D4Science Infrastructure - Task #8040

Task # 7900 (Closed): Generating Darwin Core Archives via SPD

Impossible to produce DWCA for a number of Families

Apr 07, 2017 04:38 PM - Gianpaolo Coro

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

Description

We cannot produce the DWCA for a number of Families because of an error from SPD. The list of Families and the error are attached. We should understand where this issue comes from and if those Families should be really discarded.

History

#1 - Apr 10, 2017 11:45 AM - Valentina Marioli

- Status changed from New to In Progress

#2 - Apr 11, 2017 06:22 PM - Valentina Marioli

- File errors.txt added

- File species.txt added

- % Done changed from 0 to 50

I've created a script to get all species given a list of scientific names through WoRMS REST webservice (http://www.marinespecies.org/rest/).
I get 783 species using the families included in the file families_list_missing.txt.
The complete list is in species.txt.
I also get 214 errors (see errors.txt) due to issues processing the URIs.
Sometimes the URI contains no data, other times the ID or scientific name seems not to exist, even when it's WoRMS REST webservice to provide such information.

Regarding the DWCAs:
Arthropoda: 1066 (AphiaID) is missing;
Mollusca: 105, 849233,18826,14503,871021, 385738, 234066 are missing;
Platyhelminthes: completed.

#3 - Apr 12, 2017 01:40 PM - Gianpaolo Coro

- % Done changed from 50 to 80

I will contact the WoRMS team to have this issue fixed. Meanwhile, using BiOnym I have checked the ASFIS species to see if they are contained in the WoRMS DWCAs (using minimum edit distance too) and it seems that there are some species present on the WoRMS site but not reached by the DWCA generation process. Strangely, I don't find some of them among the errors.

#4 - Apr 13, 2017 11:50 PM - Gianpaolo Coro

As usual, it was nice to talk with VLIZ technicians: they have said that empty families and genus are normal in WoRMS, because, for example, they do not harvest FishBase completely. They are not going to look into the issues we have highlighted, thus we should solve them by ourselves. I think that the DWCA process on SPD should be revised to avoid failure in the case of empty taxonomic branches. Perhaps, avoiding a consistency check for the DWCA could work for the issues listed in the errors file.

#5 - Apr 19, 2017 11:38 PM - Gianpaolo Coro

We are struggling to produce the DWCA for some crucial species (ca. 14400) but there are some issues:

1 - there are some species (e.g. WoRMS:183256) that are no more on the WoRMS system but, although the SPD service reports an exception internally, it continues to stay in the "running" state indefinitely;
2 - after some requests, the SPD service leaves all the requests in a "pending" state. I guess this is related to point 1 somehow.

I can see if I can find a workaround at client side, but I guess this issue affects the SPD work.

#6 - Apr 20, 2017 01:08 PM - Pasquale Pagano

- Priority changed from Normal to High

Please @valentina.marioli@isti.cnr.it analyses this issue and if some changes to the code has to be implemented this has to be tracked as a task issue. This incident is opened since several days and we should find a solution to close the issue.

#7 - Apr 20, 2017 04:14 PM - Gianpaolo Coro

It seems something happens when the service meets a taxonomic name which is either "quarantined" or deleted. Further, at a certain point the service does not accept jobs anymore and I see this exception in the ghn.log:

```
16:08:42.968 [spd-job-thread-9] WARN  AbstractLocalReader: the queue is empty
java.lang.InterruptedException: null
        at java.util.concurrent.locks.AbstractQueuedSynchronizer$ConditionObject.reportInterruptAfterWait(Abst
ractQueuedSynchronizer.java:2017) ~[na:1.7.0_80]
        at java.util.concurrent.locks.AbstractQueuedSynchronizer$ConditionObject.awaitNanos(AbstractQueuedSync
hronizer.java:2095) ~[na:1.7.0_80]
        at java.util.concurrent.ArrayBlockingQueue.poll(ArrayBlockingQueue.java:389) ~[na:1.7.0_80]
        at org.gcube.data.spd.plugin.fwk.readers.LocalReader.hasNext(LocalReader.java:21) ~[spd-plugin-framewo
rk-3.1.0-4.3.0-144690.jar:na]
        at org.gcube.data.spd.executor.jobs.dwca.MapDwCA.createTaxaTxt(MapDwCA.java:135) [MapDwCA.class:na]
        at org.gcube.data.spd.executor.jobs.dwca.MapDwCA.createDwCA(MapDwCA.java:45) [MapDwCA.class:na]
        at org.gcube.data.spd.executor.jobs.dwca.DWCAJobByIds.execute(DWCAJobByIds.java:93) [DWCAJobByIds.clas
s:na]
        at org.gcube.data.spd.executor.jobs.SpeciesJob.run(SpeciesJob.java:43) [SpeciesJob.class:na]
        at org.gcube.common.authorization.library.AuthorizedTasks$2.run(AuthorizedTasks.java:75) [common-autho
rization-2.0.2-4.3.0-144378.jar:na]
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1145) [na:1.7.0_80]
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:615) [na:1.7.0_80]
        at java.lang.Thread.run(Thread.java:745) [na:1.7.0_80]
```

I have run a web crawler that excludes deleted and quarantined names from the submitted IDs, but we still get job pending issues.

#8 - Apr 20, 2017 04:16 PM - Gianpaolo Coro

I also see many accounting issues in some threads:

```
java.lang.NullPointerException: null
        at org.gcube.data.spd.executor.jobs.SpeciesJob.generateAccounting(SpeciesJob.java:57) [SpeciesJob.clas
s:na]
        at org.gcube.data.spd.executor.jobs.SpeciesJob.run(SpeciesJob.java:48) [SpeciesJob.class:na]
        at org.gcube.common.authorization.library.AuthorizedTasks$2.run(AuthorizedTasks.java:75) [common-autho
rization-2.0.2-4.3.0-144378.jar:na]
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1145) [na:1.7.0_80]
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:615) [na:1.7.0_80]
```

#9 - May 29, 2017 03:12 PM - Gianpaolo Coro

Shall we continue to investigate this issue? We will need to periodically update our resources and we cannot lose weeks each time. Valentina, could you please make a plan (perhaps together with Lucio) to investigate the issue?

#10 - Jun 23, 2017 03:40 PM - Valentina Marioli

- Tracker changed from Incident to Task

- Status changed from In Progress to Paused

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

- Status changed from Paused to Closed

Files

Incident.txt	10.6 KB	Apr 07, 2017	Gianpaolo Coro
families_list_missing.txt	1.44 KB	Apr 07, 2017	Gianpaolo Coro
errors.txt	18.4 KB	Apr 11, 2017	Valentina Marioli
species.txt	14.8 KB	Apr 11, 2017	Valentina Marioli