# StocksAndFisheriesKB - Task #12861

Task # 12858 (Closed): Purge old records from public GRSF VRE

## Purge old records of the public GRSF Catalogue

Nov 12, 2018 11:55 AM - Francesco Mangiacrapa

| | | | |
|---|---|---|---|
| **Status:** | Closed | **Start date:** | Nov 12, 2018 |
| **Priority:** | Urgent | **Due date:** | |
| **Assignee:** | Aureliano Gentile | **% Done:** | 100% |
| **Category:** | | **Estimated time:** | 0.00 hour |
| **Target version:** | GRSF | | |

**Description**

We need to remove all records of the public GRSF Catalogue (it's https://ckan-grsf.d4science.org/)

**Related issues:**

| | | | |
|---|---|---|---|
| Related to D4Science Infrastructure - Incident #12944: Error Connecting on rs... | **Closed** | **Nov 27, 2018** | **Nov 27, 2018** |
| Related to StocksAndFisheriesKB - Bug #12994: Group are not created in GRSF VRE | **Rejected** | **Dec 06, 2018** | |
| Related to D4Science Infrastructure - Task #13087: Please upgrade grsf-publi... | **Closed** | **Dec 27, 2018** | |

### History

**#1 - Nov 12, 2018 12:35 PM - Francesco Mangiacrapa**

*- Status changed from New to In Progress*

**#2 - Nov 12, 2018 05:30 PM - Francesco Mangiacrapa**

*- Status changed from In Progress to Feedback*

*- % Done changed from 0 to 100*

All records have been removed from https://ckan-grsf.d4science.org/
@marketak@ics.forth.gr and @minadakn@ics.forth.gr let me know if I have to remove the groups too (you can see them at: https://ckan-grsf.d4science.org/group)

**#3 - Nov 13, 2018 11:41 AM - Yannis Marketakis**

If we remove them they are will be re-created when we publish data. Is that correct?

**#4 - Nov 13, 2018 03:02 PM - Francesco Mangiacrapa**

Yannis Marketakis wrote:

> If we remove them they are will be re-created when we publish data. Is that correct?

Yes, It is correct. Except for bug, when records will be re-published the groups will be created on-the-fly (and the records added to them) for the fields that have the property isGroup = Yes (see at https://wiki.gcube-system.org/gcube/GCube_Data_Catalogue_for_GRSF), so may I proceed with removing the groups?

**#5 - Nov 13, 2018 04:36 PM - Yannis Marketakis**

Yes please do.
Thanks

**#6 - Nov 13, 2018 05:01 PM - Francesco Mangiacrapa**

*- Status changed from Feedback to Closed*

done https://ckan-grsf.d4science.org/group

**#7 - Dec 03, 2018 04:41 PM - Aureliano Gentile**

*- Status changed from Closed to In Progress*

New records have been published in the GRSF VRE, the entry page is not showing "Browse by Organisations" and "Browse by Groups". Any reasons why?

#### #8 - Dec 04, 2018 12:59 PM - Aureliano Gentile

As per interaction with @francesco.mangiacrapa@isti.cnr.it I would need assistance from:

@marketak@ics.forth.gr for confirming that such approved GRSF records are associated to groups
@luca.frosini@isti.cnr.it to check why groups are missing.

#### #9 - Dec 04, 2018 03:52 PM - Yannis Marketakis

@aureliano.gentile@fao.org grouping is a facility offered by the catalogue. Such information is not stored in the GRSF KB. The published records are added in groups during publishing.

#### #10 - Dec 04, 2018 05:06 PM - Aureliano Gentile

thanks, @francesco.mangiacrapa@isti.cnr.it you have your answer from FORTH.

#### #11 - Dec 04, 2018 05:38 PM - Francesco Mangiacrapa

Yannis Marketakis wrote:

> @aureliano.gentile@fao.org grouping is a facility offered by the catalogue. Such information is not stored in the GRSF KB. The published records are added in groups during publishing.

Sure and thanks @marketak@ics.forth.gr
The question is: have you already published some records that had to be added to groups during the publishing? If yes, could you attach (to this ticket) a GRSF record (as JSON source of input) already published? It will be used by @luca.frosini@isti.cnr.it to check why the groups were not created...

