (again) as a parent for /1910 and /1914 (as well as already being a parent for two of the original nodes.)
>>
>> Does that help?
>>
>> Jen
>>
>> ---

From: Eli Agbayani [eagbayani@eol.org]
Sent: Wednesday, September 05, 2018 11:47 AM

To: Hammock, Jennifer
Subject: Re: task: summary data resources

Hi Jen,
I may need a fresh start tomorrow but here is what I have now.

If I DON'T fix the 2nd discrepancy (row 6 highlighted). It will be the only discrepancy we'll have.
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.marineregions.org/mrgid/1914
https://www.wikidata.org/entity/Q41228
https://www.wikidata.org/entity/Q186198
https://www.wikidata.org/entity/Q41228
http://www.geonames.org/6255151
http://www.geonames.org/6255151
https://www.wikidata.org/entity/Q41228
https://www.wikidata.org/entity/Q41228

If I fix the 2nd discrepancy, it will actually mess the other choices we previously gotten right (rows 7 & 8 highlighted):
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.geonames.org/6255151
http://www.marineregions.org/mrgid/1914
http://www.marineregions.org/mrgid/1914
---> (discrepancy fixed; previously https://www.wikidata.org/entity/Q41228)
http://www.marineregions.org/mrgid/1904
---> (previously https://www.wikidata.org/entity/Q186198)
http://www.marineregions.org/mrgid/1910
---> (previously https://www.wikidata.org/entity/Q41228)
http://www.geonames.org/6255151

http://www.geonames.org/6255151
https://www.wikidata.org/entity/Q41228
https://www.wikidata.org/entity/Q41228

So there seem to be a contradiction in rules between cases. What do you think?
Thanks,
Eli

On Tue, Sep 4, 2018 at 3:25 PM, Hammock, Jennifer <HammockJ@si.edu> wrote:

> Cool!
>
> Okay, let's see. I think this will be incomplete, because I'm not sure how to construct your whole tree from the information below.
>
> For our first and discrepancy,
>
> http://www.marineregions.org/mrgid/1902 is an ancestor of http://www.geonames.org/3370751, but so is http://www.marineregions.org/mrgid/1914, and the latter is a shared parent with http://www.marineregions.org/gazetteer.php?p=details&id=1914. The latter relationship comes from the preferred-synonym file.
>
> This is related to the second discrepancy. You chose https://www.wikidata.org/entity/Q41228 as the parent of http://www.marineregions.org/gazetteer.php?p=details&id=1914 instead of making it a grandparent, via http://www.marineregions.org/mrgid/1914, as I did.
>
> In the third discrepancy, I think you found:
>
> http://www.marineregions.org/gazetteer.php?p=details&id=1904 shares an ancestor with several other values: https://www.wikidata.org/entity/Q186198
>
> which I think is also perfectly correct, and I missed it.
>
> And the added row at the bottom: yes, my process was to add each new shared parent node I discovered (eg: http://www.marineregions.org/mrgid/1914) as a

child looking for a parent. I only ended up with that one additional row, because all the other parent nodes turned out to be roots, that could not be connected to each other through a shared parent.

Is your selection among multiple possible parents based on a ranking of how many descendants could be matched to each? I was not taking that into account. My goal was to connect all possible pairs of nodes. I think this may produce children with multiple lines of ancestry, which I think is ok.

Let me know what you think,

Jen

From: Eli Agbayani [eagbayani@eol.org]
Sent: Tuesday, September 04, 2018 2:26 PM
To: Hammock, Jennifer

Subject: Re: task: summary data resources

Hi Jen, I got same parents 9 out of 12. I put asterisk ** those three where we have different parents.

CHOSEN PARENT:
http://www.geonames.org/6255151 --->
[http://www.marineregions.org/gazetteer.php?p=details&id=australia]:
CHOSEN PARENT:
http://www.geonames.org/6255151 --->
[http://www.marineregions.org/gazetteer.php?p=details&id=4366]:
CHOSEN PARENT:
http://www.geonames.org/6255151 --->
[http://www.marineregions.org/gazetteer.php?p=details&id=4364]:
CHOSEN PARENT~:
http://www.geonames.org/6255151 --->
[http://www.geonames.org/2186224]:
**CHOSEN PARENT~:
http://www.marineregions.org/mrgid/1902 --->
[http://www.geonames.org/3370751]:
**CHOSEN PARENT:

https://www.wikidata.org/entity/Q41228 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=1914]: **CHOSEN PARENT: https://www.wikidata.org/entity/Q186198 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=1904]: CHOSEN PARENT: https://www.wikidata.org/entity/Q41228 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=1910]: CHOSEN PARENT: http://www.geonames.org/6255151 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=4276]: CHOSEN PARENT: http://www.geonames.org/6255151 ---> [http://www.marineregions.org/gazetteer.php?p=details&id=4365]: CHOSEN PARENT~: https://www.wikidata.org/entity/Q41228 ---> [http://www.geonames.org/953987]: CHOSEN PARENT~: https://www.wikidata.org/entity/Q41228 ---> [http://www.marineregions.org/mrgid/1914]:

I also attached (choosing_parent.txt) for calculation for each of the 12 terms.
This is also the ranking I computed to choose the parent:

[http://www.geonames.org/6255151] => 6
[https://www.wikidata.org/entity/Q41228] => 5
[http://www.marineregions.org/mrgid/14289] => 4
[https://www.wikidata.org/entity/Q186198] => 4
[http://www.marineregions.org/mrgid/1903] => 3
[http://www.marineregions.org/mrgid/1902] => 2
[http://www.marineregions.org/mrgid/1910] => 2
[http://eol.org/schema/terms/Australasia]

=> 2
[http://eol.org/schema/terms/Afrotropical]
=> 1
[http://www.geonames.org/6255146] => 1
[http://www.geonames.org/3358844] => 1
[http://www.marineregions.org/mrgid/1914]
=> 1
[http://www.marineregions.org/mrgid/1904]
=> 1

*Please advise on how are we going to proceed with the 3 discrepancy.

