D4Science Infrastructure - Task #7900

Generating Darwin Core Archives via SPD

Mar 30, 2017 10:35 AM - Gianpaolo Coro

Status:	Closed	**Start date:**	Apr 07, 2017
Priority:	High	**Due date:**	
Assignee:	Valentina Marioli	**% Done:**	100%
Category:	Application	**Estimated time:**	0.00 hour
Target version:	Community Support		
Infrastructure:	Production		

Description

I started a Job on SPD to produce the WoRMS-Animalia Darwin Core Archive. The job is blocked at the Arthropoda Phylum and at the service side there is a Bad-Gateway error. I assume that either WoRMS has blocked access from our domain or it is not possible to run such a large job. In the latter case, I can split the request into smaller requests and eventually merge, but I need an example of how to interrogate SPD asking for producing the Darwin Core Archive of the children of a certain taxon.

Subtasks:

Task # 8040: Impossible to produce DWCA for a number of Families	**Closed**

History

#1 - Mar 30, 2017 11:55 AM - Gianpaolo Coro

Indeed, we have not been banned because SPD is now able to generate the DWC-A of carcharodon group. Perhaps the failure for WoRMS-Animalia was an issue of the WoRMS service. We can go for the smaller jobs solution.

#2 - Mar 30, 2017 02:48 PM - Pasquale Pagano

@valentina.marioli@isti.cnr.it, this is a support request and as such it deserves your analysis and resolution in a short timeframe. Could you check the logs and reports if the WoRMS plugin can do something better than fail?

#3 - Mar 30, 2017 03:08 PM - Lucio Lelii

@pasquale.pagano@isti.cnr.it we are talking about a job that downloads the whole WoRMS repository (starting from animalia and requesting all children).
This job retries every call to the plugin 10 times waiting 2 secs between the retries.
The job failed because we had 10 consecutive connection timeouts contacting the WoRMS webservice, It meas that something was not working very well in the WoRMS webservice.

#4 - Mar 30, 2017 03:41 PM - Pasquale Pagano

- Status changed from New to In Progress

Ok, this what I was expecting. For me the ticket can be closed if we cannot do better than this.

#5 - Mar 30, 2017 04:54 PM - Valentina Marioli

- % Done changed from 0 to 70

I've added some classes to my private project "createDwCA" to generate DwCA by ID (including children) using threads and a class to generate a standard DwCA by Scientific Names. Please, let me know if it's working for your purpose.

#6 - Mar 31, 2017 10:13 AM - Gianpaolo Coro

- Assignee changed from Valentina Marioli to Gianpaolo Coro

#7 - Mar 31, 2017 01:48 PM - Gianpaolo Coro

I'm running the DWC-A generation process of Valentina on 1013 families of interest to FAO, by invoking the SPD service.

#8 - Apr 03, 2017 11:01 AM - Valentina Marioli

Here https://goo.gl/hai08h you can find all the direct children of Animalia in Darwin Core Archive format. The archives have been generated by the scripts in my private project "createDwCA" using threads.

#9 - Apr 04, 2017 11:27 AM - Gianpaolo Coro

- % Done changed from 70 to 90

Another script is running focussing on the families in ASFIS. This has not finished yet. Anyway, we were successful at generating one Taxonomic Authority File out of the Animalia Darwin Core Archives. I will close this ticket as soon as the families job finishes.

#10 - Apr 04, 2017 12:17 PM - Gianpaolo Coro

39 out of 45 WoRMS Phylum are present in the Darwin Core Archives. We should understand which ones are missing and regenerate them. Meanwhile, the per-family process is still running.

#11 - Apr 04, 2017 04:01 PM - Valentina Marioli

A complete list of direct children of Animalia:
18814 - Acanthocephala
765220 - Animalia incertae sedis
882 - Annelida
391884 - Aschelminthes
1803 - Brachiopoda
146142- Bryozoa
22699 - Cephalorhyncha
2081 - Chaetognatha
1821 - Chordata
1267 - Cnidaria
1248 - Ctenophora
22586 - Cycliophora
14221 - Dicyemida
1806 - Echinodermata
1271 - Entoprocta
2078 - Gastrotricha
14262 - Gnathostomulida
1818 - Hemichordata
162564 - Nematomorpha
152391 - Nemertea
14220 - Orthonectida
1789 - Phoronida
14260 - Rotifera
1268 - Sipuncula
1276 - Tardigrada
592916 - Xenacoelomorpha
799 - Nematoda
51 - Mollusca *
1065 - Arthropoda *
22737 -Placozoa *
793 - Platyhelminthes *
558 - Porifera *

(*)DwCA missing

Not accepted Phylums:
(AphiaID - Scientific Name -> status -> Accepted name)
152230 - Coelenterata -> nomen nudum -> 1268 Sipuncula
2601 - Ectoprocta -> unaccepted -> 146142 Bryozoa
162561 - Kinorhyncha -> unaccepted -> 101060 Kinorhyncha (Class)
162577 - Lophophorata -> unaccepted
162562 - Loricifera -> unaccepted -> 101061 Loricifera (Class)
14218 - Mesozoa -> unaccepted -> 14221 Dicyemida (Phylum)
596753 - Nemata -> unaccepted -> 799 Nematoda (Phylum)
1053 - Nemertina -> alternate representation -> 152391 Nemertea
246812 - Nemertini -> unaccepted -> 152391 Nemertea
196279 - Pentastomida -> unaccepted -> 22602 Pentastomida (Subclass)
1270 - Pogonophora -> unaccepted -> 129096 Siboglinidae (Family)
163280 - Rhynchocoela -> unaccepted -> 152391 Nemertea
152283 - Sipunculida -> unaccepted -> 1268 - Sipuncula

The direct children of Animalia are 45 Phylums, but maybe we should not consider not accepted Phylums.

#12 - Apr 06, 2017 02:28 PM - Valentina Marioli

New DwCAs generated:

- 1789 - Phoronida
- 799 - Nematoda

There are still some missing Phylums:

- 51 - Mollusca
- 1065 - Arthropoda
- 22737 -Placozoa
- 793 - Platyhelminthes

The process to generate the DwCAs of the direct children of Arthropoda is still running.
The other processes to generate the missing Phylums have failed, maybe because of network issues.

#13 - Apr 06, 2017 05:51 PM - Valentina Marioli

New DwCA generated:

- 22737 - Placozoa

Missing Phylums:

- 51 - Mollusca
- 1065 - Arthropoda
- 793 - Platyhelminthes

#14 - Apr 06, 2017 06:06 PM - Pasquale Pagano

- Tracker changed from Support to Task

#15 - Apr 12, 2017 03:52 PM - Gianpaolo Coro

The last three Phylum cannot be generated due to issues on the WoRMS service. I have asked Unesco people to give it a look.
Meanwhile, I have obtained the list of currently unmatched ASFIS species through BiOnym. Some of them are indeed contained in WoRMS but were not included in the DWCA for collateral issues related to the ones reported in #8040. Valentina is generating the DWCA for them, where possible.

#16 - Apr 21, 2017 04:00 PM - Gianpaolo Coro

- Assignee changed from Gianpaolo Coro to Valentina Marioli

Issues with the service remain, but at least we were able to generate the DWCAs to meet FAO requirements. The main workaround was to avoid submitting to SPD taxa names with "deleted" or "quarantine" status on the WoRMS service.

I guess the remaining families in the generation of the Animalia DWCA contain such kind of taxa names. Once the service has been fixed, Valentina can run the the DWCA process for the missing Phyla again.

#18 - Jul 31, 2018 11:05 AM - Pasquale Pagano

- Status changed from In Progress to Closed