Automatic publication of Open Data from OGC services: the use case of TRAFAIR project

Javier Nogueras-Iso*, Héctor Ochoa-Ortiz*, Manuel Ángel Jañez*, José R. R. Viqueira†, Laura Po‡ and Raquel Trillo-Lado*
*Universidad de Zaragoza, Spain
Email: {jnog,719509,731321,raqueltl}@unizar.es
†Universidade de Santiago de Compostela, Spain
Email: jrr.viqueira@usc.es
‡Università degli Studi di Modena e Reggio Emilia, Italy
Email: laura.po@unimore.it

***Abstract*—This work proposes a workflow for the publication of Open Spatial Data. The main contribution of this work is the automatic generation of metadata extracted from OGC spatial services providing access to feature types and coverages. Besides, this work adopts the GeoDCAT-AP metadata profile for the description of datasets because it allows for an appropriate crosswalk between the annotation requirements in the spatial domain and the metadata models accepted in general Open Data portals. The feasibility of the proposed workflow has been tested within the framework of the TRAFAIR project to publish monitoring and forecasting air quality data.**

Keywords–Environmental data; Open Data; GeoDCAT-AP; metadata; geospatial services; OGC.

I. Introduction

TRAFAIR (Understanding traffic flows to improve air quality),[1] is a European project co-financed by the Connecting Europe Facility of the European Union (Project Nr. 2017-EU-IA-0167), whose main objectives are the monitoring of air quality in urban areas, and the development of forecasting air quality services based on meteorological predictions and urban traffic flows [1]. Also, the project aims to publish monitoring and forecasting air quality data as Open Data and to develop client applications to make both citizens and public administrations aware of the air quality and the responsible use of private transport.

To facilitate the visualization and download of monitoring and forecasting data, project partners have chosen the use of Geoserver software, which facilitates the setting up of servers accessible through the standardized service interfaces compliant with Open Geospatial Consortium (OGC) specifications: Web Mapping Services (WMS) for visualization, Web Feature Services (WFS) for the download of feature data, and Web Coverage Services (WCS) for the download of coverage data. Figure 1 shows a layered architecture with the data and service components managed in the TRAFAIR project.

However, the simple publication of OGC services cannot be considered as Open Data publication. To make data really accessible as Open Data, we need to register datasets in official Open Data portals. Furthermore, the publication of datasets in the European Data Portal (EDP)[2] is a requirement of the project. To register as Open Data the TRAFAIR outcomes, the first step has been to select an appropriate metadata profile compliant with the metadata models accepted in the Open Domain context. Taking into account the spatial character of data managed in TRAFAIR, we have adopted the GeoDCAT-AP metadata profile [2]. GeoDCAT-AP is a metadata profile that extends DCAT-AP, a metadata profile designed by the European Commission to describe public sector data. GeoDCAT-AP metadata properties have been designed to assure compliance with the metadata requirements of the European INSPIRE directive for establishing a spatial information infrastructure in Europe [3].

On the other hand, to minimize the effort of creating metadata and the registration of these data in Open Data portals, we decided to automate this process employing a software that retrieves the capabilities of OGC services and converts this information into metadata records that are later ingested in a CKAN-like Open Data server. CKAN [3] is the most widely used Open source platform to support Open Data portals, which includes the necessary plug-ins to exchange metadata in RDF format (the serialization format used for GeoDCAT-AP).

The objective of this work is to describe the workflow that we have proposed for the publication of Open Spatial Data integrating the automatic generation of GeoDCAT-AP metadata. The remainder of this paper is structured as follows. Section II introduces background information on the GeoDCAT-AP metadata model. Section III describes our proposed workflow for the publication of Open Spatial Data. Section IV describes the feasibility of the application of the proposed workflow in the cities of Modena, Santiago de Compostela, and Zaragoza. Section V reviews related works in the literature. Lastly, this paper ends with some conclusions and an outline of future work.

II. GeoDCAT-AP: a metadata profile for Open Spatial Data

ISO 19115 is the international standard for geographic metadata proposed by the International Organisation for Standardization (ISO) [4], which has been widely adopted during the last decade in the geographic information community in both public and private sectors.