#### #12 - Dec 05, 2018 08:00 AM - Yannis Marketakis

@francesco.mangiacrapa@isti.cnr.it the records published in GRSF VRE are replicas of GRSF records found in GRSF Admin VRE. As such all of them fall under certain groups.

You will find the JSON contents of all these GRSF records (597 in total) at https://goo.gl/wQDAV2

#### #13 - Dec 05, 2018 10:05 AM - Francesco Mangiacrapa

Yannis Marketakis wrote:

> @francesco.mangiacrapa@isti.cnr.it the records published in GRSF VRE are replicas of GRSF records found in GRSF Admin VRE. As such all of them fall under certain groups.
>
> You will find the JSON contents of all these GRSF records (597 in total) at https://goo.gl/wQDAV2

Thanks a lot, @marketak@ics.forth.gr.
@luca.frosini@isti.cnr.it could you check this issue (why the groups were not created during the publishing) asap?

#### #14 - Dec 06, 2018 11:15 AM - Luca Frosini

*- Related to Incident #12944: Error Connecting on rstudio2.d4science.org added*

#### #15 - Dec 06, 2018 11:16 AM - Luca Frosini

*- Status changed from In Progress to Closed*

I created the ticket #12994 for the Group issue

#### #16 - Dec 06, 2018 11:26 AM - Luca Frosini

*- Related to Bug #12994: Group are not created in GRSF VRE added*

#### #17 - Dec 06, 2018 06:31 PM - Francesco Mangiacrapa

*- Status changed from Closed to In Progress*

*- % Done changed from 100 to 90*

@marketak@ics.forth.gr, @aureliano.gentile@fao.org

I just copied via script to the GRSF Catalogue (see at https://ckan-grsf.d4science.org/group) the GRSF groups already existing on GRSF-ADMIN Catalogue (https://ckan-grsf-admin2.d4science.org/group).

Now, we need to:

1. purge all GRSF records published on https://ckan-grsf.d4science.org/dataset;
2. republish approved GRSF records again to the GRSF VRE. With republishing the groups association should work fine.

Let me know if 2. is feasible, so I'll go with 1.

#### #18 - Dec 07, 2018 09:01 AM - Yannis Marketakis

As soon as there's no other way we will republish them

#### #19 - Dec 07, 2018 12:46 PM - Pasquale Pagano

*- Priority changed from High to Urgent*

This task is critical since Aureliano already sent the invitation to the GRSF colleagues for reviewing the content of GRSF.

Please take any reasonable action to fix this issue today if possible.

Thanks

#### #20 - Dec 07, 2018 03:21 PM - Luca Frosini

We cleaned the catalogue.
@marketak@ics.forth.gr can you republish the records? Thanks a lot

#### #21 - Dec 07, 2018 04:03 PM - Yannis Marketakis

Thanks Luca.
I am re-publishing right away

#### #22 - Dec 07, 2018 09:40 PM - Aureliano Gentile

thanks, no one was imaging the need to purge the grsf vre again, or at least nobody warned me. Anyway thanks a lot and sorry for that.
At the moment the group of source records are empty (i.e. Groups Stock - FIRMS or Groups Stock - FIRMS RAM, etc.). But maybe it is due to the publishing still in progress?

#### #23 - Dec 10, 2018 08:24 AM - Yannis Marketakis

The publishing of the records was completed a couple of hours after updating the ticket.
However, I still see that some groups are empty. @francesco.mangiacrapa@isti.cnr.it and @luca.frosini@isti.cnr.it can you please check?

#### #24 - Dec 10, 2018 11:43 AM - Francesco Mangiacrapa

Yannis Marketakis wrote:

> The publishing of the records was completed a couple of hours after updating the ticket.
> However, I still see that some groups are empty.

Hi @marketak@ics.forth.gr, @luca.frosini@isti.cnr.it,

checking the published records (at https://ckan-grsf.d4science.org/dataset) that were added to groups (at https://ckan-grsf.d4science.org/group), the situation seems to me the following:

1. the groups for legacy records ('Stock - FIRMS', 'Stock FIRMS FishSource' and so on) are empty because no legacy records have been published... and they will not be published. Is it right? @aureliano.gentile@fao.org, do we want to remove such groups from GRSF Catalogue?

