Activity 2 - Task 2.2, 2.3 & 2.4

Progress Report / Deliverable Report

Co-financed by the Connecting Europe Facility of the European Union

1 March 2022

GEOE3

Revision History

Version	Status	Author	Modifications
0.1	Initial notes	Laura Alemany Gómez	Notes for first draft
0.2	First draft	Lexi Rowland	Additions and details for specific tasks
0.9	Draft for PMG	Dorus Kruse	Review remarks for PMG
0.95	Approved draft by PMG	Dorus Kruse	Last comments and approval added
1.0	Final version	Dorus Kruse	Links in appendix corrected.

Distribution

Version	Status	Recipient
0.9	draft	PMG
0.95	Approved draft by PMG	PMG
1.0	Final version	PMG

Table of Contents

Appendices

1. Introduction

The overall objective for Activity 2 is to discover and evaluate the data and APIs needed for the creation of new quality tools and to support new methodologies on data quality; therefore making metadata more accessible and visual. Having provided in depth descriptions of the tasks to be carried out in previous documentation, this document intends to present these descriptions and provide a status on the progress made in working towards these tasks as deliverables. In this deliverable report, the results of carrying out an initial investigation into the following tasks within Activity 2 are presented.

- Task 2.2 National and European Data Portal Connectivity

  This task will develop the connection between the national infrastructure and the European Data Portal (EDP). The Action will provide tools for publishing service metadata on the European Data Portal and, on the other hand, retrieve and re-use the portal content for the selected use cases. The metadata will be published using the DCAT metadata profile, which enables data sharing and re-use between sectors and between ecosystems. The CEF eTranslation building block tool will be used to provide translations between the national language and the target language. Content discovery service will be published as an Open Geospatial Consortium (OGC) API - Catalogues interface.

- Task 2.3 Semantic Search

  This task will investigate existing national profiles of metadata with the aim of subsequently using the profiles for the provision of content discovery and evaluation services. New tools for enhancing discovery and evaluation of geospatial data will be developed and these will include the use of ontologies and semantic content, utilizing INSPIRE-based content registries. The feasibility and possible alignments of the GEMET thesauri (applied in INSPIRE), or other alternatives for deploying a common ontology will also be investigated.

- Task 2.4. Support for Search Engines

  This task will develop mechanisms for making data discovery services available through Internet search engines, like Google.

The following sections of this report will provide an overview of the product description for the specified tasks within this Activity including the intended output and will then provide a description of the actual activities carried out over the course of the project and the results achieved to date. Please note, this report is accompanied by a set of files in the appendices of this report which present some of the results identified here. Where applicable these are referenced in the sections that follow.

2. Product Description

Because Task 2.2, 2.3 and 2.4 are closely related to each other, the following section will provide a single product description for all three tasks. This product description is closely aligned with the product description delivered in May 2021. This report and the accompanying files represent the finalisation of these tasks in line with deliverables outlined in the product description.

2.1. Purpose

For each task, the purpose in carrying out the activities are defined.

2.1.1. Task 2.2: National and European Data Portal Connectivity

Provide the tools for publishing service metadata on the EDP based on a connection between the national infrastructure (National Geoportals) and the EDP. This requires DCAT-AP implementation for National Geoportals.

2.1.2. Task 2.3: Semantic Search

With the aim of improving discovery and evaluation services, this task aims to enrich existing metadata profiles using existing ontologies and semantic INSPIRE-based content registries. Feasibility of enriching profiles with the GEMET thesauri or other alternatives will be investigated with the goal of deploying a common ontology for semantic enrichment of metadata. Within the context of this task, the opportunity to enrich the current metadata profiles with quality information will also be assessed.

2.1.3. Task 2.4: Support for Search Engines

Based on the output from 2.2, the resultant metadata profiles will be mapped to schema.org (SDO) with the goal of improving Google findability. This effectiveness of this should be evaluated by performing a Rich Results Test.

2.2. Quality Criteria

In carrying out the tasks presented above, the following quality criteria were defined. The success in achieving these criteria in carrying out the task activities are discussed in the sections that follow.

2.2.1. Connection to EDP

A minimum criteria with respect to the EDP connectivity should ensure that a connection between the National Geoportals and the EDP can be made.

