D4Science Infrastructure - Task #4879

Re-harvesting of FAO Geonetwork products

Aug 25, 2016 04:34 PM - Emmanuel Blondel

Status:	Closed	**Start date:**	Aug 25, 2016
Priority:	Low	**Due date:**	
Assignee:	Fabio Sinibaldi	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	UnSprintable		
Infrastructure:	Production		

Description

Is it possible to re-harvest the FAO Geonetwork FIGIS products so they appear in the iMarine catalogue (https://i-marine.d4science.org/group/imarine-gateway/data-catalogue) and GeoExlorer ? Several updates (including bug fixing) have been applied in the set of GIS metadata.

Thanks in advance

History

#1 - Aug 26, 2016 11:42 AM - Leonardo Candela

- Assignee set to Francesco Mangiacrapa

As far as I know the harvesting should be automatic (there might be some delay due to the scheduling interval).

I'll add in CC some colleagues that can say more on how the harvesting is configured.

#2 - Aug 30, 2016 03:13 PM - Pasquale Pagano

- Assignee changed from Francesco Mangiacrapa to Gianpaolo Coro

@gianpaolo.coro@isti.cnr.it can you clarify the harvesting rules and proceed with the update? Thanks.

#3 - Aug 30, 2016 03:20 PM - Gianpaolo Coro

- Assignee changed from Gianpaolo Coro to Fabio Sinibaldi

As far as I have understood, with the new management of GeoNetwork, with dynamic users and groups, it is impossible for a human manager to re-harvest an external GeoNetwork. Basically, I don't know to which group(s) the layers should be assigned. The groups are managed as numbers from the libraries/services and as alpha-numeric strings on the GN interface. How to get the mapping is not clear to me. I guess only Fabio (@fabio.sinibaldi@isti.cnr.it) can indicate a solution for this. Perhaps there is a trivial solution I cannot see now.

#4 - Aug 30, 2016 03:32 PM - Fabio Sinibaldi

Hi,
well as usual the harvesting of fao GN metadata is set for the reserved group All (rights "view" and "interactive map"), and the category "fao". However there's no need to configure it again to make it run. At the moment it's set to run just once for performance issues we had in the past, but whoever has admin privileges can launch it and/or configure it. Anyway I just launched it.

#5 - Aug 30, 2016 03:35 PM - Fabio Sinibaldi

- Status changed from New to In Progress

#6 - Aug 30, 2016 03:37 PM - Gianpaolo Coro

What VRE/VO visibility is associated to the all group? All the root/VO/VRE levels?

#7 - Aug 30, 2016 03:43 PM - Fabio Sinibaldi

According to GeoNetwork documentation "A public metadata is a metadata that has the view privilege for the group named all.", the harvested metadata is public.

#8 - Aug 30, 2016 04:50 PM - Fabio Sinibaldi

- Status changed from In Progress to Feedback

Harvesting seems to be completed. Can you please check if every expected metadata has been imported by using geoexplorer portlet?
NB : ckan catalogues might take longer to update, so don't rely on that information for this purpose.

#9 - Aug 30, 2016 06:38 PM - Emmanuel Blondel

it's ok in Geoexplorer, but not in the data catalogue which i was looking primarily (due to its search capacities). Can you clarify how the information is harvested in the latter? It's not clear why it is harvested in Geoexplorer and not in the data catalogue.

Question: i would have questions/feedback about the data catalogue, and the way our products are handled there. Can you please tell me how you want me to proceed? single support ticket? distinct set of support tickets? thanks

#10 - Aug 31, 2016 10:37 AM - Fabio Sinibaldi

Well, simply put : the data catalog performs harvesting of GeoNetwork information, and until that the catalog content is not updated.
GeoExplorer performs direct queries on our GeoNetworks, so you directly access that information.

About further questions/feedback I think it really depends on what you mean. Probably @gianpaolo.coro@isti.cnr.it can tell which way would be better.