2. no fishery records have been published, then the "GRSF Fishery" group is empty (see at https://ckan-grsf.d4science.org/group/grsf-fishery)

3. We should investigate on it (see attached screenshot)... 595 published records are "Assessment Unit" and they all had to be added to:
   ** "GRSF Assessment Unit", they are 593 - 2 records are missing;
   ** GRSF Stock", they are 592 - 3 records are missing;

If this analysis sounds for you, I can go immediately with Luca to find missing published records in those groups...

#### #25 - Dec 10, 2018 12:03 PM - Francesco Mangiacrapa

*- File grsf-publishing-add-to-groups-missing.png added*

#### #26 - Dec 10, 2018 12:11 PM - Aureliano Gentile

*- File screenshot-bluebridge.d4science.org-2018.12.10-12-05-02.png added*

1: indeed if no legacy records are published it make sense get rid of that group (but to be kept in the GRSF Admin VRE)

2: this is correct, since so far we approved only stocks records. I guess no action is needed for the time being

3: there are not consistent figures, if by types we have 597 records, by groups we should have 597 GRSF Stock and 595 GRSF assessment units and 2 GRSF marine resource, Please see also attached screenshot. (Vice-versa, if groups are correct, then types are wrong).
This is an example of what I mean when I am saying the application is not so much stable, reliable and in any release we need manual checks/fixes, and to understand the underlying reasons...

#### #27 - Dec 12, 2018 05:26 PM - Francesco Mangiacrapa

*- File LegacyGroupsToBeRemovedFromGRSFVRE.png added*

@aureliano.gentile@fao.org about the point 1.
I'm going to remove the legacy groups from GRSF Catalogue (they will be kept in the GRSF Admin VRE).
The list of legacy groups is reported (in red) in the attached image. Could you confirm the list?

#### #28 - Dec 12, 2018 05:38 PM - Aureliano Gentile

Thanks, sorry but I think there is a misunderstanding, that list is for GRSF records and those groups should be there and populated with the current numbers of approved records. None of those items should be removed. if you want tomorrow we can have a brief call on that.

#### #29 - Dec 14, 2018 04:41 PM - Luca Frosini

Hi all,

The field ***refers_to*** is used to automatically create the field ***Database Source*** and to add the GRSF record to some groups.

As the wiki report (see https://wiki.gcube-system.org/gcube/GCube_Data_Catalogue_for_GRSF#Common_Metadata), the field ***refers_to*** contains:

*"A list of objects of the format {"url": "http://", "id": "..."} that allows the aggregated GRSF records to point to their source records **already published in the catalogue. The url and the id are both mandatory and are the ones returned by the services when a source record is published."*

The code retrieves the referred records and uses their information to create the field ***Database Source*** and to add the record to the appropriates groups.
Unfortunately, the referred records (which are legacy records) are not present in GRSF hence the ***Database Source*** is not present and the groups are not added.

#### #30 - Dec 17, 2018 07:11 PM - Pasquale Pagano

*- Assignee changed from Francesco Mangiacrapa to Yannis Marketakis*

This issue has to be analyzed by FORTH. As reported, GRSF misses an information that is key for its users. This information is present in GRSF Admin but not reported in GRSF. We need shortly to find a solution to solve this issue:

- an additional field extracted by the knowledge base and specified at submission time may do the job;
- a link to the GRSF Admin record could also work but in this case, the user will find a link from GRSF to GRSF Admin and s/he will not have the rights to access it.

Please let us know.

#### #31 - Dec 18, 2018 09:11 AM - Yannis Marketakis

*- Assignee changed from Yannis Marketakis to Aureliano Gentile*

I do not see any technical difficulties here. It is clearly a matter of decision.

I would expect that colleagues from FAO come up with a decision about this and we (the technical team) proceed with this.

The alternatives I see are:

- Add only the source of the legacy record that contributed for this GRSF record (i.e. FIRMS, RAM, FishSource)
- Add the URL of the original legacy record (i.e. http://firms.fao.org/firms/resource/13792/en) or the URI if such a URL does not exist (i.e. for RAM records http://www.bluebridge-vres.eu/ram/ANCHMEDGSA16)
- Add the catalogue URL of the legacy records from the GRSF_Admin VRE
- Publish the legacy records in GRSF VRE as well

Personally speaking, I think that options 3 and 4 are not so elegant. I would like to mention again that there are no technical issues in implementing any of the above. Its clearly a decision to be made.