However, in the Open Data domain, more general and simple metadata schemas are needed to facilitate the publication of datasets from different disciplines in the same metadata

[1] http://trafair.eu/
[2] https://www.europeandataportal.eu/en
[3] https://ckan.org/

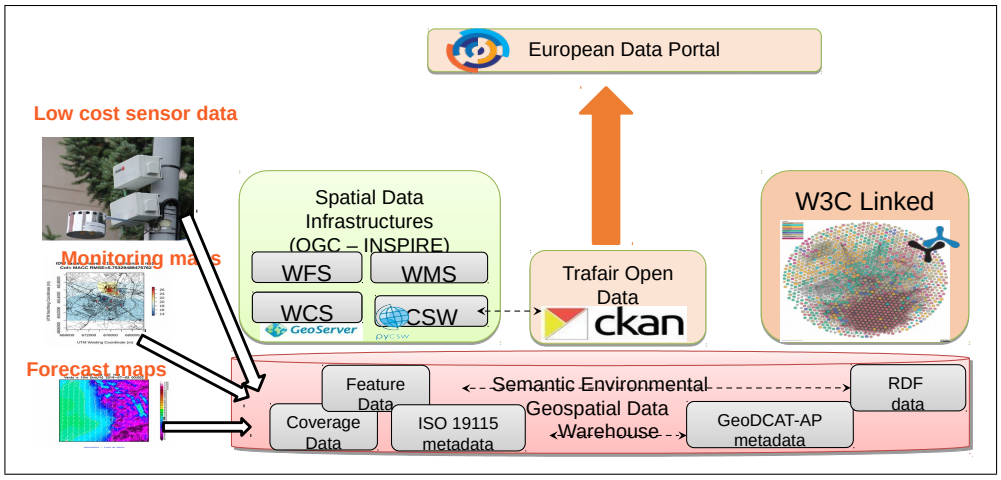

Figure 1. Architecture of data and services components in TRAFAIR project

repository. DCAT is the acronym for W3C's Data Catalogue vocabulary) [5] and can be considered as a basic and general core of metadata properties shared by the different metadata schemas used in various Open Data initiatives. In the case of Europe, the European Union proposed in 2013 DCAT-AP [6], a specification based on DCAT for describing public sector datasets in Europe. Compared to DCAT, DCAT-AP provides stricter definitions of catalogs, datasets, distributions, and other objects.

As mentioned in the introduction, within the context of this project, we have selected GeoDCAT-AP v1.01 [2]. This extension of DCAT-AP [6] was designed for the description of spatial data and its metadata properties have an exact mapping with the main elements of ISO 19115 metadata. This mapping assures the transformation of GeoDCAT-AP records into equivalent ISO 19115 metadata records compliant with INSPIRE requirements [7].

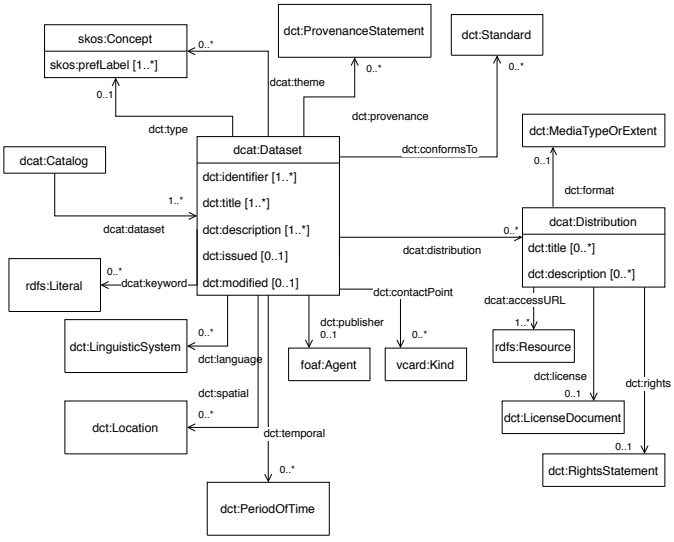

Figure 2. Entities and properties used from GeoDCAT-AP

The description of datasets according to GeoDCAT-AP is mainly focused on providing information about three main entities: a *Catalog* that is published through an Open Data portal containing *Datasets* and the associated *Distribution* forms of each dataset. Besides, GeoDCAT-AP makes a distinction between core and extended properties. The core set is the selection of DCAT-AP metadata properties that have a direct binding with ISO 19115 and INSPIRE metadata. The extended set is a superset of the core set, including additional metadata properties to provide a complete binding with ISO 19115 and INSPIRE metadata. In some cases, these additional properties belong to other metadata vocabularies. In other cases, although the properties belong to DCAT-AP, they are classified as extended because they only provide a partial binding with ISO 19115 and INSPIRE.

