The Twelfth International Conference on Advanced Geographic Information Systems, Applications and Services
GEOProcessing 2020
21-25 November 2020, Valencia (Spain)

TRAFAIR
www.trafair.eu

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, Raquel Trillo-Lado

(*) Aragon Institute of Engineering Research (I3A), Universidad de Zaragoza (Spain)
jnog@unizar.es

Javier Nogueras-Iso biography

- Javier Nogueras-Iso holds MS and PhD degrees in Computer Science from the University of Zaragoza.
- In 1998, he started his research at the Advanced Information Systems Laboratory of the University of Zaragoza. Currently, he is an Associate Professor of Computer Science at that University.
- Between 2011 and 2017 he was Director of Catedra Logisman on 'Technological Document Management', and between 2015 and 2019 he was Associate Director of the Aragon Institute of Engineering Research (I3A).
- His research interests are focused on Information Retrieval and Semantic Web technologies applied to different domains, although with a special emphasis on Geographic Information Infrastructures.

Outline

- 1. Introduction
- 2. A metadata profile for Open Spatial Data
- 3. Workflow for the publication of Open Spatial Data
- 4. Deployment of Open Data in Zaragoza, Santiago and Modena
- 5. Conclusion and future work

1. Introduction

TRAFAIR
www.trafair.eu

- TRAFAIR: Understanding traffic flows to improve air quality
  - European project co-financed by the Connecting Europe Facility of the European Union
  - Project Nr. 2017-EU-IA-0167 (Nov 2018 – Oct 2020)
- 10 partners based in 6 cities from 2 European countries
  - 4 universities, 2 research centres
  - 3 public administrations
  - 1 regional in-house providing company

Project goals

TRAFAIR
www.trafair.eu

- Design and develop the necessary infrastructure to **estimate** the pollution level on urban scale
  - Test the infrastructure in 6 European cities of different size
- Sub-goals:
  - Provide real-time **monitoring** of air pollution in the city on an urban scale
  - Develop an air quality **forecasting** service based on the weather forecasts and the urban traffic flow
  - Publish monitoring and forecasting air quality and traffic data as **open data**
  - Develop applications for end-users and public administrations

Focus of this work

TRAFAIR
www.trafair.eu

- **Publication of Open Data**
  - Monitoring data
    - Low cost sensor data
    - Coarse urban air pollution maps in semi-real time
  - Forecast maps
    - Dispersion maps of pollutants
- Given the spatial character of data, it was clear that Spatial Data Infrastructures (SDI) could provide the necessary technology for the management/publication of data
- However, publication through typical SDI services compliant with OGC is not enough for Open Data
  - Data must be registered in Open Data portals
  - TRAFAIR data should be visible through the European Data Portal (EDP)

Objectives of this work

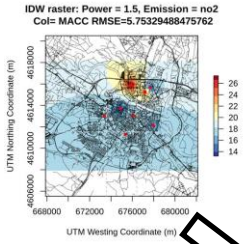
- Propose a workflow for the publication of Open Spatial Data
- Desirable features:
  - Automate as much as possible the generation of metadata without human interaction
  - Adopt a metadata model compatible with both the Open Data domain and the spatial domain
    - Use of GeoDCAT-AP
  - Provide a solution based on Open Source software packages
    - GeoServer: management of spatial data
    - CKAN: platform for deploying Open Data portals

Software components for the management and publication of data

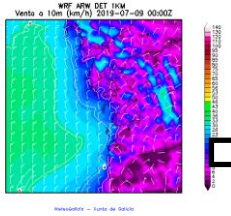
TRAFAIR
www.trafair.eu

Low cost sensor Data

Monitoring maps



Forecast maps



European Data Portal

Spatial Data Infrastructures (OGC – INSPIRE)