2.2.2. Metadata Profiles

For each metadata quality profile, there exist a list of required categories which should be included in order to be compliant with the profile. As such, the quality criteria for the metadata profiles should be minimum compliance with all metadata profiles and can be automatically validated using SHACL shape technologies. The quality criteria for the mapping between profiles should be defined by experts, reported as part of these tasks and assessed herein.

2.2.3. Metadata Quality

- For DCAT-AP: Interview/questionnaire conducted with the EDP and the National Open Data Portals to see whether the output file from the developed tool can be read.
- Semantic search enrichment: SPARQL-based validation to see whether vocabularies have been accurately applied.
- Schema.org Schema: Rich Results Test (Google based validation tool for metadata profile).

2.3. Quality Tolerance

In defining the quality criteria, the following quality tolerance was defined as necessary and the extent to which this was achieved in carrying out the tasks is discussed in the sections that follow.

2.3.1. Connection to EDP

Whether connection can be made from National Geoportals to both the EDP and the Open Data Platforms.

2.3.2. Semantic Search

Whether or not an ElasticSearch instance provides a result based on a given query. No result highlights poor (and intolerable) quality.

2.3.3. Schema.org Schema

Google's tooling will provide feedback on whether the profile defined is sufficient or not for search engine feasibility. Negative notification highlights poor (and intolerable) quality.

2.4. Quality Method

In carrying out the quality checks required for these tasks and in implementing the criteria and tolerances defined about, the following methods were used to ensure a quality in the results. Where applicable, a more extensive description of these methods and their results per task are presented in the sections that follow.

2.4.1. Metadata Quality

- For DCAT-AP: Interview/questionnaire conducted with the EDP and the National Open Data Portals to see whether the output file from the developed tool can be read
- Semantic search enrichment: SPARQL-based validation to see whether vocabularies have been accurately applied and are, therefore, findable
- Schema.org Schema: Rich Results Test (Google based validation tool for metadata profile)

2.5. Deliverable Output

Based on the product description, the following deliverable outputs were defined. This report and the associated files are in line with this deliverable output and represent a completion of these.

2.5.1. Deliverable Report

The deliverable report should include:

- Description of the tools required and/or developed for the purpose of achieving the aims of the tasks

- Description of the mapping between the different metadata profiles

The report presented in the sections that follow serve to meet this criteria.

2.5.2. Excel Sheet

The Excel sheet which accompanies the deliverable report in Appendix E should include 3 profiles/schema mappings:

- Elements from the INSPIRE profiles (format: xml) → I think Lexi has the last version of this.
- Corresponding elements from the DCAT-AP profile (format: RDF)
- Enriched DCAT-AP profile (format: RDF)
- Corresponding elements from the SDO profile (format: RDF)

The Excel file fulfilling these requirements has been included as an accompanying file to this report as available in Appendix E.

2.5.3. Outputs

In order to provide example outputs of the results of these tasks to date, the following is required.

- Example 1: An XML file in INSPIRE Profile (TG 2.0).
- Example 2: A RDF file in DCAT-AP profile.
- Example 3: A RDF file in GeoDCAT-AP profile.
- Example 4: An enriched DCAT-AP profile to support semantic search
- Example 5: DCAT-AP to SDO mapping in excel format (includes INSPIRE-DCAT-AP mapping)

These examples are presented in the Appendix sections of this report.

- Example 1: Appendix A
- Example 2: Appendix B
- Example 3: Appendix C
- Example 4: Appendix D
- Example 5: Appendix E

In the sections that follow, the results of each task in this Activity are presented. Wherever applicable, the quality methods, criteria and tolerances are referenced and discussed.

3. Results: Task 2.2

In order to discover and evaluate the appropriate datasets for specific purposes, good quality metadata is essential. The activities carried out within the context of this task sought to improve the accessibility and discoverability of quality metadata.

3.1. Methods

The results and findings of this task were achieved through the following methods:

- Conducting meetings between various European partners to explore the problem context, country-specific infrastructures and intended outcome of this task.
- Email contact with the European Data Portal to assess existing and planned infrastructure changes.

