A SKOS Taxonomy of the UN Global Geospatial Information Management Data Themes

Beyza Yaman,[1] Kevin Thompson[2] and Rob Brennan[1]

[1] ADAPT Centre, Dublin City University, Dublin, Ireland
[2] Ordnance Survey Ireland, Dublin, Ireland
{beyza.yaman,rob.brennan}@adaptcentre.ie, kevin.thompson@osi.ie

Abstract. Complex data domains increase the difficulty of structuring, sharing, discovering and governing information. For the geospatial domain common models such as INSPIRE have been established in the European Union. The United Nations initiative on Global Geospatial Information Management (UN-GGIM) draws together national and regional capacities. Interoperability is the main principle behind these initiatives. Nonetheless there is a lack of published research to date on mapping agency geospatial linked data leveraging the UN-GGIM taxonomy of information management data themes. Thus, we have identified use cases and defined a Simple Knowledge Organization System (SKOS)[3] taxonomy expressing the UN-GGIM data themes for national spatial infrastructure. This has been applied in a metadata generation and reporting tool for Ordnance Survey Ireland (OSi) which underpinned improved governance and reporting infrastructure in OSi. This demonstrated the contribution of Semantic Web technology to spatial data governance as well as its importance for data publishing. This paper presents a documented open license SKOS taxonomy for the UN-GGIM data themes that follows Linked Data best practices. It provides a set of three use cases, an overview of UN-GGIM theme definitions and an example application of the taxonomy for deployment in OSi for DCAT metadata generation and data publishing pipeline reporting.

1 Introduction

Geospatial data is essential in part due to its importance in social, economic, and environmental policy formation and decision making. In the geospatial data domain, large organizations -especially, ones working at a National level- have to confront high data heterogeneity due to the need to collect, analyze and share information within national, regional and global policy frameworks. Interoperable aggregation and reporting of this geospatial statistical information is key to avoid data quality problems and to streamline data transfers and management. Interoperability requires common vocabularies, models, meta-data and interfaces for creating, reporting on and curating geospatial statistical data.

In these circumstances, developing standards and norms becomes crucial to providing a coherent service infrastructure and to create meaningful relations

among data integrated from multiple organizations. In order to boost countries' geospatial data activities, the United Nations has developed a framework called the Global Geospatial Information Management (UN-GGIM) Data Themes [4]. These themes are a minimum set of concepts to be used to label datasets in order to enable interoperability between national mapping agencies. They form a foundation to support global geospatial information management and an integrated geospatial information framework to strengthen geospatial information sharing among other global initiatives[4]. They are similar in concept to the EU INSPIRE data themes but with global scope. However, to date it was seen that no paper was published establishing a geo Linked Data approach to supporting UN-GGIM metadata and no documented UN-GGIM data themes use cases for Semantic Web technology.

Thus this paper explores the research question of to what extent a geo Linked Data approach could increase the interoperability of geospatial data by modelling UN-GGIM concepts and using standardized tools to report on national geospatial data production in terms of UN-GGIM data themes in an understandable manner. In order to solve this problem, we created a taxonomy for the UN-GGIM Data Themes to classify the data and provide a meaningful relationship among the data. The contribution of this paper is a new SKOS vocabulary [3] that can be used to enhance geospatial datasets with the UN-GGIM data themes concepts to increase the interpretability of the data.

The remainder of this paper is structured as follows: Section 2 discusses Related Work and Section 3 describes the Use Case for Ordnance Survey Ireland (OSi) deployment. Section 4 introduces the UN-GGIM Data Themes and Section 5 demonstrates the UN-GGIM taxonomy development and data theme solution for the OSi National Map dataset (Prime2). Finally, the paper provides conclusions in Section 6.

2 Related Work

The INSPIRE Data Themes[3] are the first set of classification themes implemented for geospatial data within the scope of the INSPIRE Directive in the EU. An RDF model version has been developed for the description of the data themes and provided to end-users such as national mapping agencies so that appropriate descriptive metadata could be generated for the datasets. The model consists of 34 INSPIRE data themes which makes the model more detailed than UN-GGIM themes.