*Another question please: In your sample.txt, in section: Shared Values Ancestry from [terms relationship files]
You've added a 12th row, was that intentional?
12th row: https://www.wikidata.org/entity/Q41228
http://www.marineregions.org/mrgid/1914

Thanks that's it for now.
Eli

On Tue, Sep 4, 2018 at 11:55 AM, Eli Agbayani <eagbayani@eol.org> wrote:

> Hi Jen,
> I see, makes sense.
> I will put that criteria on script and see if I get the same parents as you have.
> Will share when done.
>
> Yes, I must get the rules/criteria right because this is the first step in the process, and like you said the trickiest.
> Thanks,
> Eli
>
> On Tue, Sep 4, 2018 at 11:26 AM, Hammock, Jennifer <HammockJ@si.edu> wrote:

Good question!

I'm afraid the answer may be computationally expensive. The criterion for choosing http://www.geonames.org/6255151 is that it is also an ancestor of other values in the original list, specifically

http://www.marineregions.org/gazetteer.php?p=details&id=4366
http://www.marineregions.org/gazetteer.php?p=details&id=4364
http://www.geonames.org/2186224
http://www.marineregions.org/gazetteer.php?p=details&id=4276
http://www.marineregions.org/gazetteer.php?p=details&id=4365

If this helps: I recorded no ancestors for http://www.marineregions.org/gazetteer.php?p=details&id=1904 because it didn't share any ancestors with the other values in the list. All of this with the caveat that I did this by hand and might have missed something...

Does that help?

Jen

From: Eli Agbayani [eagbayani@eol.org]
Sent: Tuesday, September 04, 2018 11:02 AM
To: Jen Hammock; Hammock, Jennifer
Subject: Re: task: summary data resources

Hi Jen, thanks for the sample.txt and further explanations.

Please bear with me. Question please:
In your 4th bullet point you mentioned:
"**However, we are only interested in ancestors that connect one of the values we started with to others from the same set**".
For example in your sample.txt under section: **Shared Values Ancestry from [terms relationship files]:**
In the first row:

http://www.geonames.org/6255151
http://www.marineregions.org/gazetteer.php?p=details&id=australia

How was **http://www.marineregions.org/gazetteer.php?p=details&id=australia** got a parent that is **http://www.geonames.org/6255151** ?

I made some computations and got these:

term in question:
[http://www.marineregions.org/gazetteer.php?p=details&id=australia]:
There are 2 preferred term(s) :
[0] => http://www.geonames.org/2077456
[1] => http://www.marineregions.org/mrgid/australia

parent(s) of http://www.geonames.org/2077456:
[0] => https://www.wikidata.org/entity/Q186198
[1] => https://www.wikidata.org/entity/Q41228
[2] => http://www.geonames.org/6255151
[3] => http://eol.org/schema/terms/Australasia

parent(s) of http://www.marineregions.org/mrgid/australia: -- NO parent

What is the criteria to pick **http://www.geonames.org/6255151** among the 4 parents of **http://www.geonames.org/2077456** ?

Thanks,
Eli

On Mon, Sep 3, 2018 at 2:07 PM, Jen Hammock <jen.hammock@gmail.com> wrote:

> Hang in there, Eli- I think this is the most complex of the methods, but it'll be important for /present and /habitat data, where there are the most records, so it'll be the most helpful method of the bunch. But yes, it's convoluted. I've attempted to work with a sample (just one taxon for /Present) to clarify somewhat. Some items to note:
>
> - You'll want to keep the record identifier (eol_pk) because some of this task is identifying existing records, to apply flags to them
> - You don't need the taxon ancestry, or the predicate ancestry, for the basal value method. The terms relationships files are only needed for the values ancestry
> - You can discard literal values
> - The trickiest part, I think, will be constructing the shared values ancestry tree. Each value may have several parents (to their left, in the parent-child and preferred-synonym files) and those parents can have parents of their own, creating multiple lines of ancestry for each value we started with. However, we are only interested in ancestors that connect one of the values we started with to others from the same set: to make the simplest possible tree that connects all five, or thirty,

values that we have for one taxon, for one predicate. This is the tree that is used in steps 1-4.

- All that being said, I think I'll need to make you a curated relationships file for /present and another for /habitat. I think I'll want to pick and choose some terms. I'll point you to the new files when they're ready, but feel free to carry on with the terms relationships files in the meantime.

Let me know if this helps with getting to the "selected values". I'll bet I still have some explaining to do about flagging or creating records for those values once you have them.

Jen

On Mon, Sep 3, 2018 at 12:14 PM Eli Agbayani <eagbayani@eol.org> wrote:

> Hi Jen,
> Help please. Using the spreadsheet, and the carnivora dataset.
>
> I'm working on for example a single predicate = "http://eol.org/schema/terms/Present", which is under 'summary process' = "basal values". Attached is a sample working file (sample.txt.zip) for this single predicate.
>
> It has 3 sections:
> 1. Similar terms from [terms relationship files]:
> 2. Taxa (with ancestry) having data for predicate in question and similar terms:
> 3. Records fr