3.2. Results

During the meetings between European partners from various countries, experts expressed the fact that most countries implemented INSPIRE and, therefore, all have datasets which are compliant with Technical Guidelines v 2.0. Additionally, all metadata files associated with these datasets are published in a country's respective National Catalogue. In Europe, most of these National Catalogues are made using Geonetwork and will be published as an Open Geospatial Consortium (OGC) API Catalogues Interface.

As is referenced in the link, the EDP supports INSPIRE Geoportal Services. When contacted, the European Data Portal responded with the following remarks:

1. The European Data Portal already connect to the National Geo Catalogues via CSW and harvest the metadata for datasets, dataset series and services themselves.
2. The European Data Portal do already have the mapping required between DCAT and INSPIRE profiles.

This is, therefore, no need to develop specific tooling for publishing service metadata on the EDP based on the existing connection between the national infrastructures (National Geoportals) and EDP because harvesting this metadata is already done from the National Geoportals via CSW. This, therefore, completes this task for all intents and purposes.

It should be noted that some of the work which supported this task in support some European partners with required transformation was existing work from SEMIC. Indeed, this project has already developed a tool which converts files from the INSPIRE profile to DCAT and GeoDCAT. This tool can be found here and is based on this schema. The EDP claims to do this transformation while harvesting from National Geoportals but this work would support partner countries with automatic transformation should this be internally required. Example outputs for both INSPIRE records and an example DCAT-AP output is presented in Appendix A and B respectively.

3.3. Follow-up Actions

As a means of performing a quality check of this connection to support the finalization of this task, a follow up action which is currently being carried out is to check if all metadata files from the use cases in this project are available through the EDP as projected. This task is preparing an excel sheet with all EDP links to these files to support the accessibility to the descriptions of this information and enable data sharing and re-use between sectors and ecosystems within the context of this project.

4. Results: Task 2.3

The activities carried out within the context of this task sought to investigate the existing national profiles of metadata with the aim of subsequently using the profiles for the provision of content discovery and evaluation services. The sandbox triple store environment used to complete this task is available here: https://data.pldn.nl/geoe3task23/-/overview

4.1. Methods

The results and findings of this task were achieved through the following methods:

- Transformation of linked data metadata records as first step to the explore task.
- Carrying out research to define the method through which to best enrich existing metadata profiles to support better semantic search across records.
- Implement method as proof of concept a set of European datasets.
- Document lessons learned and the ways in which the method defined may or may be suitable for other European profiles.

4.2. Results

Metadata profiles from various European partners representing datasets associated with addresses and buildings were identified and manually transformed to linked data. These profiles where then uploaded to an instance of a TriplyDB triple store maintained by Platform Linked Data Nederland for demonstration purposes. The following steps were then carried out to demonstrate the potential of enhancing semantic search through metadata profile enrichment.

1. Once transformed to linked data and before uploading to the triple store, each metadata record was enriched with a set of appropriate keywords by adding this set to the existing keyword list. This list of additional keywords were performed manually, highlighting an area for improvement in following iterations of this task. An example of an enriched profile can be seen in Appendix D.
2. The new enriched records were uploaded to the triple store, combined into a single graph including the graph relating to the GEMET thesauri and an Elasticsearch service was started on this dataset. The starting of this service is easily done within the TriplyDB interface and does not require any coding or scripting knowledge. The simplicity of using this tool can be tested here: https://data.pldn.nl/geoe3task23/GeoE3-Digital-Assets/search/GeoE3-Digital-Assets-1
3. For demonstration purposes, a data story was also created within the triple store where, based on predefined SPARQL queries and showcases how metadata records could be queried based on a common keyword and returned as a table list highlighting all European datasets which may have similar information contained within. This data story can be found here: https://data.pldn.nl/geoe3task23/-/stories/geoe3-semanticsearch-task2-3

Overall, the demonstration environment highlights the usefulness of enriching metadata records in order to support a more in depth semantic search of metadata records of datasets provided across a wide range of European partners. As such, the following conclusions can be made:

1. At minimum, it is necessary to produce a standard metadata profile where it is compulsory to include an extensive list of keywords using the GEMET profile. This extension of keywords should include the data

model associated with the dataset in question but can be extended to include themes and topics with which the dataset can be associated.

