EDIT - bug #9143

Distribution data sources are missing if multiple distributions exist

07/10/2020 05:18 PM - Andreas Kohlbecker

Status:	Closed	**Start date:**	
Priority:	Highest	**Due date:**	
Assignee:	Andreas Müller	**% Done:**	100%
Category:	cdm-dataportal	**Estimated time:**	0:00 hour
Target version:	Release 5.20		
Severity:	normal	**Found in Version:**	

Description

Distribution data sources are missing if multiple distributions exist for the same area and the same status. This is the case in the below example where aggregated distributions without sources are merged with explicit distributions with sources.

ERS:

etwas anderes ist jetzt komisch: einige Quellen für Verbreitungsdaten werden nicht (mehr) angezeigt, obwohl eindeutig vorhanden.

Vergleiche: (Aster alpinus) Czech Republic, France, Greece.... keine Quellen!

http://www.europlusmed.org/cdm_dataportal/taxon/229a548c-fec8-4f15-9058-8d340bb97c25

Related issues:

Related to EDIT - feature request #4366: Transmissionengine Distribution: imp...	**Duplicate**
Related to EDIT - feature request #5050: revise the subAreaPreference rule fo...	**Closed**

Associated revisions

Revision f20e21fb - 02/06/2021 08:45 PM - Andreas Müller

ref #9143 remove preferComputed rule as default from calling DefilterDistributions.filterDistributions(...) methods

Revision a8de3c3c - 02/06/2021 09:03 PM - Andreas Müller

ref #9143 rename variable name for preferComputed->preferAggregated and update javadocs

History

#1 - 07/10/2020 05:35 PM - Andreas Kohlbecker

- Status changed from New to In Progress

- Assignee changed from Andreas Kohlbecker to Andreas Müller

AM:

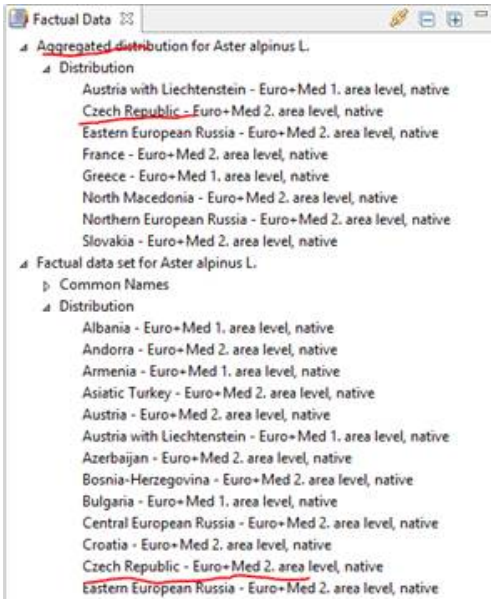
also da sind in der DB auch keine Quellen dran, da kann im Portal auch nichts rauskommen.

Übrigens sehen die Verbreitungsdaten mir aus, als ob die durch einen Import reinkamen oder so. Jedenfalls ist kein CreatedBy genannt. War aber erst am 27.11. letztes Jahr. Müssten wir nochmal checken, was da war, falls dir dazu nichts einfällt, Eckhard.

#2 - 07/10/2020 06:45 PM - Andreas Müller

- File picture103-1.png added

so, jetzt seh ich das ein bisschen genauer. Das Problem ist, dass es sich um aggregierte Daten handelt:



Aber nicht nur. Z.B Czech Republic hat sowohl aggregierte Daten als auch Originaldaten an dieser Art. Für aggregierte Daten hatten wir mal gesagt, dass diese keine Quellen anzeigen sollen, also weder einen Link zu den darunter liegenden Taxa noch zu den Originalquellen. Das könnte aber geändert werden.