Figure 2 shows a UML diagram with the properties from GeoDCAT-AP that are needed for describing datasets and distributions in TRAFAIR. Most of these properties belong to the core set of GeoDCAT-AP. The only exceptions are the *dct:type* property of *Datasets* and the *dct:description* property of *Distributions*. *dct:type* is employed to indicate whether the described resource is a dataset or a dataset series. *dct:description* allows the description of the spatial resolution of associated distributions.[4]

III. Proposed Workflow for the Publication of Open Spatial Data

Figure 3 shows an activity diagram with the main five steps of the workflow that we have proposed for the publication of Open Spatial Data. For steps 1, 3, and 4, we have developed software to automate as much as possible the automatic generation and release of metadata. Steps 2 and 5 are accomplished thanks to the use of existing software packages.

The first step is the ingestion of layers in a spatial data warehouse. Geoserver is the software package selected for managing the publication of spatial data layers, either discrete feature data or coverage data. In the context of this project, we have developed specific software in Java and R languages to ingest feature types (supported in a spatial database) and

[4]Although GeoDCAT-AP proposes *rdfs:comment* as a provisional property to fill this resolution information, there is no direct mapping of this property to CKAN fields and we have considered *dct:description* as a valid alternative.

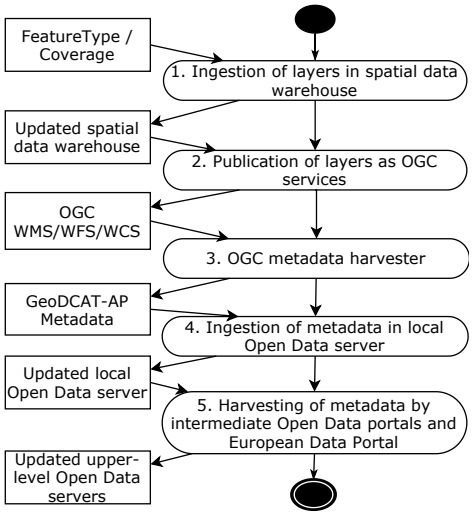

Figure 3. Workflow for publication of Open Spatial Data

coverages in Geoserver through its REST API. [5]. Concerning metadata generation, this software takes care of sending to the REST API the appropriate values for the tags enumerated in the *Geoserver* column of table I.

The second step is the publication of layers as OGC services. This step is directly achieved thanks to the Geoserver software, which provides access to layers through different OGC services. Feature types such as observations retrieved from traffic and air quality sensors may be downloaded through a WFS service. In the case of coverages for air quality monitoring (interpolations of geo-referenced sensor observations) or coverages for predicting quality (the result of applying a Lagrangian model for the dispersion of pollutants called GRAL [8]), a WCS service is used to retrieve these raster data. Beyond WFS and WCS, some layers are also available to perform server-side map rendering using a WMS service.

The third step is the harvesting of metadata from OGC services through its *GetCapabilities* operation. To implement this step, we have developed a Python program that takes profit of OWSLib,[6] a Python package for client programming with OGC web service interface standards and their related content models. This software interacts with the OGC interface, instead of the GeoServer REST API, because we wanted to make this software scalable enough to integrate in the future other layers managed by software packages different from Geoserver. The algorithm behind this software generates a *Dataset* instance for every layer published in WFS or WCS services. In addition, each *Dataset* has at least one associated *Distribution* instance in the form of a link to a WFS or WCS service. In some cases, if the layer is also rendered through a WMS, a second distribution linking to the WMS service is generated. The *OWSLib* column in table I shows the fields retrieved with OWSLib package to generate the corresponding GeoDCAT-AP property.

The fourth step is the ingestion of metadata in the Open Data server of the institution in charge of publishing the air quality data of the local area. As a continuation of the software in the previous step, our Python program transforms the information retrieved in the previous step into a dictionary with the required items to construct a dataset and its associated resources, which are immediately inserted in the CKAN-based local Open Data server through its REST API.[7] The *CKAN* column in table I indicates the tags that are used in this dictionary data structure to generate later RDF metadata based on GeoDCAT-AP.[8]

The final step is the harvesting of metadata in the local servers by regional and national Open Data portals until the EDP finally harvests metadata. This step is beyond the scope of the TRAFAIR project. However, we assume that upper-level portals are based on CKAN technology (or have a similar mechanism for the harvesting of subscribed lower level catalogs). On the one hand, the *ckanext-dcat* plugin of CKAN allows the publication of datasets metadata as RDF in compliance with DCAT-AP vocabularies. On the other hand, the *ckanext-harvest* plugin of CKAN allows us to harvest the contents of different types of catalog sources.