2. Because metadata records for datasets are already being delivered, it is necessary to support the enrichment of these existing records according to this standard profile. This should be automated and applied by partner countries without the need for custom code development.

The demonstration environment was also developed using a triple store instance that is generally available for prototype purposes. For developing a production environment, a different set of tools or proprietary licenses to the tools used here are necessary.

4.3. Follow-up Tasks

This explore task highlights the potential that this enrichment has for better semantic search. As noted in the results section, the follow activity to this task is to develop an automated ETL and enrichment step to support partner countries with enrichment of existing profiles.

4.4. Tooling

The following section provides a list of tools used in carrying out this task. Please note, not all tools documented here are intended for production use and might require some licensing or further development should a production environment be desired.

1. Instance of TriplyDB triple store. A triple store instance is available for prototype use on request through Platform Linked Data Nederland (PLDN).
2. Although no automation was done for this task, an automated transformation script for the transformation of records to linked data as well as an automated enrichment step could be developed for this purpose.

5. Results: Task 2.4

The activities carried out within the context of this task sought to investigate mechanisms for making data discover services available through internet search engines such as Google.

5.1. Methods

The results and findings of this task were achieved through the following methods:

- Perform a metadata service mapping by providing a mapping from a DCAT profile to a SDO profile. This mapping should be made available to all partner countries.
- Carrying out research to define the method through which to make metadata for datasets findable through the Google search engine. The research conducted was done within the context of Kadaster.
- Implement the method as a proof of concept on Dutch datasets.
- Investigate and document the lessons learned for this task and the ways in which this method may or may not be suitable for other European partners.

5.2. Results

The results presented here are the results of following a number of steps, each of which are outlined as follows. The metadata records being used in the process that follows can be found in the triple store maintained by the Kadaster Data Science Team here: https://data.labs.kadaster.nl/pdok/metadata/.

5.2.1. DCAT-SDO Mapping Definition

Based on the results of Task 2.2 where a required DCAT-AP profile was defined to support EDP connectivity, it was necessary to define a mapping of this profile to the schema.org specification in order to support the search engine findability requirements for Google. Elements of the GeoDCAT profile was also included in this mapping. This mapping was delivered and is included in Appendix E.