Das Problem ist allerdings, dass es wohl einen Fehler im Anzeigealgorithmus gibt, wenn das gleiche Areal 2x vorkommt, 1x mit Quellen und 1x ohne. Da werden die Quellen nicht zusammengezogen, wenn es sich um den gleichen Status handelt. Bei abweichendem Status ist das glaube ich gewollt, dass nur der „höchste“ Werte genommen wird und auch nur die Quellen von diesem. Bei gleichrangigen Quellen ist das aber natürlich Quatsch, da muss man die Quellen addieren.

#3 - 07/10/2020 09:26 PM - Andreas Müller

- Subject changed from Aparranly distribution data source are missing to Distribution data sources are missing if multiple distributions exist

- Description updated

#4 - 07/14/2020 04:40 PM - Andreas Müller

- Assignee changed from Andreas Müller to Andreas Kohlbecker

- Priority changed from New to Highest

The problem is already described in #4366. However, there it is handled as pure data aggregation (transmission engine) issue. But I don't think it is an aggregation issue. If data exists from aggregation and from explicit data sources they should be combined and not taken either from one or the other distribution record (except for use-cases where this is required).

So the problem is in DescriptionUtility.filterDistributions(). There is a rule applied that computed distributions are prefered over explicit distributions:

```
        // 2) remove not computed distributions for areas for which computed
        //    distributions exists
        //
        if(preferComputed) {
```

So this means that even if the status of the computed distribution is less it is still taken (e.g. if computed -from taxonomic children- status is "introduced" and the explicit distribution is "native" the computed status "introduced" is taken) which I do not understand.

@Andreas K., as you have written this code do you remember why it was written like this?
@Eckhard: is there a rule why it should be like this?

The consequence of this rule is that the sources of explicit distribution are never shown if a computed distribution exists.

See also DescriptionUtilityTest.testFilterDistributions_computed() for the currently expected behavior.

#5 - 07/14/2020 04:40 PM - Andreas Müller

- Status changed from In Progress to Feedback

#6 - 07/14/2020 04:40 PM - Andreas Müller

- Related to feature request #4366: Transmissionengine Distribution: implement rules for source references added

#7 - 08/17/2020 07:30 PM - Andreas Kohlbecker

- Description updated

#8 - 08/17/2020 07:47 PM - Andreas Kohlbecker

- Assignee changed from Andreas Kohlbecker to Andreas Müller

Andreas Müller wrote:

> The problem is already described in #4366. However, there it is handled as pure data aggregation (transmission engine) issue. But I don't think it is an aggregation issue. If data exists from aggregation and from explicit data sources should be combined and not taken either from one or the other distribution record (except for use-cases where this is required).
>
> So the problem is in DescriptionUtility.filterDistributions(). There is a rule applied that computed distributions are prefered over explicit distributions:
>
> ```
> // 2) remove not computed distributions for areas for which computed
> //    distributions exists
> //
> if(preferComputed) {
> ```
>
> So this means that even if the status of the computed distribution is less it is still taken (e.g. if computed -from taxonomic children- status is "introduced" and the explicit distribution is "native" the computed status "introduced" is taken) which I do not understand.
>
> @Andreas K., as you have written this code do you remember why it was written like this?
> @Eckhard: is there a rule why it should be like this?
>
> The consequence of this rule is that the sources of explicit distribution are never shown if a computed distribution exists.
>
> See also DescriptionUtilityTest.testFilterDistributions_computed() for the currently expected behavior.

As far as I can remember it was meant like that:

Once the transmission engine distribution has run, the computed information is more complete and thus more valuable than the entered or imported data. This fact shoud also be clear after reading the parameter documentation:

Prefer computed rule: Computed distributions are preferred over entered or imported elements. (Computed description elements are identified by the MarkerType.COMPUTED()). This means if a entered or imported status information exist for the same area for which computed data is available, the computed data has to be given preference over other data.

When there is a situation in which the edited/imported distribution status is "native" in contrast to the computed status with value "introduced" I would assume that this is explicitly wanted or that there is a problem in the aggregation algorithm. Otherwise the edited/imported distribution status would in many cases overwrite the computation result. And in this case the question has to be raised why this rule is not incorporated into the aggregation algorithm.

