NEXTGEOSS

European Data Hub and Platform

Data and Service Cataloging

NextGEOSS User Guide

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730329.

Ingestion of metadata records into CKAN catalogues

Third-party data providers identified by the NextGEOSS Pilots as providers of Open Data resources useful for their Pilot application can be included as registered resources, accessible through the NextGEOSS DataHub.
Also, Partners in charge of supporting the Pilots (e.g. DLR, NOA, CLS/CMEMS...) with their data offering can have their data servers included as registered resources, accessible through the NextGEOSS DataHub.
To this end, ad-hoc software connectors are designed as "CKAN extensions", and implemented in the NextGEOSS DataHub, in order to harvest the Data Provider's metadata records (note: this is an online operation, running over web service endpoints) that are describing the relevant data collections.

Using connectors as CKAN remote harvesting extensions

Within NextGEOSS, the evolvability of the CKAN Open Source software is exploited in order to create ad-hoc software connectors, designed as "CKAN extensions".
A CKAN connector is implemented according to a specification describing a web service endpoint hosting metadata (typically, a catalogue endpoint) that documents the data collections and data products to be exploited by a NextGEOSS Pilot.
When ready, a connector is implemented in the NextGEOSS DataHub, and is operated in order to feed (and update) the NextGEOSS dataHub with metadata records. Such records allow search operations to select a specific part of a data collection (datasets matching search criteria such as geolocation, creation dates, keywords...). The search results include the information about the datasets access endpoints, where actual datasets can be fetched by a Pilot application (especially, the data processing component of it).
Operational Metadata harvesters tested within the NextGEOSS DataHub are currently:

- Sentinel 1, 2, 3
    - Domain: ground surface parameters, clouds, aerosols

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730329.

  - Connector developed with redundant "multi-repository" harvesting, encompassing dataset sources from the Copernicus SciHub, the Hellenic Sentinel DataHub (NOA) and the German Copernicus collaborative ground segment (CODE-DE)
- MetOp/[AB]-GOME-2
  - Domain: atmospheric trace gases e.g. ozone, NO2, SO2, ...
- CMEMS
  - Domain: Marine environment

The guidance for the integration of a new CKAN Connector is available from:

- https://docs.ckan.org/en/ckan-2.7.3/extensions/tutorial.html (CKAN documentation)
- https://github.com/ckan/ckanext-harvest (ckanext-harvest - Remote harvesting extension, a common harvesting framework for ckan extensions)
- NextGEOSS deliverable D2.1.1 Data Ingestion Guide - Version 2 (NextGEOSS Connectors specific aspects: harvesting from multiple repository sources, use of Dublin Core as common information model, ...)

Using metadata enrichment techniques - the ws-iTag service

iTag is an external application that can be queried during the harvesting process, to get additional metadata for datasets based on their spatial coordinates.

It is a web service meant for the semantic enhancement of Earth Observation products, i.e. the tagging of products with additional information about the covered area, regarding for example geology, water bodies, land use, population, countries, administrative units or names of major settlements, added as extra fields in the common information model of a set of metadata records.

The guidance for the integration of the iTag service is available from the following GitHub repositories:

- https://github.com/jjrom/itag (original iTag repository)
- https://github.com/Terradue/ws-itag (Terradue's iTag repository, which has a fix for the installation done for NextGEOSS, and provides guidance for data ingestion)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730329.

The metadata enrichment mechanism relies on submitting the footprint of a dataset or resource to the ws-iTag service, which returns a list of tags that are relevant for that footprint.

The metadata enrichment workflow consists in the following operations:

1. Harvest a dataset's metadata record from the source (see the section "Using connectors as CKAN remote harvesting extensions")
2. Query from within the harvester's code the ws-iTag service using the dataset's footprint
3. Update the dataset's metadata using the information returned by ws-iTag (i.e. add the retrieved information to the harvested metadata records)
4. Store the result (enriched metadata records) in CKAN

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730329.