IV. Deployment of Open Data in the cities of Modena, Santiago de Compostela and Zaragoza

Figure 4 shows the deployment of Open Data portals in the cities of Modena (Italy), Santiago de Compostela (Spain) and Zaragoza (Spain). The figure also shows how the local GeoServers are queried with the software described in steps 3 and 4 of the proposed workflow to feed the contents of the local Open Data portals. In addition, the figure shows the Open Data portals at regional, national, and European level that harvest the contents of the local Open Data portals.

In the case of Modena, the Open Data contents managed by the municipal government of Modena (*Commune di Modena*)[9] are directly ingested in the CKAN-based Open Data server provided by the regional government of Emilia-Romagna.[10] The contents of this portal are harvested by the Italian Government Open Data portal (*dati.gov.it*).

In the case of Santiago de Compostela, the Open Data portal is managed by the municipal government of Santiago de Compostela (*Concello de Santiago*).[11] Later, the contents of this portal are harvested by the Spanish Government Open Data portal (*datos.gob.es*).

The case of Zaragoza is more complicated. There is an Open Data portal based on CKAN maintained by the researchers of the University of Zaragoza involved in the TRAFAIR project. However, the Open Data contents of the University are published through a different kind of repository (called Zaguan) based on MARC metadata and accessible through the OAI-PMH protocol. In this case, we had to develop a specific program to upload the GeoDCAT-AP metadata periodically in MARC format at Zaguan portal (see figure 5).[12] Then, these metadata records are harvested by the regional

[5]See https://docs.geoserver.org/latest/en/api/#/latest/en/api/1.0.0/coverages.yaml (API for coverages) and https://docs.geoserver.org/latest/en/api/#/latest/en/api/1.0.0/featuretypes.yaml (API for features).

[6]https://geopython.github.io/OWSLib/

[7]https://docs.ckan.org/en/2.8/api/index.html

[8]The mapping between CKAN fields and RDF properties has been obtained from https://github.com/ckan/ckanext-dcat#rdf-dcat-to-ckan-dataset-mapping.

[9]https://www.comune.modena.it/opendata

[10]https://dati.emilia-romagna.it/dataset?organization=comune-di-modena

[11]https://datos.santiagodecompostela.gal/es

[12]https://zaguan.unizar.es/collection/opendata-investigacion-medioambiental?ln=en

TABLE I. Correspondence between Geoserver tags (contained in the body of a POST request to create a featureType/coverage), layer fields retrieved with OWSLib from a *GetCapabilities* response, CKAN tags (contained in the body of a POST request to create a Dataset), and GeoDCAT-AP properties

Geoserver	**OWSLib**	**CKAN**	**GeoDCAT-AP**
featureType/name, coverage/name	*layerName*	extra:identifier	Dataset/dct:identifier
featureType/title, coverage/title	contents[*layerName*].title	title	Dataset/dct:title
featureType/description, coverage/description *(software in step 1 introduces predefined descriptions according to name patterns)*	contents[*layerName*].abstract	notes	Dataset/dct:description
	("series" for OGC services with temporal dimension, or "dataset" without temporal dimension)	extra:dcat_type	Dataset/dct:type
	(default language proposed in step 3)	extra:language	Dataset/dct:language
	(default INSPIRE data themes and ISO 19115 topic categories proposed in step 3)	extra:theme	Dataset/dcat:theme
(some default keywords are automatically introduced by Geoserver)	contents[*layerName*].keywords	tags	Dataset/dcat:keyword
(computed automatically by Geoserver)	contents[*layerName*].boundingBoxWGS84	extra:spatial	Dataset/dct:spatial
(start date and end date are automatically updated by Geoserver)	contents[*layerName*].timepositions	extra:temporal_start + extra:temporal_end	Dataset/dct:temporal
		extra:issued *(automatically inserted with first ingestion in CKAN)*	Dataset/dct:issued
		extra:modified *(automatically updated with every update of a dataset in CKAN)*	Dataset/dct:modified
	(default provenance proposed in step 3)	extra:provenance	Dataset/dct:provenance
	(default INSPIRE conformance and coordinate reference system proposed in step 3)	extra:conforms_to	Dataset/dct:conformsTo
(contact information is directly introduced by administrators at Geoserver configuration page)	contents[*layerName*].provider.contact. organization	extra:publisher_name	Dataset/dct:publisher
	contents[*layerName*].provider.contact.name + contents[*layerName*].provider.contact.email	extra:contact_name + extra:contact_email	Dataset/dcat:contactPoint
(OGC service URL generated automatically by Geoserver)	*(OGC service URL)*	resource:url	Distribution/dcat:accessURL
featureType/name, coverage/name	*layerName*	resource:name	Distribution/dct:title
featureType/serviceConfiguration, coverage/serviceConfiguration	*("wfs", "wcs" or "wms" according to OGC service type)*	resource:format	Distribution/dct:format
	(default licence proposed in step 3)	resource:license	Distribution/dct:license
	(default rights proposed in step 3)	resource:rights	Distribution/dct:rights
	(default resolution proposed for project datasets)	resource:description	Distribution/dcat:description