5.2.2. Implementation of ETL

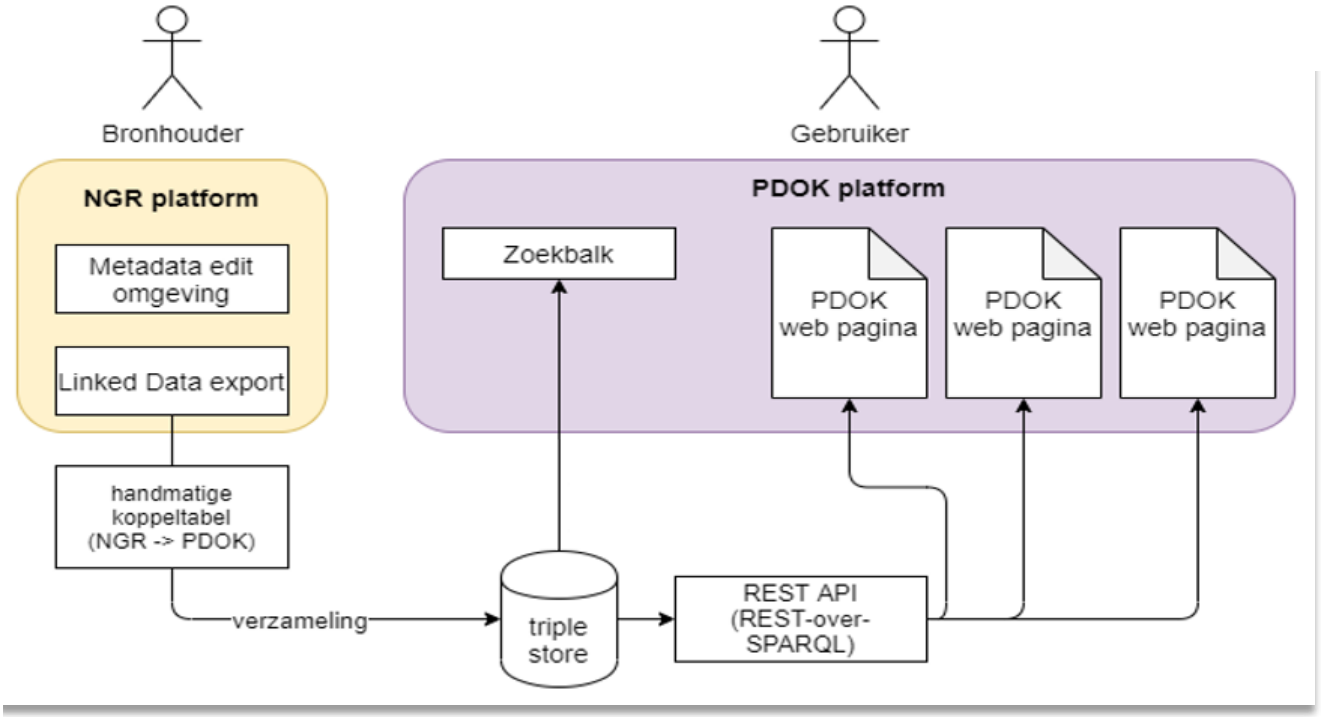
In order to research the manner in which search engine optimization for datasets could be performed in practice, an explore task was initiated within Kadaster based on a set of Dutch metadata records maintained by the Dutch National Georegister (Nationaal Georegister) and associated with Kadaster. The Dutch National Georegister (NGR) is the national source for all geospatial datasets in the Netherlands and it is through this catalogue that all metadata records are produced. As one of its responsibilities, Kadaster is responsible for the maintenance of PDOK (Public Services on the Map) which serves as a catalogue for all geoinformation related to the Dutch government. Because Kadaster has responsibility for maintaining this catalogue, it also has easier access to the infrastructure of PDOK which was necessary to carry out this explore task but the source of the metadata records stored on PDOK is always NGR and, therefore, are always updated by NGR in real time. As such, the metadata records for relevant geospatial datasets related to the Dutch government were identified from NGR itself but these records were transformed through the following process and integrated into PDOK for greater flexibility in exploring the results of this transformation for findability. The ETL process which first transforms the metadata records to the schema.org profile has a number of steps:

1. Scrape HTML pages of NGR for PDOK-NGR relations for metadata records. This association is done using a coupling table which highlights the associations between PDOK datasets and the associated dataset in NGR. For the most part, this is done based on matching GUIDs.

2. Import NGR data from NGR's RDF API into TriplyDB based on identified associations, the triple store instance used for this explore task.
3. Transform NGR data for Google compliance with SPARQL Construct queries based on the mapping provided in the previous step of this task.
4. Validate the transformed data using SHACL to ensure compliance to the metadata profile mapping defined.
5. Reserialize transformed data into JSON-LD files.
6. Upload JSON-LD files as assets in the TriplyDB instance.

Once the relevant transformations have been applied, a REST-over-SPARQL API imports this transformation into the HTML of each associated dataset page in PDOK. This ETL is currently automatically performed every 24 hours based on an open source script which ensures that any updates in the NGT metadata records associated with PDOK pages are delivered as close to real-time as possible. The following figure provides a visual overview of this process.

Figure 1. ETL Process for Including SDO Metadata in Webpages of Datasets on PDOK



It should be noted that this process is recognized to be somewhat specific to the internal processes at Kadaster. However, overall this process makes use of generic tooling and open source scripting which could be offered to any partner wishing to perform this implementation. This tooling includes:

- The use of a triple store of which any could be used in practice. For partners wishing to implement this ETL as an explore task, the PLDN triple store which is also a TriplyDB instance, is available for use as a prototype implementation. For more extensive and ongoing use, an alternative would need to be found per partner.
- The ETL script could be made available as open source.
- The SPARQL Construct queries used to perform the transformation from DCAT to SDO could also be made available.