WFS

WMS

WCS

CSW

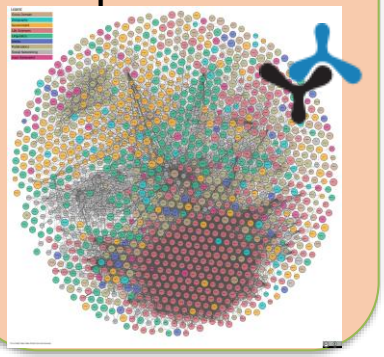
GeoServer

pycsw

Trafair Open Data

ckan

W3C Linked Open Data



Vector Data

Raster Data

ISO 19115 metadata

Semantic Environmental Geospatial Data Warehouse

GeoDCAT-AP metadata

RDF data

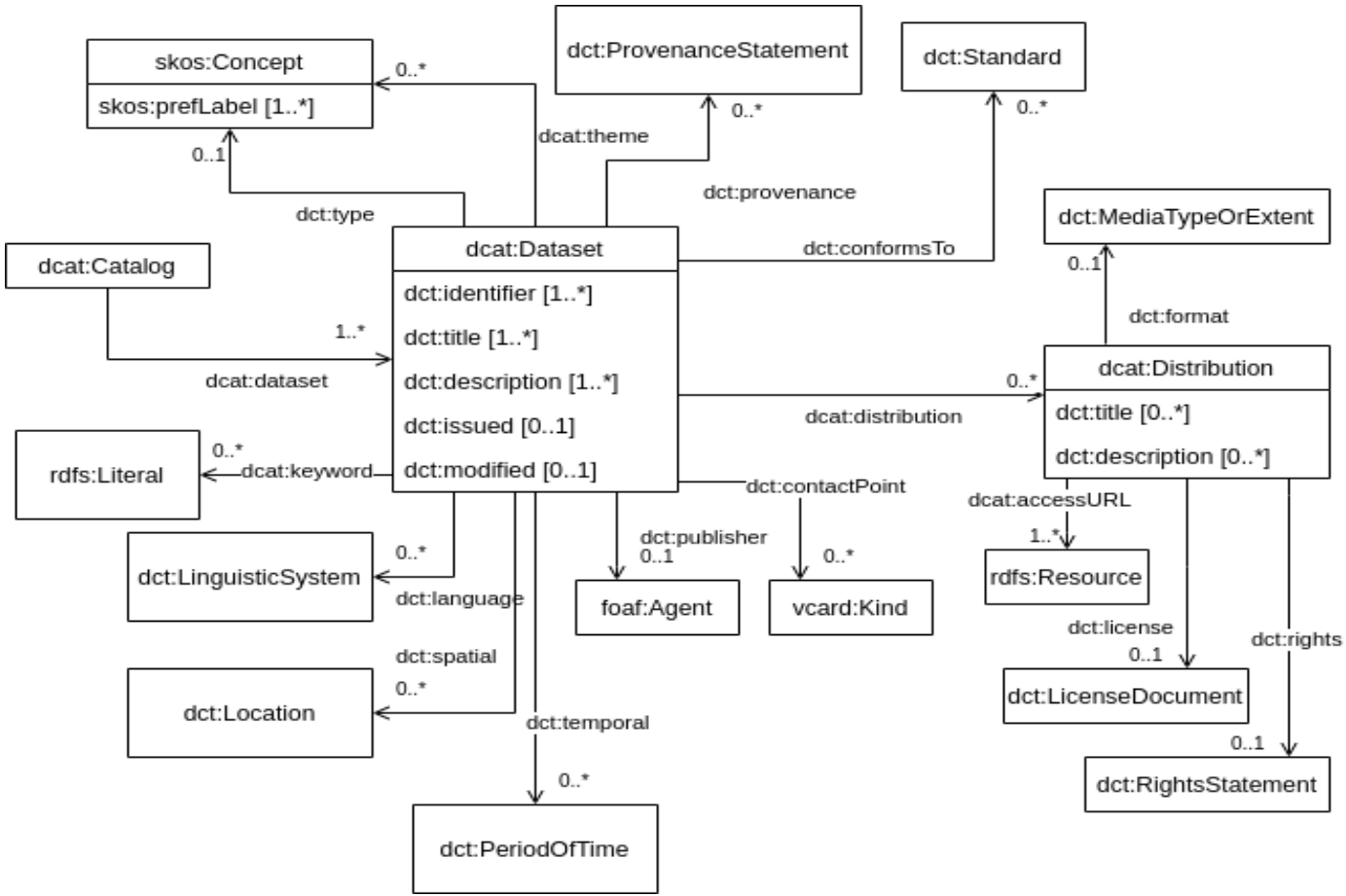
2. A metadata profile for Open Spatial Data

- Metadata model alternatives
  - ISO 19115: traditional metadata for geographic information datasets and services
    - large and complex model, whose automation is complicate
  - DCAT: Data Catalogue vocabulary proposed by W3C for the description of Open Data
  - DCAT-AP: European Application Profile of DCAT for public sector datasets
  - GeoDCAT-AP: extension of DCAT-AP for geographic information
    - Easily accepted in Open Data software
    - Designed to assure compliance with European INSPIRE directive for establishing a spatial information infrastructure in Europe (and ISO 19115)