government of Aragon, and later by the Spanish Government Open Data portal.

Apart from having the possibility of querying all the datasets contributed by different TRAFAIR partners at the European Data Portal,[13] the TRAFAIR partners have decided to deploy also a central Open Data portal that harvests the contents of the individual portals in each city. This central Open Data portal (its access will be provided through the main site of the project) is also based on CKAN software platform and serves as a unified catalog to have a global perspective of all the Open Data contents contributed by the different project partners.

V. Related work

There are several examples of works trying to crawl the contents of OGC services and automate the generation of metadata items that are later ingested in catalogs compliant with the OGC Catalog Services for the Web (CSW) specification. For instance, Nogueras-Iso et al. [9] proposed a mechanism to derive metadata from the capabilities information returned by OGC services (e.g., WMS, WFS or WCS) and create entries in a catalog of geographic information services. A related solution is the CSW - ISO 19115 community module of GeoServer software [10]. This GeoServer extension allows the browsing of GeoServer layers through a CSW API, but no details are provided about the OGC services providing access to the layers. This extension is also comparable to the harvesting possibility offered by Geonetwork (a software for deploying geographic metadata catalogs) to use the GetCapabilities response of an OGC service (e.g., WMS, WFS or WCS) to generate ISO 19115 metadata for the resources delivered by the service [11]. Another example studying in more detail the layers advertised in a *GetCapabilities* response is the one proposed by Florczyk et al. [12]. This work describes a method for the automatic detection of the orthoimage layers discovered in the *GetCapabilities* responses of Web Map Services, which was used to feed the contents of a virtual catalog of orthoimages.

Concerning the synchronized publication of data and metadata, there are also examples of works trying to define workflows for the joint publication of datasets and metadata. For instance, Gil-Altaba et al. [13] proposed a service framework that used the GeoServer REST API to create a new data store accessible through OGC services, and immediately ingest the associated metadata to this datastore into a CSW catalog supported with Geonetwork software.

The previous works are focused on generating ISO 19115 - compliant metadata. However, for publication of geographic information as Open Data, solutions generating DCAT-based metadata are required. Perego et al. [14] describe uses cases for profile-based content negotiation and publishing metadata on the web where a GeoDCAT-AP API has been developed to transform original content according to ISO 19115 metadata standard into DCAT-AP metadata in various formats. A similar approach is provided through the *ckanext-spatial* plugin of CKAN [15]. This plugin allows harvesting the ISO 19115