#9 - 08/19/2020 02:30 PM - Andreas Müller

- Assignee changed from Andreas Müller to Andreas Kohlbecker

To me your answer seems to imply that computed data and "normal" data are based on the same data and therefore should be equal. But this is not (always) the case.

Example: We have data for A. alba from 2 sources. One has data for A. alba subsp. alba and A. alba subsp. pinus. The other source only handles species level data and has data for A. alba. Source 1 handles both subspecies as introduced, the second source handles them as native (for what ever reason both sources come to this result). Then we have on species level the computed (aggregated) data from source 1 with introduced and the data from source 2 with native. In this case source 2 should win as it has the higher status. But with the above rule it does not. And I can't see a logical reason for this.

@Eckhard: do you agree that in the above case status "native" (from source 2) should win? Or do I oversee something here?

#10 - 08/19/2020 02:33 PM - Andreas Müller

Andreas Müller wrote:

> The problem is already described in #4366. However, there it is handled as pure data aggregation (transmission engine) issue. But I don't think it is an aggregation issue. If data exists from aggregation and from explicit data sources should be combined and not taken either from one or the other distribution record (except for use-cases where this is required).

Do we agree here? My suggestion is that sources need to be combined and not only taken from the computed data. Adapting the above example and saying that also source 1 uses status native this means that the distribution for A. alba gets status native and both sources, source 1 and source 2, are attached, not only source 1 (as it is the case now).

#11 - 08/19/2020 02:34 PM - Andreas Müller

- Target version changed from Release 5.18 to Release 5.19

I move this to 5.18 as it won't be fixed today.

#12 - 08/19/2020 08:46 PM - Andreas Müller

- Target version changed from Release 5.19 to Release 5.18

#13 - 09/02/2020 08:48 AM - Andreas Kohlbecker

Andreas Müller wrote:

> Andreas Müller wrote:
>
> > The problem is already described in #4366. However, there it is handled as pure data aggregation (transmission engine) issue. But I don't think it is an aggregation issue. If data exists from aggregation and from explicit data sources should be combined and not taken either from one or the other distribution record (except for use-cases where this is required).
>
> Do we agree here? My suggestion is that sources need to be combined and not only taken from the computed data. Adapting the above example and saying that also source 1 uses status native this means that the distribution for A. alba gets status native and both sources, source 1 and source 2, are attached, not only source 1 (as it is the case now).

Eckhard wrote:

Yes, I agree. Both sources should be attached.

#14 - 09/14/2020 03:25 PM - Andreas Müller

- Status changed from Feedback to New

- Assignee changed from Andreas Kohlbecker to Andreas Müller

- % Done changed from 0 to 10

#15 - 02/01/2021 05:44 PM - Andreas Müller

- Target version changed from Release 5.18 to Release 5.21

#16 - 02/06/2021 08:56 PM - Andreas Müller

- Related to feature request #5050: revise the subAreaPreference rule for filtering Distributions added

#17 - 02/06/2021 09:54 PM - Andreas Müller

- Status changed from New to Resolved

- Target version changed from Release 5.21 to Release 5.20

- % Done changed from 10 to 50

#18 - 02/09/2021 10:02 AM - Andreas Müller

- Status changed from Resolved to Closed

- % Done changed from 50 to 100

I tested on test https://test.e-taxonomy.eu/dataportal/preview/euromed/cdm_dataportal/taxon/229a548c-fec8-4f15-9058-8d340bb97c25 and the result looks as expected for Czech Republic, France, Greece so I close this ticket.

Eckhard, please let us know if the above link does not show the expected result or if after the next release in production there are still sources missing.

#19 - 02/09/2021 10:55 AM - Andreas Müller

ERS: Hi, this looks good now on the test portal for the mentioned areas.

Files

picture103-1.png	128 KB	07/10/2020	Andreas Müller