GeoDCAT-AP

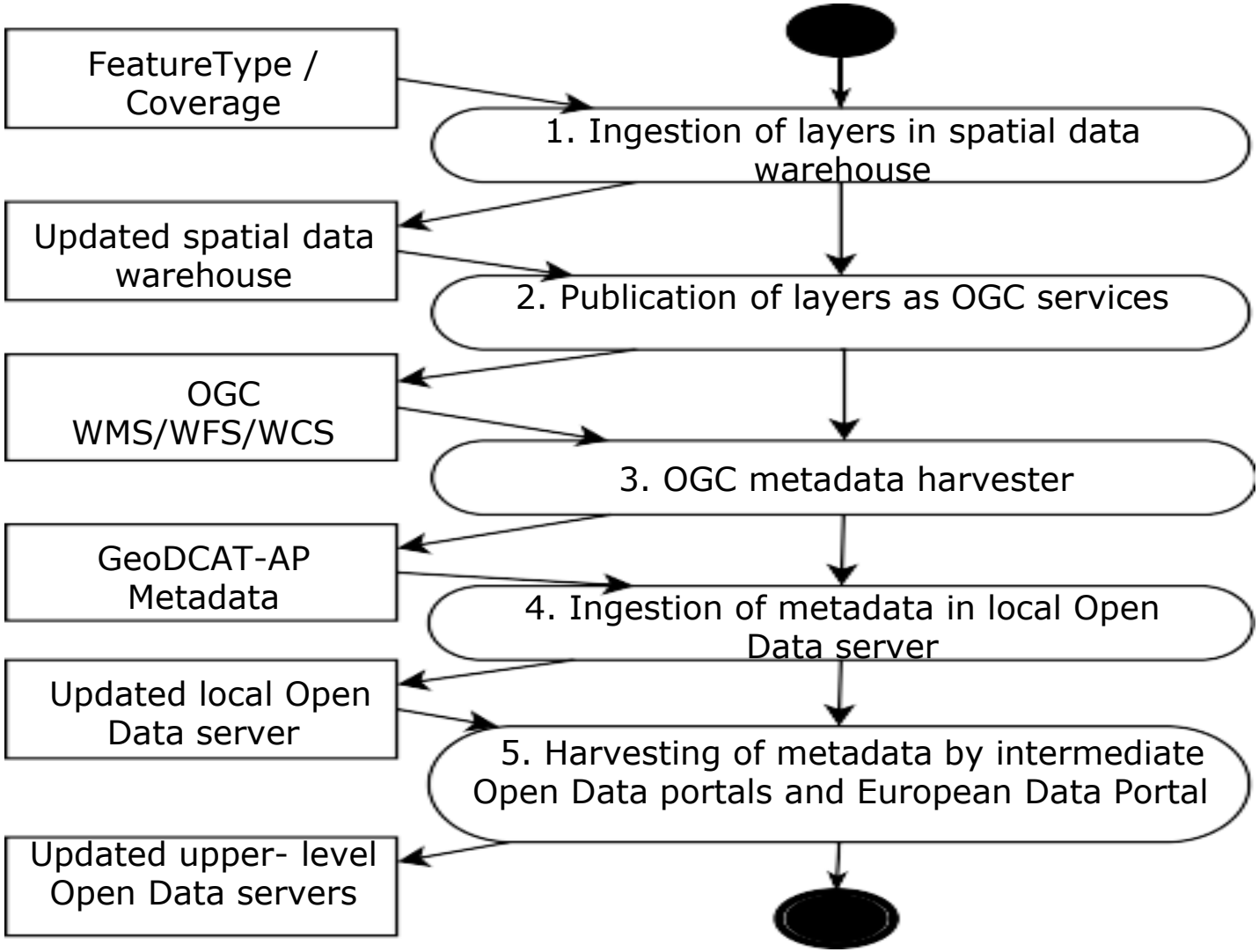
TRAFAIR
www.trafair.eu

- Light adoption of GeoDCAT-AP
  - Use mostly core properties (direct binding with ISO 19115-INSPIRE)
  - All properties compatible with DCAT-AP
  - All elements editable with CKAN software



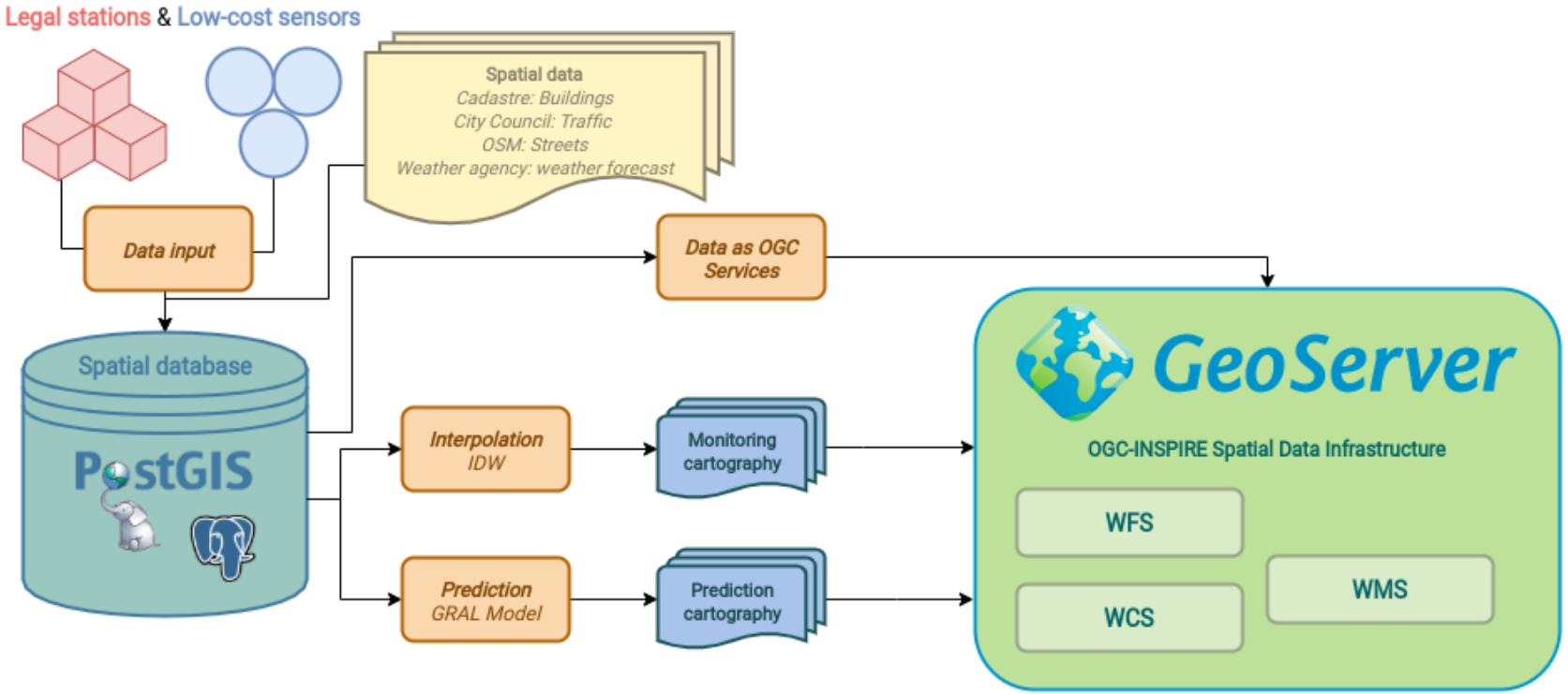
3. Workflow for the publication of Open Spatial Data