[13] https://www.europeandataportal.eu/data/datasets?locale=en

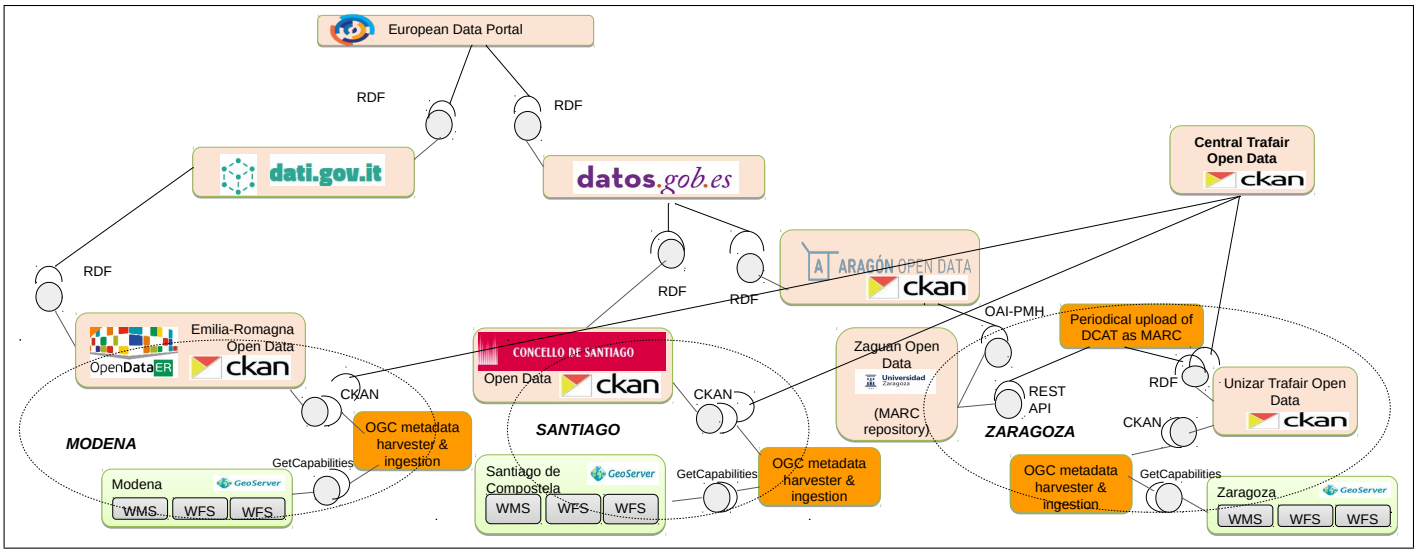

Figure 4. Deployment of Open Data servers in the cities of Modena, Santiago and Zaragoza

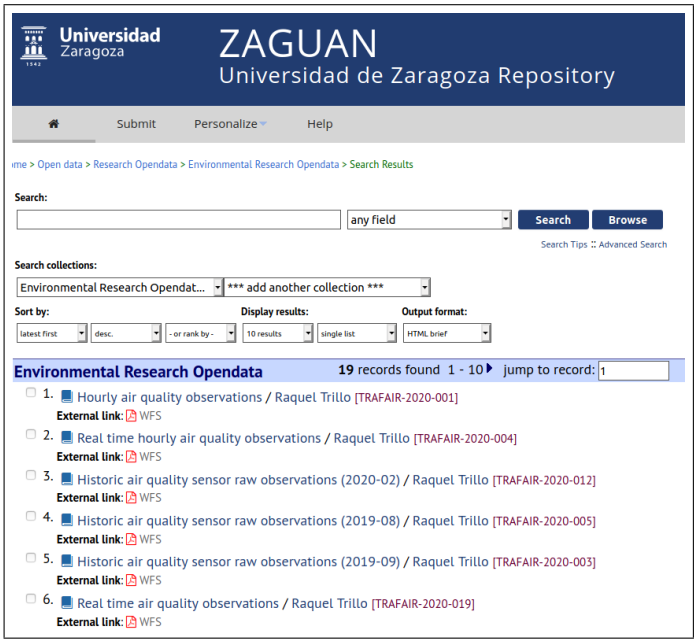

Figure 5. TRAFAIR datasets at Zaguan repository (University of Zaragoza)

contents of CSW catalogs. Nevertheless, none of these two approaches derive metadata automatically from services.

The workflow for the publication of open data proposed in this work contributes to the state of the art as it provides an integrated approach to solve jointly three challenges: the automatic generation of metadata from the *GetCapabilities* responses of OGC services; the generation of DCAT-based metadata; and the synchronized publication of data and metadata.

VI. Conclusions

We have proposed the workflow for the publication of Open Spatial Data that can be customized to other projects dealing with spatial data that must be publicly accessible. Besides, we have demonstrated how GeoDCAT-AP metadata can be applied in a real use case to describe more specifically spatial data than other more general metadata vocabularies based on DCAT.

As future work, we plan to integrate the software that we have developed for the automatic generation and publication of metadata as a new plugin of CKAN, or as an extension of existing *ckanext-spatial* plugin. Another work in progress is the evaluation of the quality of metadata according to several approaches like the Metadata Quality Assurance methodology [16] or the ISO 19157-based method for metadata quality analysis [17].