#### #32 - Dec 18, 2018 09:33 AM - Pasquale Pagano

Waiting for @aureliano.gentile@fao.org, please see below my personal opinion.

> The alternatives I see are:

- Add only the source of the legacy record that contributed for this GRSF record (i.e. FIRMS, RAM, FishSource)

I think this may confuse the user accessing GRSF.

- Add the URL of the original legacy record (i.e. http://firms.fao.org/firms/resource/13792/en) or the URI if such a URL does not exist (i.e. for RAM records http://www.bluebridge-vres.eu/ram/ANCHMEDGSA16)

Not so elegant as well since for RAM we have not a persistent URL (BB domain is related to the project and it will not be maintained for so many additional years)

- Add the catalogue URL of the legacy records from the GRSF_Admin VRE

I think this is safe since we maintain both GRSF Admin and GRSF. Those URLs are persistent and we can properly advise the user that their access require specific privileges.

- Publish the legacy records in GRSF VRE as well

This is clearly fine if FAO decides to select this option.

> Personally speaking, I think that options 3 and 4 are not so elegant.

As you can see above, I think that only 3 and 4 are viable.

#### #33 - Dec 18, 2018 09:43 AM - Yannis Marketakis

Thanks for your answers Lino. See some comments below:

Pasquale Pagano wrote:

> Waiting for @aureliano.gentile@fao.org, please see below my personal opinion.
>
> > The alternatives I see are:
> >
> > - Add only the source of the legacy record that contributed for this GRSF record (i.e. FIRMS, RAM, FishSource)
>
> I think this may confuse the user accessing GRSF.

I do not see why they will be confused.

> > - Add the URL of the original legacy record (i.e. http://firms.fao.org/firms/resource/13792/en) or the URI if such a URL does not exist (i.e. for RAM records http://www.bluebridge-vres.eu/ram/ANCHMEDGSA16)
>
> Not so elegant as well since for RAM we have not a persistent URL (BB domain is related to the project and it will not be maintained for so many additional years)

I agree with this.

> > - Add the catalogue URL of the legacy records from the GRSF_Admin VRE
>
> I think this is safe since we maintain both GRSF Admin and GRSF. Those URLs are persistent and we can properly advise the user that their access require specific privileges.

I agree it is safe however the problem here is that users registered in GRSF VRE should also register in GRSF-ADMIN VRE to check them out.

> > - Publish the legacy records in GRSF VRE as well
>
> This is clearly fine if FAO decides to select this option.
>
> > Personally speaking, I think that options 3 and 4 are not so elegant.
>
> As you can see above, I think that only 3 and 4 are viable.

#### #34 - Dec 18, 2018 10:56 AM - Aureliano Gentile

Thanks to all, appreciated. I discussed the matter also with Anton and I showed him also the citation aspect. We think that it would be enough to have under the box "Data and Resources" simply the list of the data source(s) as appropriate. In the public VRE the link to the legacy record was envisaged as confusing and indeed it was asked to be omitted. if you consider the citation https://support.d4science.org/issues/12278 , following the GRSF record citation we are envisaging something like this "Database sources: [FIRMS]" which then is followed by the original source record citation.

In conclusion, if feasible, at this stage it would be enough to have listed the sources and the "Groups" populated with that information. Admin users have the opportunity to browse legacy records and make all the checks while for the public could be enough like that. The citation, when completed, will give access to the source URL in the data owner websites.

Opening this pilot release to other users and the discussion at FSC11 will give further directions, if needed.

Does it make sense for you? Thanks.

#### #35 - Dec 18, 2018 11:09 AM - Yannis Marketakis

I am OK with that.

Technically speaking this means that we should include the database sources in the JSON serialization (as we do with legacy records). For example:

```
"database_sources" : [ {
    "name" : "FIRMS",
    "description" : "Fisheries and Resources Monitoring System aims to ...",
    "url" : "http://firms.fao.org/firms/en"
  },
  {
    "name" : "FishSource",
    "description" : "FishSource is an online information resource about ...",
    "url" : "http://www.fishsource.com"
  },
  {
    "name" : "RAM",
    "description" : "RAM Legacy Stock Assessment Database is ...",
    "url" : "http://ramlegacy.org"
  }
],
```

CNR colleagues is this OK with you?