TRAFAIR
www.trafair.eu



3.1. Ingestion of layers in spatial data warehouse

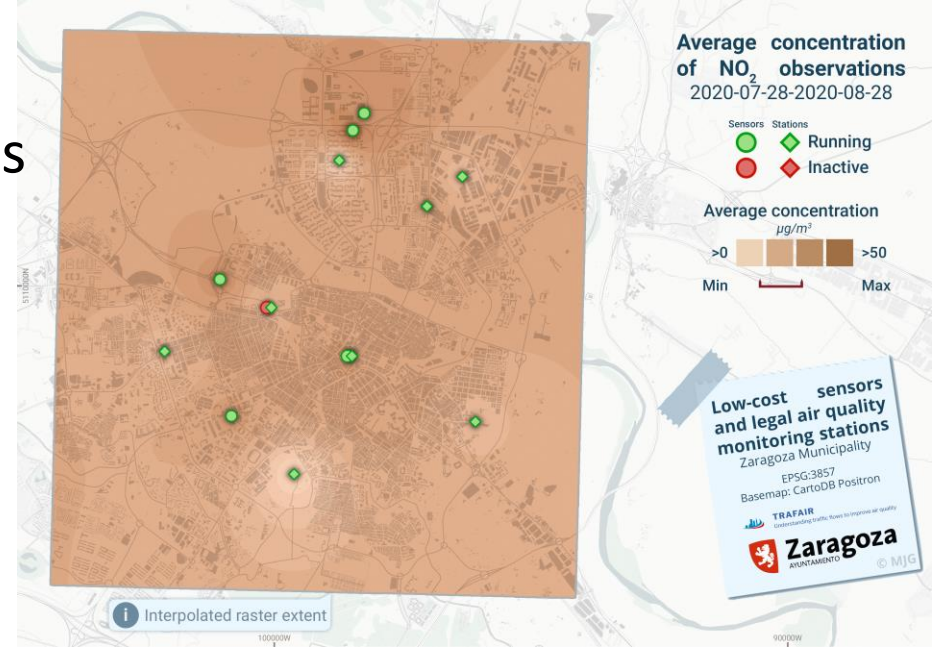
- We have developed specific software in Java and R languages to ingest layers in GeoServer
  - Feature types (monitoring observations)
  - Coverages (interpolation of observations, predictions based on a Lagrangian model for dispersion of pollutants)



3.2. Publication of layers as OGC services

TRAFAIR
www.trafair.eu

- During the ingestion of layers, we can setup GeoServer to provide access to some layers through OGC services
- In the TRAFAIR case
  - Feature Types are downloaded through a Web Feature Service (WFS)
  - Coverages are downloaded through a Web Coverage Service (WCS)
  - In addition, some layers (coverages, traffic models , …) are also available as maps through a Web Mapping Service (WMS)



3.3. OGC metadata harvester

TRAFAIR
www.trafair.eu

- Metadata is harvested from OGC services through its *GetCapabilities* operation
- Python program using OWSLib package to retrieve *GetCapabilities*
- For each WCS/WFS layer we generate one Dataset and one Distribution
  - A second optional distribution if the layer is rendered through a WMS

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 ref	extra:conforms to	Dataset/dct:conformsTo

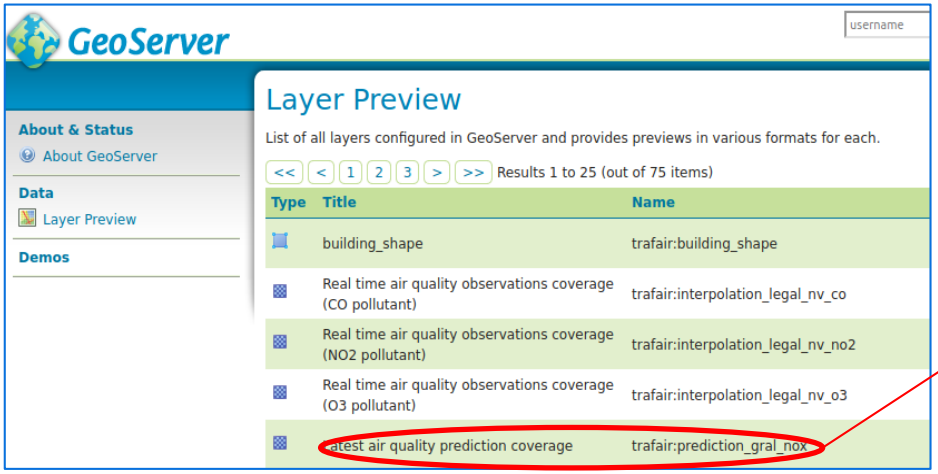
3.4. Ingestion of metadata in local Open Data server

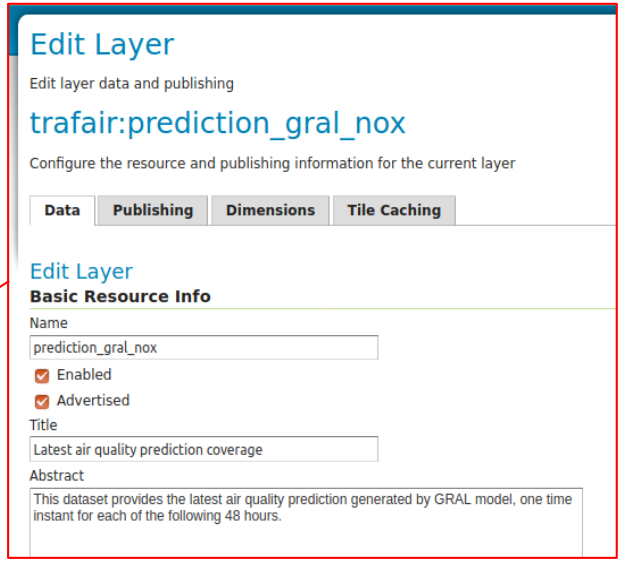
- Our Python program transforms the information retrieved in the previous step into a dictionary with the required items to construct a dataset and its associated resources, which can be ingested through the REST API of CKAN
- We select the appropriate tags to assure that GeoDCAT-AP RDF can be generated with *ckanext-dcat* plugin

