

StocksAndFisheriesKB - Bug #18070

Tags containing a double space (maybe multiple spaces) are not searchable

Nov 14, 2019 05:07 PM - Francesco Mangiacrapa

Status:	Closed	Start date:	Nov 14, 2019
Priority:	High	Due date:	
Assignee:	Francesco Mangiacrapa	% Done:	100%
Category:		Estimated time:	0.00 hour
Target version:	GRSF		
Description			
The tags saved in the GRSF-ADMIN Catalogue like:			
Code 21.3.P.s System FAO Name Atlantic Northwest 21.3.P.s Code 21.3.L System FAO Name Atlantic Northwest 21.3.L			
that have a double space (maybe multiple spaces) in the tag value (in the above example the double space between 'Northwest' and '21.3.P.s', 'Northwest' and '21.3.L'), are not searchable by using the filtering for tags (provided by CKAN). This case generates an error in the CKAN querying system.			
We must identify where the bug is located in the CKAN source code: either fix it directly or upgrade the CKAN version to fix it			

History

#2 - Nov 25, 2019 11:06 AM - Francesco Mangiacrapa

- Status changed from New to Feedback

- % Done changed from 0 to 100

The bug reported by @aureliano.gentile@fao.org

We are currently encountering a strange behavior with GRSF tags.

When filtering using the species and area tag there are no records available, but when removing the area tag the record shows up again.

Please see the below examples where there are issues with these GRSF tags.

Short Name: Atlantic herring - Fortune Bay GRSF

Semantic identifier: asfis:HER+fao:21.3.P.S

Record URL:

http://data.d4science.org/ctlq/GRSF_Admin/9358b94d-c26a-3dbb-bbea-8ca8d946118b

[1]

When filtering by the tag "Code HER Classification System ASFIS Scientific Name Clupea harengus" the result shows, but when including the tag "Code 21.3.P.s System FAO Name Atlantic Northwest 21.3.P.s" there is no such record.

Another such record is the following:

Short Name: Atlantic herring - Bonavista - Trinity bay

GRSF Semantic identifier: asfis:HER+fao:21.3.L

Record URL:

http://data.d4science.org/ctlq/GRSF_Admin/f6b8e3f4-8413-3b5b-87dc-35e853bfdcd8

[2]

should be fixed on GRSF and GRSF-ADMIN. Please, Aureliano, could you confirm (and if OK), close this ticket?

I just released on production a patch (to the CKAN source code, see #18090) that should solve the filtering performed by Tags with more spaces.

#3 - Dec 03, 2019 10:08 PM - Aureliano Gentile

Sorry for the delayed answer but we were busy with the FAO Symposium then I was on duty travel.

I checked the above examples and seems fine.

Also, in the tag list, I see that those tags are no longer duplicated, am I correct?

If so, then the ticket can be closed unless other comments.

#4 - Dec 04, 2019 10:20 AM - Francesco Mangiacrapa

Hi Aureliano,

Aureliano Gentile wrote:

Sorry for the delayed answer but we were busy with the FAO Symposium then I was on duty travel.

don't worry...

I checked the above examples and seems fine.

Also, in the tag list, I see that those tags are no longer duplicated, am I correct?

I've not performed any action regarding the tag list and tags duplicated... I identified and fixed a bug (for tag containing multiple spaces, see #18090, those tags now are searchable) in the querying system provided by CKAN.

If you notice repeated tags in the tag list (belonging to the same GRSF record), please provide us with the record/s involved and we will take actions needed to check the issue in the source json or grsf updater service.

If so, then the ticket can be closed unless other comments.

In my opinion, this ticket can be closed.

#5 - Dec 05, 2019 06:38 PM - Aureliano Gentile

- File duplicatedTAG_5dec2019.JPG added

Thanks, for example these two tags are the same but they do retrieve different records:

Code BLU Classification System ASFIS Scientific Name Pomatomus saltatrix
Code BLU Classification System ASFIS Scientific Name pomatomus saltatrix

See also attached screenshots.

They are written in a different way, where "Pomatomus" is in upper case or lower case but they are the same.

I copy @marketak@ics.forth.gr , he may advice since this is also a content related issue.

In any case, all tags with scientific names should have the first letter of the first word in capital and all the rest in lower case.

Regarding TAG creation, upper or lower case could be ignored. But I understand it is not straight forward to build the rule identifying the right way to do it. Unless we put everything in upper case or in lower case...

#6 - Dec 06, 2019 10:47 AM - Yannis Marketakis

As regards species name, they are shown exactly as they appear in their original sources.

However, we could normalize them on the GRSF KB side so that they appear properly (Genus with first capital letter, and specific epithet in lower case). I think this would resolve the issue (at least for species tags)

#7 - Dec 06, 2019 12:51 PM - Francesco Mangiacrapa

Hi Aureliano, Yannis

Aureliano Gentile wrote:

Thanks, for example these two tags are the same but they do retrieve different records:

Code BLU Classification System ASFIS Scientific Name Pomatomus saltatrix
Code BLU Classification System ASFIS Scientific Name pomatomus saltatrix

In any case, all tags with scientific names should have the first letter of the first word in capital and all the rest in lower case.

as reported by Yannis, after a normalization performed on the GRSF KB side, so before submitting a record in the GRSF/GRSF-ADMIN Catalogue, all tags with scientific names will be identical.

Regarding TAG creation, upper or lower case could be ignored. But I understand it is not straight forward to build the rule identifying the right way to do it. Unless we put everything in upper case or in lower case...

just a clarification as regards filtering for TAG and if it should be case sensitive or case insensitive.

Now, on catalogue side, the filtering for TAG is case sensitive but there should be no problems from the technical/technological point of view to change/update it to case insensitive one (by updating the indexing schema).

In my opinion, the question is... what is the main use and semantic of a TAG in the GRSF/GRSF-ADMIN context?

In my understanding, if a TAG attached to a GRSF Record refers to scientific "data" (e.g. species scientific name, fao area, etc.) a normalization on such data should be performed before of submitting them in the GRSF/GRSF-ADMIN catalogue and the Catalogue should provide (as it is now) the filtering for TAG that is case sensitive.

#8 - Dec 06, 2019 05:21 PM - Aureliano Gentile

Thanks to all, I understand that normalization should take place in the knowledge base and leave the TAG mechanism as it is now. if so, then it is FORTH to follow up, I guess into a new ticket and this one can be closed.

#9 - Dec 09, 2019 09:36 AM - Yannis Marketakis

- Status changed from Feedback to Closed

Sure, The new ticket is [#18232](#)

Files

duplicatedTAG_5dec2019.JPG	21.5 KB	Dec 05, 2019	Aureliano Gentile
----------------------------	---------	--------------	-------------------