#11 - Aug 31, 2016 03:29 PM - Francesco Mangiacrapa

Re-Harvesting of GeoNetwork for FAO Layers on D4science Data Catalogue has finished. You can see new layers here:
https://ckan-d4s.d4science.org/organization/fao

FAO FAGIS is displaying 1,364 products instead of 1368 (layers harvested on GeoNetwork, see:
http://geonetwork.d4science.org/geonetwork/srv/en/main.home and then click on "fao" category). This mismatch is due to 4 Dataset schema (gmx.xsd) Validation Error:

content: https://ckan-d4s.d4science.org/harvest/object/fbd7d354-de03-44df-a62e-23d8692c0bbd

```
4c9e8750-dbba-11dc-9d70-0017f293bd28
Dataset schema (gmx.xsd) Validation Error
Element '{http://www.isotc211.org/2005/gco}DateTime': '2008-02-15' is not a valid value of the atomic type 'xs
:dateTime'. (line 113)
Element '{http://www.isotc211.org/2005/gco}Real': '0,004' is not a valid value of the atomic type 'xs:double'.
 (line 230)
```

content: https://ckan-d4s.d4science.org/harvest/object/92c2107c-332b-490b-bcbf-419b03dbd9af

```
vliz-eez-map-8314
Dataset schema (gmx.xsd) Validation Error
Element '{http://www.isotc211.org/2005/gmd}MD_TopicCategoryCode': [facet 'enumeration'] The value 'OCEAN' is n
ot an element of the set
{'farming', 'biota', 'boundaries', 'climatologyMeteorologyAtmosphere', 'economy', 'elevation', 'environment',
'geoscientificInformation',
'health', 'imageryBaseMapsEarthCover', 'intelligenceMilitary', 'inlandWaters', 'location', 'oceans', 'planning
Cadastre',
'society', 'structure', 'transportation', 'utilitiesCommunication'}. (line 472)
Element '{http://www.isotc211.org/2005/gmd}MD_TopicCategoryCode': 'OCEAN' is not a valid value of the atomic t
ype
'{http://www.isotc211.org/2005/gmd}MD_TopicCategoryCode_Type'. (line 472)
```

content: https://ckan-d4s.d4science.org/harvest/object/0aac85fa-22c9-4fda-b953-7df1b95f9a42

```
84017a9f-09bf-4b18-a346-dff0e661d86f
Dataset schema (gmx.xsd) Validation Error
Element '{http://www.isotc211.org/2005/srv}SV_ServiceIdentification': This element is not expected. Expected i
s one of
( {http://www.isotc211.org/2005/gmd}AbstractMD_Identification, {http://www.isotc211.org/2005/gmd}MD_DataIdenti
fication,
{http://www.isotc211.org/2005/gmd}MD_ServiceIdentification ). (line 80)
```

content: https://ckan-d4s.d4science.org/harvest/object/3ac2ac42-7224-4455-82fa-02e3f4b4efdb

```
ac02a460-da52-11dc-9d70-0017f293bd28
Dataset schema (gmx.xsd) Validation Error
Element '{http://www.isotc211.org/2005/gco}Integer': '' is not a valid value of the atomic type 'xs:integer'.
(line 120)
Element '{http://www.isotc211.org/2005/gco}DateTime': '2008-02-15' is not a valid value of the atomic type 'xs
:dateTime'. (line 150)
```

#12 - Aug 31, 2016 04:04 PM - Emmanuel Blondel

Thanks for your feedback on validation issues. FYI, these has been fixed, as follows:

- 4c9e8750-dbba-11dc-9d70-0017f293bd28 --> fixed numeric field
- vliz-eez-map-8314 --> deleted metadata (old draft)
- 84017a9f-09bf-4b18-a346-dff0e661d86f --> deleted metadata
- ac02a460-da52-11dc-9d70-0017f293bd28 --> add missing count of vector objects

Thanks for the update on data catalogue. It's not clear if the 2 harvesting (Geonetwork <- FAO Geonetwork, CKAN <- Geonetwork) are automatic, and if yes what is the periodicity. Thanks in advance if you can clarify (maybe documenting it somewhere would be useful, if not yet done).

NB: On the data catalogue (which is very useful in term of search capacity): my first reaction is that i see the products are under a "FAO" group, and "FIGIS" subgroup, but strangely i'm not part of them while i'm the individual owner/editor of the harvested metadata. I will look more into the data catalogue, and send feedback if needed.

#13 - Aug 31, 2016 07:56 PM - Pasquale Pagano

The catalog is a new product released in August. There is not yet a scheduled harvesting so far. We run the harvesters and check the results to avoid problems (we are verifying how it works with metadata updates, with deleted metadata, and so on). In the future (release 4.1) we will schedule the harvesters and they will run every night. Documentation will be added as well.

#14 - Sep 05, 2016 06:17 PM - Fabio Sinibaldi

- Status changed from Feedback to Resolved

I suppose we can close this one. Please, re open it if you need to.