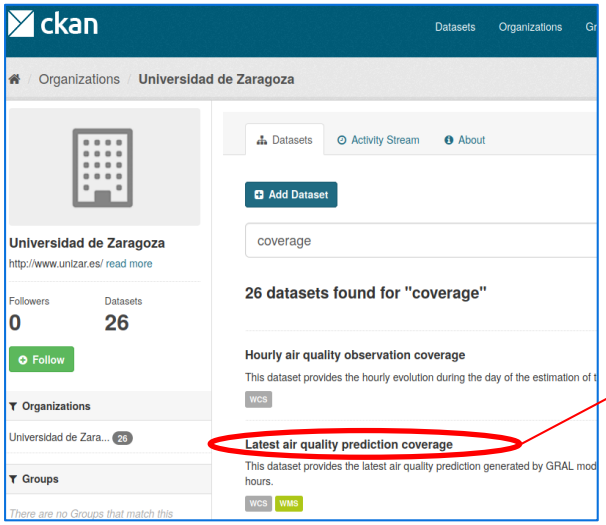
GeoServer	**OWSLib**	**CKAN**	**GeoDCAT-AP**
featureType/name, coverage/name	*layerName*	extra:identifier	Dataset/dct:identifier
featureType/title, coverage/title	contents[*layerName*].title	title	Dataset/dct:title
featureType/description, coverage/description *(software in step 1 introduces predefined 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 ref	extra:conformsTo	Dataset/dct:conformsTo

Example of steps 1-4

TRAFAIR
www.trafair.eu







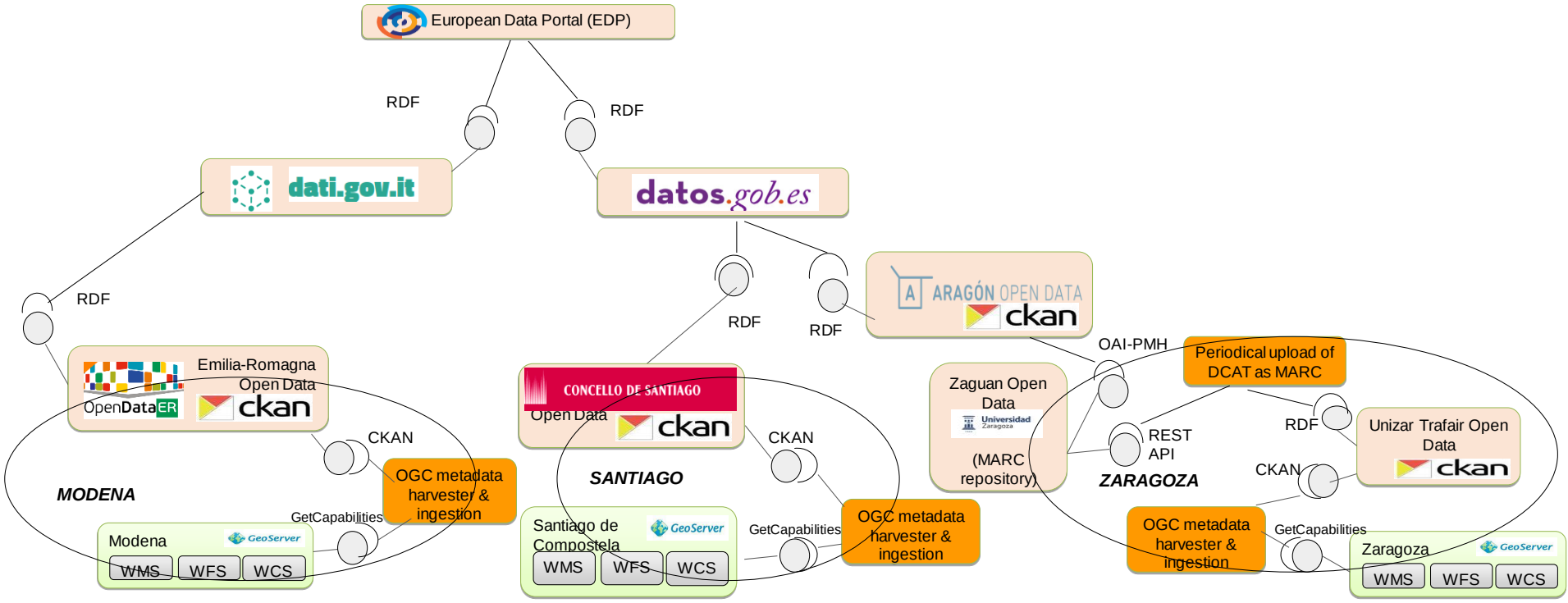

```
<dcat:Dataset rdf:about="http://atila.unizar.es:3394/dataset/dbb3799e-dc88-4e8f-995b-7d8
  <dct:title>Latest air quality prediction coverage</dct:title>
  <dct:description>This dataset provides the latest air quality prediction generated by
  <dct:type>http://inspire.ec.europa.eu/metadata-codelist/ResourceType/series</dct:type>
  <dcat:theme rdf:resource="http://inspire.ec.europa.eu/metadata-codelist/TopicCategory/
  <dct:temporal>
    <dct:PeriodOfTime rdf:nodeID="N4229e255693543359f8bdad8d23747c4">
      <schema:startDate rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2019-07
      <schema:endDate rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2020-10-3
    </dct:PeriodOfTime>
  </dct:temporal>
  <dcat:theme rdf:resource="http://inspire.ec.europa.eu/theme/ac"/>
  <dcat:keyword>TRAFAIR</dcat:keyword>
  <dcat:keyword>prediction_gral_nox</dcat:keyword>
```

3.5. Harvesting of metadata by intermediate Open Data portals and European Data Portal

- This step is beyond the scope of the TRAFAIR project.
- Our local CKAN servers include the installation of *ckanext-dcat* plugin
  - This allows the publication of datasets metadata as RDF in compliance with DCAT-AP vocabularies
- We assume that upper-level portals are based on CKAN technology (or have a similar mechanism for the harvesting of subscribed lower level catalogs)
  - The *ckanext-harvest* plugin of CKAN allow to harvest the contents of different types of catalog sources
  - Most regional/national portals based on DCAT interact with lower-level CKAN servers through its catalog end-point to retrieve contents as RDF