Acknowledgment

This research has been supported by the TRAFAIR project 2017-EU-IA-0167, co-financed by the Connecting Europe Facility of the European Union. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.

References

[1] L. Po, F. Rollo, J. R. R. Viqueira, R. T. Lado, A. Bigi, J. C. López, M. Paolucci, and P. Nesi, "TRAFAIR: understanding traffic flow to improve air quality," in 2019 IEEE International Smart Cities Conference, ISC2 2019, Casablanca, Morocco, October 14-17, 2019. IEEE, 2019, pp. 36–43. [Online]. Available: https://doi.org/10.1109/ISC246665.2019.9071661

[2] European Commission, "GeoDCAT Application profile for data portals in Europe, GeoDCAT-AP v1.0.1," https://joinup.ec.europa.eu/release/geodcat-ap/101, 2016.

[3] ——, "COMMISSION REGULATION (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata," European Union, Tech. Rep., 2008.

[4] International Organization for Standardization (ISO), "ISO 19115-1:2014. Geographic information - Metadata - Part 1: Fundamentals," Geneva, CH, Tech. Rep., 2014.

[5] W3C, "Data Catalog Vocabulary (DCAT). W3C Working Draft,12 March 2013," http://www.w3.org/TR/2013/WD-vocab-dcat-20130312/, 2013.

[6] European Commission, "DCAT Application Profile for data portals in Europe, DCAT-AP v2.0.0," https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe/release/200, 2019.

[7] INSPIRE MIG, "Technical Guidelines for implementing dataset and service metadata based on ISO/TS 19139:2007," INSPIRE Maintenance and Implementation Group (MIG), INSPIRE Maintenance and Implementation Group (MIG). Version 2.0.1, 2017, http://inspire.ec.europa.eu/id/document/tg/metadata-iso19139.

[8] D. Öttl, P. Sturm, G. Pretterhofer, M. Bacher, J. Rodler, and R. Almbauer, "Lagrangian dispersion modeling of vehicular emissions from a highway in complex terrain," Journal of the Air and Waste Management Association, vol. 53, 2003, pp. 1233–1240.

[9] J. Nogueras-Iso, J. Barrera, A. Rodríguez-Pascual, R. Recio, C. Laborda, and F. Zarazaga-Soria, SDI Convergence: Research, Emerging Trends, and Critical Assessment. The Netherlands Geodetic Commission (NGC), 2009, ch. Development and deployment of a services catalog in compliance with the INSPIRE metadata implementing rules.

[10] Open Source Geospatial Foundation, "Catalog Services for the Web (CSW) - ISO Metadata Profile, GeoServer Community Module," https://docs.geoserver.org/stable/en/user/community/csw-iso/index.html, 2020.

[11] ——, "GeoNetwork User Manual v2.10.4-0, Harvesting OGC Services," https://geonetwork-opensource.org/manuals/2.10.4/eng/users/managing_metadata/harvesting/ogcwxs/index.html#ogcwxs-harvester, 2020.

[12] A. J. Florczyk, J. Nogueras-Iso, F. J. Zarazaga-Soria, and R. Béjar, "Identifying orthoimages in web map services," Computers & geosciences, vol. 47, 2012, pp. 130–142.

[13] J. Gil-Altaba, L. Díaz-Sánchez, C. Granell-Canut, and J. Huerta-Guijarro, "Open source based deployment of environmental data into geospatial information infrastructures," International Journal of Applied Geospatial Research, vol. 3, no. 2, 2012, p. 6–23.

[14] A. Perego, A. Friis-Christensen, and M. Lutz, "GeoDCAT-AP: Use cases and open issues," in Smart Descriptions & Smarter Vocabularies (SDSVoc) workshop. Amsterdam, 30 Nov - 1 Dec 2016, https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_25.

[15] Open Knowledge, "ckanext-spatial - Geo related plugins for CKAN," https://docs.ckan.org/projects/ckanext-spatial/en/latest/, 2015.

[16] Publications Office of the European Union, "Metadata Quality Assessment Methodology. How EDP measures the quality of harvested metadata," https://www.europeandataportal.eu/mqa/methodology, 2020.

[17] M. Ureña-Cámara, J. Nogueras-Iso, J. Lacasta, and F. Ariza-López, "A method for checking the quality of geographic metadata based on iso 19157," International Journal of Geographical Information Science, vol. 33, no. 1, 2019, pp. 1–27.