#15 - Sep 05, 2016 06:23 PM - Emmanuel Blondel

Dont' know why but now FAO products are no longer available in the data catalogue, CKAN based, not the CSW one. I'm referring to the 'production' environment.
I see this as notification: "Francesco Mangiacrapa deleted the product Geonetwork Harvester for Fao Figis Layers"

#16 - Sep 05, 2016 07:03 PM - Pasquale Pagano

- Status changed from Resolved to In Progress

- Priority changed from Normal to High

#17 - Sep 06, 2016 11:01 AM - Francesco Mangiacrapa

Hi Emmanuel,
FAO products with its Geonetwork Harvester have been moved in the BB catalogue: https://ckan-bb.d4science.org/organization/fao
The main catalogue https://ckan-d4s.d4science.org/ will be only a "simple" container, that is a CKAN harvester for other CKAN instances (ckan-bb, ckan-sobigdata).
However, FAO products (with FAO Organization, etc.) will reappear again in the main catalogue when CKAN D4Science will restart harvesting for ckan-bb and ckan-sobigdata
ok?

#18 - Sep 06, 2016 11:20 AM - Emmanuel Blondel

Thanks, i understand this diff, but what are we supposed to have in the i-Marine portal https://i-marine.d4science.org/group/imarine-gateway/data-catalogue? BB or the whole D4S? IMHO it is the BB one, that's why it's not clear why FAO products are not available in the i-Marine data catalogue which should point to https://ckan-bb.d4science.org/ or no?

#19 - Sep 06, 2016 12:17 PM - Leonardo Candela

- Priority changed from High to Low

- % Done changed from 0 to 90

What are the products "published" by each catalogue is a matter of configuration, each catalogue should be oriented to serve the needs of a designated community.

As previously commented, the CKAN-based Catalogue is a just released service that will evolve in the coming months. Moreover, it exists in many incarnations, the D4Science catalogue, the BB Catalogue, the Catalogue. Each catalogue instance, can be populated by both

- explicit publishing, adding datasets (and their descriptions) directly to the catalogue;
- harvesting, collecting products from other catalogues (including GN).

Having clarified this, I assume it is quite easy to figure out how challenging might be to properly configure such an array of services and the flow of information among them.

The initial idea is to have CKAN catalogues organised thus to form a two-levels hierarchy having in D4Science catalogue its root. The root catalogue is expected to collect products out of the BB and catalogues. The point is that we would like to not introduce duplicates in the root. In order to guarantee this (with the current version of the technology) we have to avoid that products appear in two "leaf" catalogues.

We are well aware on the fact that there are possible limitations in this approach and are studying proper configuration strategies that will be in place

in the next releases.

For FAO products, the approach is to harvest them in the D4Science GN (they will be collected from here to populate the various catalogues).

I'm going to lower the priority of this ticket right now since:

- the ticket is about the re-harvesting of FAO dataset (this is done, i.e. datasets are in GN);
- the FAO datasets are in the BB catalogue (and will appear in the D4Science catalogue soon);

Re what is the "right" catalogue to attach to the iMarine portal, your comment is correct it should be the BB CKAN instance. We will modify this and once done will close the ticket.

#20 - Sep 06, 2016 12:24 PM - Emmanuel Blondel

perfect, thanks

#21 - Sep 20, 2016 12:15 PM - Fabio Sinibaldi

- Status changed from In Progress to Feedback

Metadata has been harvested. Please provide feedback if needed or close this ticket if everything is ok.

#22 - Sep 20, 2016 12:31 PM - Emmanuel Blondel

I still don't see products in the data catalogue https://i-marine.d4science.org/group/imarine-gateway/data-catalogue, there is no 'FAO' organization. If harvested through D4Science, the latter organization mentions '0 products'.

#23 - Sep 20, 2016 12:44 PM - Francesco Mangiacrapa

At the moment, Fao products are available on CKAN BB instance: https://ckan-bb.d4science.org/organization/fao.
CKAN BB is reachable in the i-marine/services portal via VRE's. For example: https://services.d4science.org/group/biodiversitylab/data-catalogue
The Fao Products will be available also on D4Science Data Catalogue after that ckan harvester cron will harvest from CKAN D4Science to CKAN BB.

#24 - Sep 20, 2016 12:54 PM - Emmanuel Blondel

Ok please let me know when done, afterwhat we will close this ticket.

#25 - Sep 22, 2016 02:18 PM - Francesco Mangiacrapa

- % Done changed from 90 to 100

Fao Layers are now available on D4Science Data Catalogue: https://i-marine.d4science.org/group/imarine-gateway/data-catalogue

#26 - Sep 22, 2016 02:22 PM - Emmanuel Blondel

- Status changed from Feedback to Closed

Many thanks

#28 - Oct 10, 2016 07:21 PM - Pasquale Pagano

- Tracker changed from Support to Task