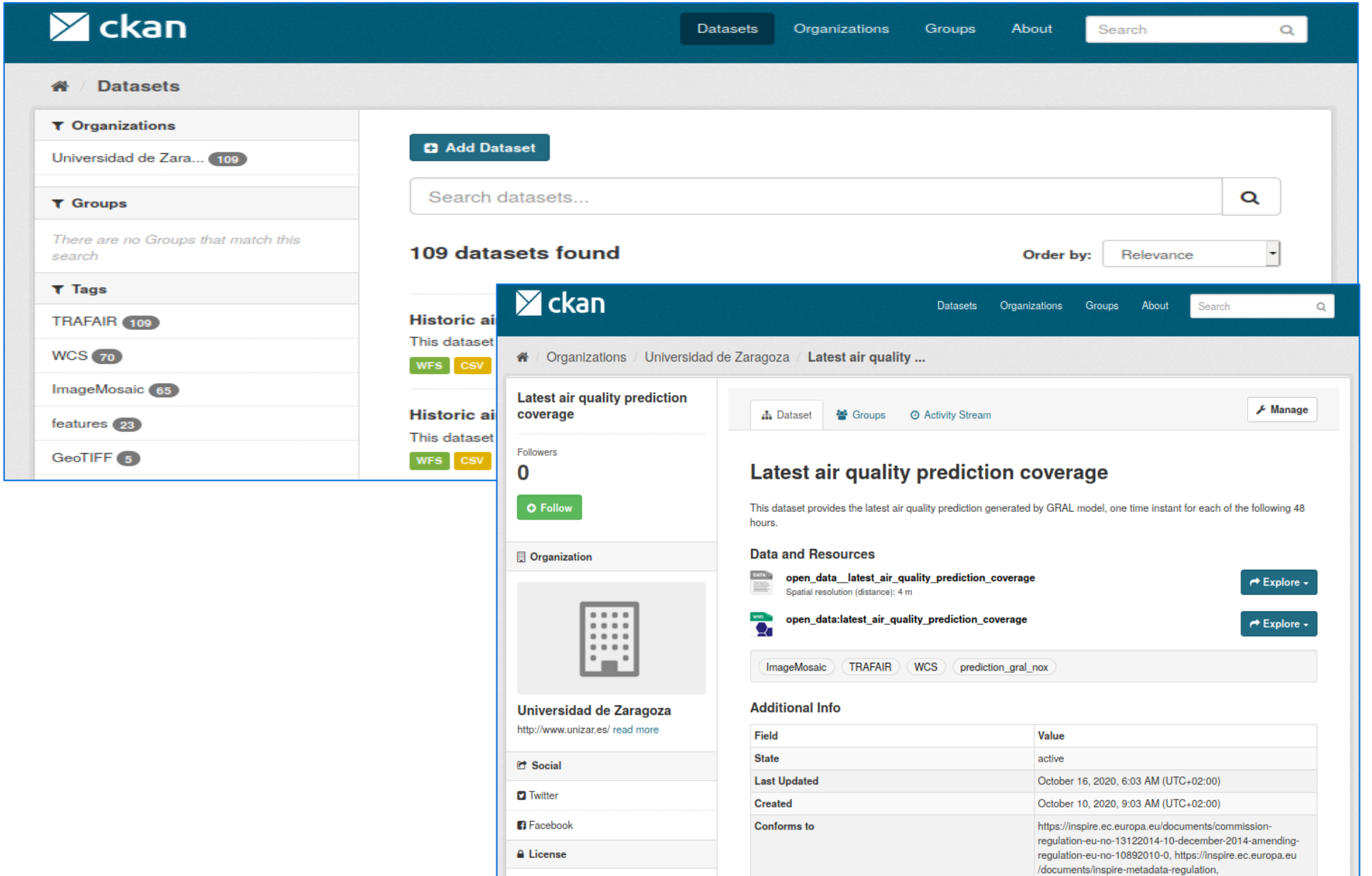
4. Deployment of Open Data in Zaragoza, Santiago and Modena