It is also important to note that, while it seems that transforming metadata records and implementing these back within the NGR environment seems to be more appropriate in this task, the NGR infrastructure, which

is made available through the Geonetwork proprietary software, made this difficult in practice. Because most partners make use of this software in the delivery of their own National Geoportals as outlined in the results of task 2.2, a similar approach as the one made here might be necessary. The feasibility and appropriateness of this should be explored and discussed.

5.2.3. Rich Results Test

Once the transformed metadata records had been appropriately integrated into the metadata of the HTML page of each dataset on PDOK, a validation step was required in order to ensure the quality of this metadata and the ability of the search engines to index these pages to the fullest extent. In all cases, the metadata included in the HTML pages was validated by the Rich Results Test and found to be, in principle, indexable by Google. As such, this process achieved the desired quality in its results as defined by the quality criteria and tolerances in the product description. Ongoing management of the quality of the metadata embedded in HTML pages for indexing can be done by using Google's Search Console functionality. This environment offers users the ability to identify dataset pages were poor metadata quality leads to a lack of indexing and highlights the appropriate fixes based on schema.org specifications. This tool is openly available to all partner countries.

5.2.4. Google Dataset Search

A further validation step, and one that was not initially defined in the product description, is to ensure that the datasets in question are actually being indexed by Google in practice. This testing can be done through making use of Google's Dataset Search and assessing the results for the presence of the dataset in question. Whilst all metadata transformed and integrated into PDOK pages performed well in the Rich Results Test, the PDOK pages which include this metadata in their HTML are not being indexed by Google.

5.3. Follow-up Actions

As noted, whilst the metadata transformed through the ETL process is of a high quality and fully validated by the appropriate Google testing tools, it seems that these pages are not currently being indexed by Google. A follow up task to identify why this is the case is being performed by Kadaster and an update should be available in due course. The outcome of this will be documented in order to prevent a similar outcome for partner countries.

5.4. Tooling

The following section provides a list of tools used in carrying out this task. Please note, not all tools documented here are intended for production use and might require some licensing or further development should a production environment be desired.

1. Instance of TriplyDB triple store. This instance is maintained by the Kadaster Data Science Team as is not available for external use but a similar instance is available for prototype use on request through Platform Linked Data Nederland (PLDN).
2. Custom script to perform the ETL process outlined above.
3. SPARQL Construct queries and SHACL validation scripting to ensure appropriate metadata profile mapping from DCAT to SDO has been performed.
4. Google Search Console to manage metadata quality and overview of indexability of metadata for a particular website.

6. Conclusion

This report highlights the successful completion and documentation of the tasks defined by the product description. Indeed, all but one deliverable was achieved in carrying out this report. The final deliverable, a demonstration of the results in a workshop, is still possible and can be arranged as a follow up task to this report.

7. Appendices

APPENDIXES A, B, C AND D ARE LISTED HERE ONLY AS AN EXAMPLE FOR DEVELOPERS WHO HAVE INSTALLED A TEXT EDITOR (FOR EXAMPLE VISUAL STUDIO CODE). THEY CONTAIN XML'S AND RDF'S (TRIPLES), APPENDIX E IS NOT ONLY FOR DEVELEOPERS. THIS IS AN EXCEL FILE.

APPENDIX A: XML FILE IN INSPIRE PROFILE

Example 1: An XML file in INSPIRE Profile (TG 2.0) For Spain

spaignLLM.xml

APPENDIX B: RDF FILE IN DCAT-AP PROFILE

Example 2: A RDF file in DCAT-AP profile for Spain

spaignLLM_DCAT.rdf

APPENDIX C: RDF FILE IN GEODCAT-AP PROFILE

Example 3: A RDF file in GeoDCAT-AP profile for Spain

spaignLLM_GeoDCAT
.rdf

APPENDIX D: ENRICHED DCAT-AP PROFILE TO SUPPORT SEMANTIC SEARCH

Example 4: An enriched DCAT-Ap profile to support semantic search

bag_metadata_enrich
ed.trig

APPENDIX E: GEODCAT-AP TO SDO MAPPING

Example 5: GEODCAT-AP to SDO mapping in Excel format (includes INSPIRE-DCAT-AP mapping)

SDO_GeoDCAT_ISO-
Mapping_Final.xlsx