The UN-GGIM Secretariat has provided an interactive web interface implemented by the ArcGIS tool employing the UN-GGIM data themes[4]. The interface demonstrates the use case examples for statistical reporting based on UN-GGIM data themes and how they can be used in the context. The use cases include population distribution, transport network maps integrating the stan-

[3] https://inspire.ec.europa.eu/theme
[4] https://www.arcgis.com/apps/Cascade/index.html?appid=4741ad51ff7a463d833d18cbcec29fff

dardised, fundamental data from national mapping agencies, statistical offices and other institutions.

The W3C's data catalog vocabulary standard, DCAT[1], allows the description of data themes and they are proposed as recommended properties by the DCAT application profile (DCAT-AP) for describing a resource (dataset, data service or dataset distribution). The DCAT vocabulary provides two different properties for describing data theme information: the `dcat:themeTaxonomy` property for the schema to be used for data themes, and the `dcat:theme` property for the specific "theme/category" from the available theme options[5]. Thus a DCAT record can be used to describe a dataset and associate it with one or more data themes in one or more data theme taxonomies.

3 Use Cases

Ordnance Survey Ireland (OSi) is the national mapping agency of Ireland that produces geospatial data. The data is first captured through surveying the land using aeroplanes and then adjusted by surveyors if necessary. This digital data is typically in image or point cloud formats and later converted and stored in an Oracle Spatial and Graph database using the Prime2 model. Prime2 is the object-oriented spatial model of over 50 million spatial objects tracked in time followed by conversion for printing as cartographic products or data sales and distribution by OSi. After going through transformation phases the data is provided in different formats either with open access or with a cost to private end-users, stakeholders, or other governmental institutions.

Acting as a source of spatial data, OSi has to comply with several statistical regulations such as National, European Directives like INSPIRE and global statistical reporting like UN agencies. OSi reporting use cases must take into account three different aspects: i) Classification of the data: The end-users need a common understanding of key characteristics of the data they are consuming. For instance, an insurance company is interested in subdatasets related to flood risk management, or a post office is interested in addresses. ii) Reporting to stakeholders: Institutions like the United Nations and the Organisation for Economic Co-operation and Development (OECD) demand the standardization of geospatial statistical data according to the field of geospatial information management. It is also mandatory for INSPIRE Directive states, to share the country spatial data and metadata created for spatial data sets and services corresponding to the themes through interoperable infrastructures. iii) Dataset metadata generation: Internally the departments need to catalog their data and provide information about suitability for intended uses. This allows high level monitoring of the data including quality scores or provenance. The lack of providing the data and metadata in an interoperable and standardized way can result in failures in the system, reduced co-operation with partners and fines or other penalties for non-compliance.

[5] https://www.w3.org/TR/vocab-dcat-2/#Property:catalog_themes

4 UN-GGIM Data Themes

The UN-GGIM created strategic pathways[6] proposing implementation and custodianship guidelines for best practices in collection and management of integrated geospatial information to establish a global geospatial data framework. This way UN-GGIM supports and guides the geospatial data infrastructure of the member countries through the approach, content, rationale, options and considerations, principles that align with actions, and sample outcomes for comparison [4]. UN-GGIM Data Themes are the set of prioritized national data themes, aligned to the globally endorsed fundamental geospatial data themes. 14 Global Fundamental Geospatial Data Themes are proposed to be used by the strategic pathways to classify geospatial data. The 14 data themes from UN-GGIM are as follows [UN-GGIM]: Global Geodetic Reference Frame (GGRF), Addresses, Buildings and Settlements, Elevation and Depth, Functional Areas, Geographical Names, Geology and Soils, Land Cover and Land Use, Land Parcels, Orthoimagery, Physical Infrastructure, Population, Distribution, Transport Networks, and Water. Theme descriptions can be found in the Global Fundamental Geospatial Data Themes document[2].

Data practitioners can follow these strategic plans and data themes to design, develop, maintain a high standards, high quality and sustainable geospatial data infrastructure. On the other hand, the data themes also allow the data to be analysed and provide statistical results w.r.t. the specific data theme such as population distribution in a specific area, e.g. for policy development. They are specifically important for organizing a country's geospatial, statistical and other information. Fundamental data themes (e.g. transportation) are required for a broad range of decision-making applications, or application data themes (e.g. flood models) required for specific studies; and socio-economic data themes that provide demographic information, such as census and population data [7]. Integrating data themes information into a data catalogue allows end-users to decide the suitability of the dataset for their purpose.