TRAFAIR
www.trafair.eu



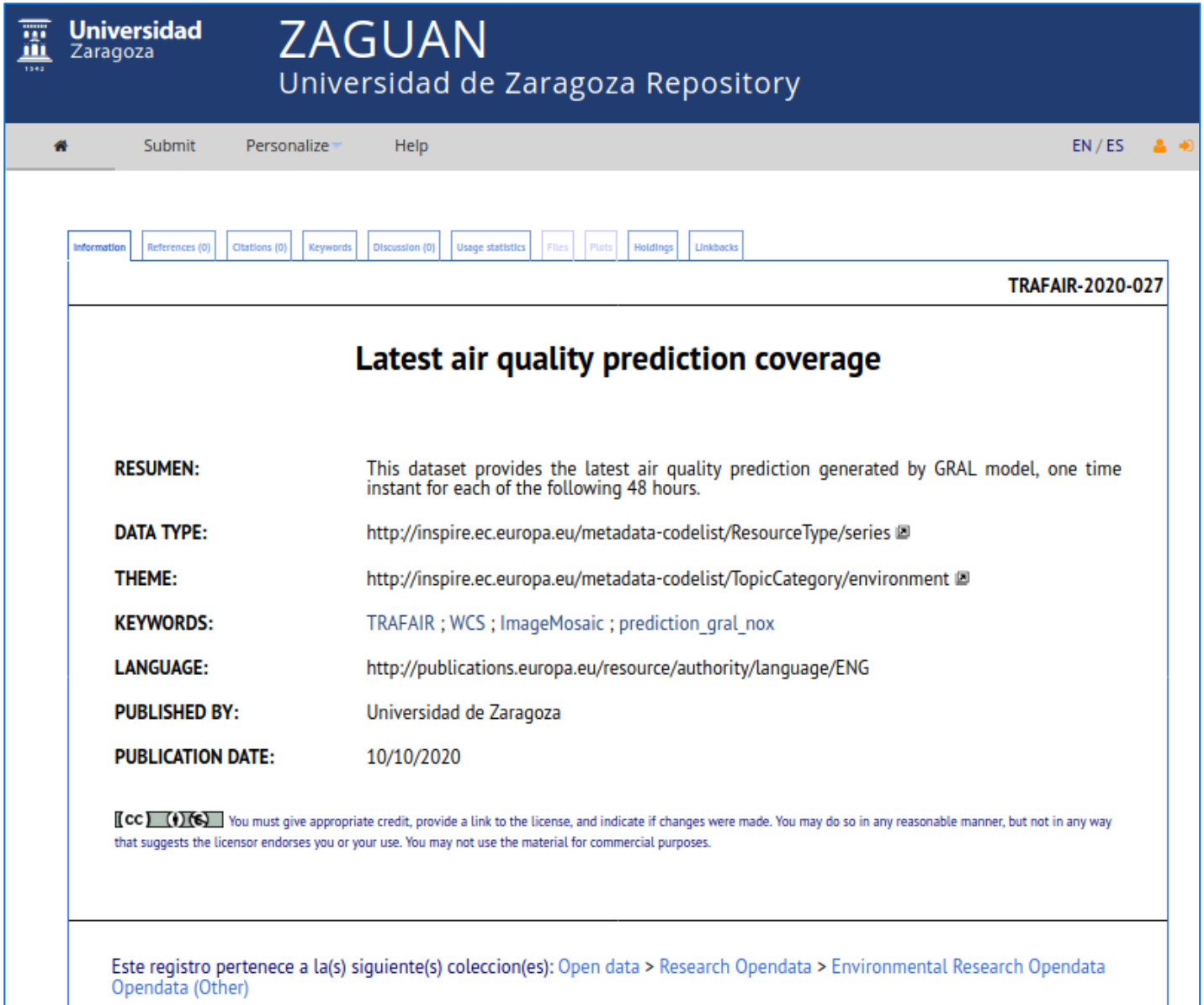
An example of the harvesting process (I)

TRAFAIR
www.trafair.eu



An example of the harvesting process (II)

TRAFAIR
www.trafair.eu



Universidad Zaragoza

ZAGUAN
Universidad de Zaragoza Repository

Submit | Personalize | Help | EN / ES

Information | References (0) | Citations (0) | Keywords | Discussion (0) | Usage statistics | Files | Plots | Holdings | Linkbacks

TRAFAIR-2020-027

Latest air quality prediction coverage

RESUMEN: This dataset provides the latest air quality prediction generated by GRAL model, one time instant for each of the following 48 hours.