#### #36 - Dec 19, 2018 11:04 AM - Luca Frosini

The solution is ok for me.
Please take into account that this behaviour will occur also on GRSF_Admin VREfor any records containing the field ***database_sources***.

If this is ok for everyone, I'll modify the code.

#### #37 - Dec 19, 2018 11:09 AM - Aureliano Gentile

I Understand this is an additional information added in the json of the grsf record, so at worst it won't be used in specific contexts. So it is fine with me. Fyi, this afternoon we'll have the first call with RAM colleagues to start validating GRSF VRE and approving new records in GRSF VRE Admin. let em know if these modifications implies erase/republish or other drastic actions in the GRSF KB.

#### #38 - Dec 19, 2018 11:43 AM - Yannis Marketakis

@@luca.frosini@isti.cnr.it we already have this feature when publishing records in the GRSF_Admin (in particular when publishing legacy records)

@aureliano.gentile@fao.org
I think we can simply update the existing records in GRSF VRE, so nothing will be removed.

#### #39 - Dec 21, 2018 03:02 PM - Luca Frosini

Yannis Marketakis wrote:

> @@luca.frosini@isti.cnr.it we already have this feature when publishing records in the GRSF_Admin (in particular when publishing legacy records)

Hi @marketak@ics.forth.gr,

sorry I lost your comment.

Looking the code seems that **database_sources**** filed is only used to create additional resources.

Legacy records are added to organizations not to group.

#### #40 - Dec 21, 2018 03:10 PM - Yannis Marketakis

So in that case, the field name should be different. Right?

#### #41 - Dec 21, 2018 03:35 PM - Luca Frosini

Yannis Marketakis wrote:

> So in that case, the field name should be different. Right?

It could. But if it is easier for you use the ***database_sources*** field I can use it.
In GRSF_admin should not cause any issues because the record is just added twice to the same group.

Just let me know what do you prefer.

#### #42 - Dec 21, 2018 04:26 PM - Yannis Marketakis

I thought it would create issues if the field name is the same.
Since it does not, then use ***database_sources***. It is fine.

#### #43 - Dec 27, 2018 04:06 PM - Luca Frosini

*- Related to Task #13087: Please upgrade  grsf-publisher-ws to latest version added*

#### #44 - Dec 28, 2018 04:18 PM - Luca Frosini

The new feature is available in the production instance.
@marketak@ics.forth.gr you can update the records when you want.
Please be sure that the update rate is limited as was agreed with Costantino.

#### #45 - Jan 07, 2019 09:17 AM - Yannis Marketakis

Hi Luca. Thanks a lot.
What exactly do you mean with the following?
Luca Frosini wrote:

> Please be sure that the update rate is limited as was agreed with Costantino.

#### #46 - Jan 08, 2019 12:05 PM - Luca Frosini

Hi Yannis,

the publishing rate should be limited to avoid failures on async operations caused by workloads.
Costantino told me you agreed on a delay between invocations.
If you are in trouble, giving that they are few records you could try to use 60 seconds.

Is that feasible?

#### #47 - Jan 08, 2019 12:43 PM - Yannis Marketakis

As far as I remember we did not have any issues with updates.
However, it is fine by me to add an idle period between updates.

Thanks

#### #48 - Jan 11, 2019 11:39 AM - Yannis Marketakis

*- Status changed from In Progress to Closed*

*- % Done changed from 90 to 100*

All the records (597 in number) in GRSF have been updated.

#### #49 - Jan 14, 2019 12:35 PM - Aureliano Gentile

I checked the GRSF VRE, groups are now available for source databases (ram, firms, fishsource) and also record pages are enriched with the box "Data and resources", similarly to GRSF Admin vre. Many thanks

### Files

| | | | |
|---|---|---|---|
| grsf-publishing-add-to-groups-missing.png | 31.6 KB | Dec 10, 2018 | Francesco Mangiacrapa |
| screenshot-bluebridge.d4science.org-2018.12.10-12-05-02.png | 122 KB | Dec 10, 2018 | Aureliano Gentile |
| LegacyGroupsToBeRemovedFromGRSFVRE.png | 20.2 KB | Dec 12, 2018 | Francesco Mangiacrapa |