5 Taxonomy Development

This section describes the implementation of the taxonomy for the OSi use case. UN-GGIM Data Themes Taxonomy was designed and implemented as a Linked Data controlled vocabulary to provide common and standardized definitions for managing the data. SKOS concept taxonomy was selected as the most appropriate modelling language for this task and it corresponds to the DCAT data theme property requirements. A 3 step methodology was followed to create this vocabulary. First, UN-GGIM Data Themes were encoded to RDF to generate the taxonomy in Linked Data format. Second, error checking process was performed to validate the data. Third, the generated vocabulary was published on the web.

[6] https://ggim.un.org/IGIF/part2.cshtml
[7] https://ggim.un.org/IGIF/documents/SP4-Data_10Jan2020_GLOBAL_CONSULTATION.pdf

```
unggim-dt:Buildings-Settlements
        a          skos:Concept ;
        skos:broader      unggim-dt:DataTheme ;
        skos:prefLabel    "Buildings and Settlements"@en;
        skos:note "A Building refers to any roofed structure permanently
constructed or erected on its site , for the protection of humans, animals
,things , or the production of economic goods. Settlements are collections
of buildings and associated features where a community carries out
socio-economic activities.".
```

Listing 1.1. Example Data Theme Snippet

Each proposed theme in the UN-GGIM data themes was defined as a SKOS concept class. All the classes in the data theme vocabulary were collected under a generic Data Theme class which was defined as a top concept of the scheme and the themes were described as the narrower concepts of this concept. The created vocabulary is available online and it can be seen partially in Listing 1.1. Since there is no direct relation between concepts they have been classified under the top concept of data theme[8].

OSi have identified the UN-GGIM data themes as an important framework for reporting to their stakeholders while using Prime2 dataset. A key issue was how to map or present the contents of the 50 million spatial objects captured in Prime2 as data themes. This would enable adding additional meaning that not only humans but also machines could use and interpret the data. In turn this would enable Prime2 data quality reports to be generated for a specific data theme or group of data themes. At first glance it seems that Prime2 (as a universal model) spans all UN-GGIM data themes. However within Prime2 there are different types of spatial classes totaling to 34, e.g. `Building` and `Locale` datasets are spatial object classes in the Prime2 dataset so each class can be considered to form a sub-dataset in Prime2. Our approach has been to create a mapping between Prime2 spatial objects and UN-GGIM data themes. Then a DCAT record can be created for each sub-dataset and this in turn enables it to be assocated with one or more data themes. This enhanced data catalog for OSi data assets enables us to write queries for parts of Prime2 that correspond to specific data themes and to associate the quality data for those spatial objects with the data theme.

It was important to take into consideration what type of processes will be performed upon these datasets and decide the relevant themes for each class. The created vocabulary is used to organize Prime2 dataset. The datasets (spatial classes) were enriched by one or more than one label using the UN-GGIM themes. The preliminary version of the assignments are performed as in Table 1. There are 3 options for a relationship between a Prime 2 spatial object class and a data theme category: partial match, full match (yes in table), no match (no in table). A full match means all the instances of a class are relevant to the data theme. A partial match means some instances of the class are relevant to the data theme. No match means it does not have any instances related to that data theme.

[8] https://linkeddataops.adaptcentre.ie/vocabularies/unggim-data-themes

Table 1. Example OSi Prime2 Sub-Datasets to UN-GGIM Data Theme Mapping

	Addresses	Buildings and Settlements	Geographical Names	Transport Networks	Water
Boundary Area	partial	partial	yes	no	no
Building	yes	yes	partial	no	no
Locale	yes	yes	yes	no	no
Site	partial	yes	yes	no	no
Water	no	no	partial	yes	yes
Way	yes	no	partial	yes	no

A catalog description (Listing 1.2) was created for the datasets including each subdatasets existing in the pipeline. This enabled us using a standard model to describe the metadata features from multiple vocabularies and having different values throughout the time. This description was used as a metadata repository to enable effective data governance controls to store and track the data practically. Listing 2 shows a DCAT description example of `Building` dataset which was enhanced by `Buildings-Settlements` theme relation. Beside the advantages adding to the data practitioners, the themes add value to the data in the international level. For instance, the data with transport network theme allows to track the transportation ways around Europe and between the continents.

The dashboard (Fig. 1) presents the reporting features based on the OSi data catalog which includes data theme, provenance and data quality metadata about each dataset and sub-dataset. This is a part of the LinkedDataOps project which aims to manage the data in a useful way [5]. The dashboard page has several filters on the left side of the page which allows users to click interactively. Users can pose various queries with different views by clicking on the filters and visualize the dataset relations semantically. The description of the data themes allowed an easy classification and exploration process of the datasets in the pipeline. Thus, while the classification of subdatasets helps OSi to better understand, classify and provide the data, it also helps the end users to consume the parts of data they need. This approach brings convenience not only to the data producer but also to the user of the data.

```
<http://ontologies.adaptcentre.ie/dataset-hierarchy#BUILDING>
    a dcat:Dataset ;
    a prov:Entity;
    dc:title "Building"@en ;
    dc:description "Building is a permanent roofed construction ,
currently or formerly used or intended for shelter . The construction
must have permanent foundations . A work under construction , is included
as a Building , if it is apparent that , on completion , it will meet the
definition for Building . A structure that is identifiable as having once
been a Building but which no longer has a roof , is included as a
Building ."@en ;
    dct:created "2019-01-09"^^xsd:date ;
    dct:modified "2020-01-09"^^xsd:date ;
    dcat:theme unggim-dt:Addresses ,unggim-dt:Buildings-Settlements .
```

Listing 1.2. Part of OSi Data Catalog

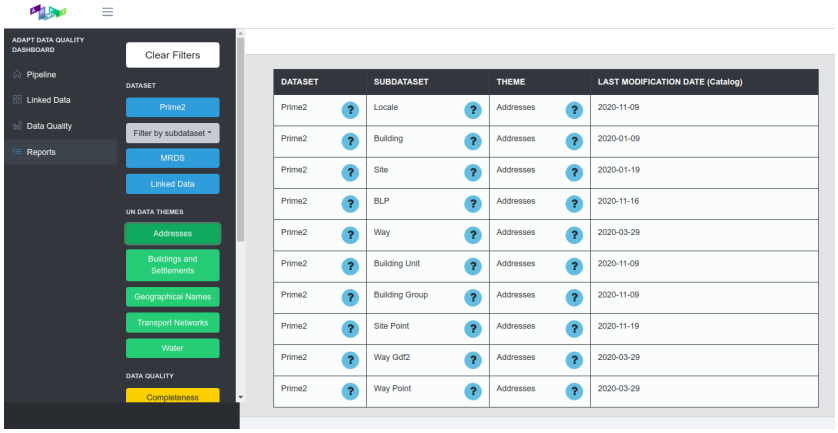

Fig. 1. OSi Dasboard for Data Catalog

6 Conclusions/Future Work

A UN-GGIM data theme taxonomy was implemented in order to increase interoperability, facilitate the integration of non-centralized data, and improve search facilities with specific data themes. It is anticipated that using this taxonomy will be used as a means of integrating geospatial data in the national and international level and will address global challenges and help build international awareness. As a future work, we would like to map the UN-GGIM data themes to INSPIRE data themes in order to improve the interoperability further.

Acknowledgement This research received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No. 801522, by Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology [grant number 13/RC/2106] and Ordnance Survey Ireland.

References

1. R. Albertoni, D. Browning, S. Cox, A. G. Beltran, A. Perego, and P. Winstanley. Data catalog vocabulary (dcat) - version 2. *World Wide Web Consortium*, 2020.
2. C. Hadley. The global fundamental geospatial data themes journey. *United Nations Committee of Experts on Global Geospatial Information Management*, 2018.
3. A. Miles and S. Bechhofer. Skos simple knowledge organization system reference. *W3C recommendation*, 2009.
4. UN-GGIM. The Global Fundamental Geospatial Data Themes. https://ggim.un.org/documents/Fundamental%20Data%20Publication.pdf, 2019. Accessed on 05.12.2020.
5. B. Yaman and R. Brennan. Linkeddataops: linked data operations based on quality process cycle. *Proceedings of the EKAW 2020 Posters and Demonstrations Session, Globally online & Bozen-Bolzano, Italy, September 17*, 2020.