DATA TYPE: http://inspire.ec.europa.eu/metadata-codelist/ResourceType/series

THEME: http://inspire.ec.europa.eu/metadata-codelist/TopicCategory/environment

KEYWORDS: TRAFAIR ; WCS ; ImageMosaic ; prediction_gral_nox

LANGUAGE: http://publications.europa.eu/resource/authority/language/ENG

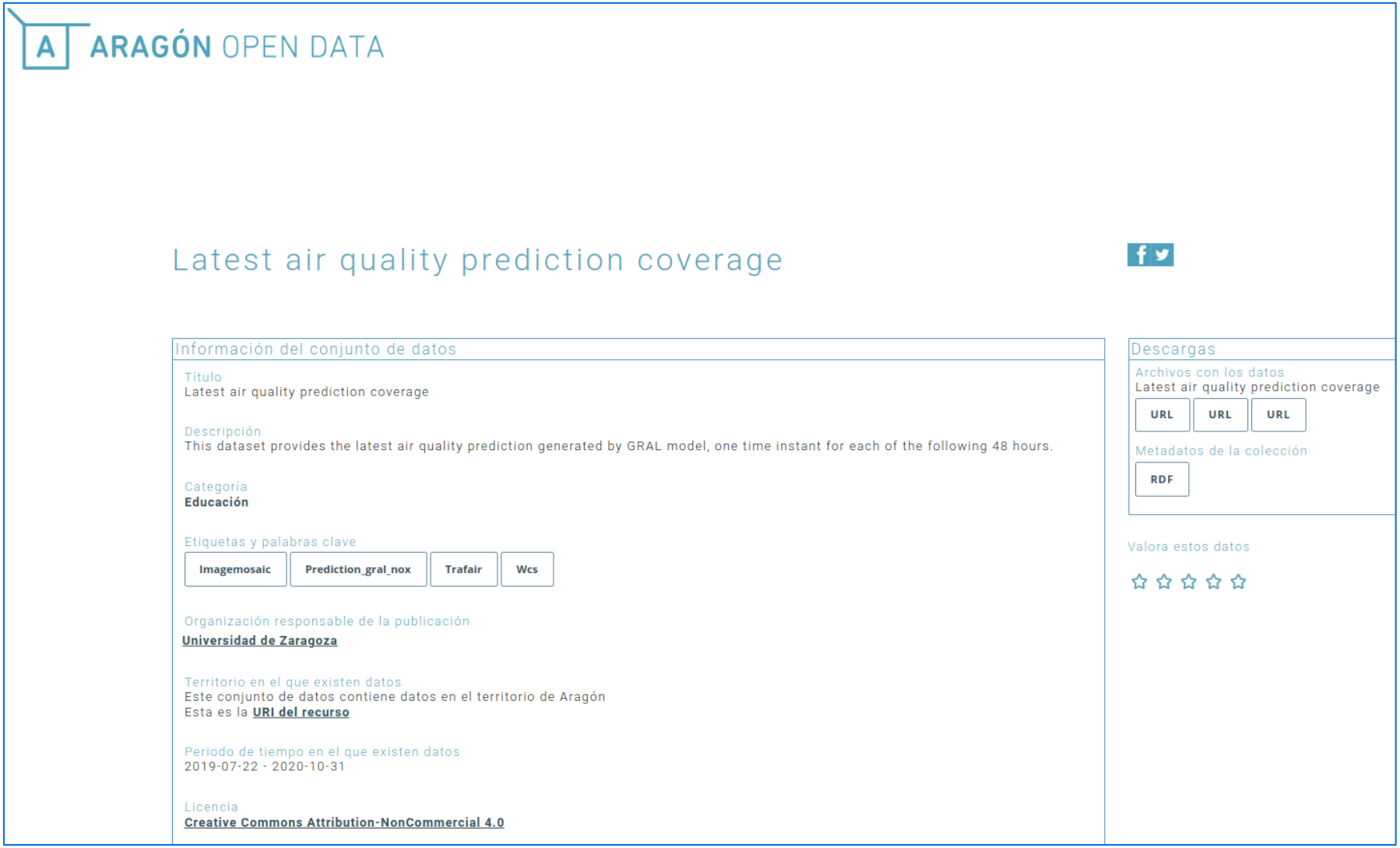
PUBLISHED BY: Universidad de Zaragoza

PUBLICATION DATE: 10/10/2020

Este registro pertenece a la(s) siguiente(s) coleccion(es): Open data > Research Opendata > Environmental Research Opendata Opendata (Other)

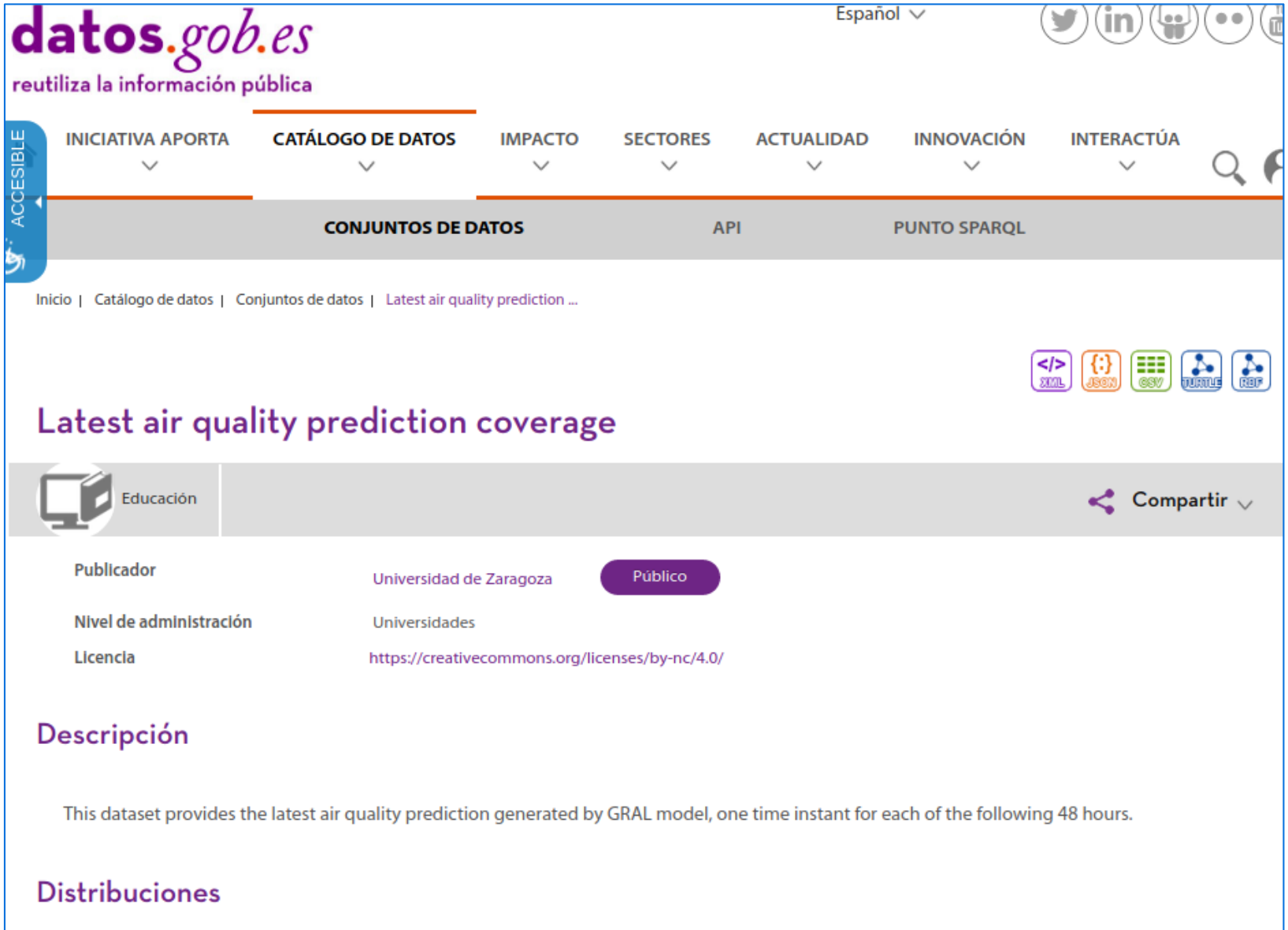
An example of the harvesting process (III)

TRAFAIR
www.trafair.eu



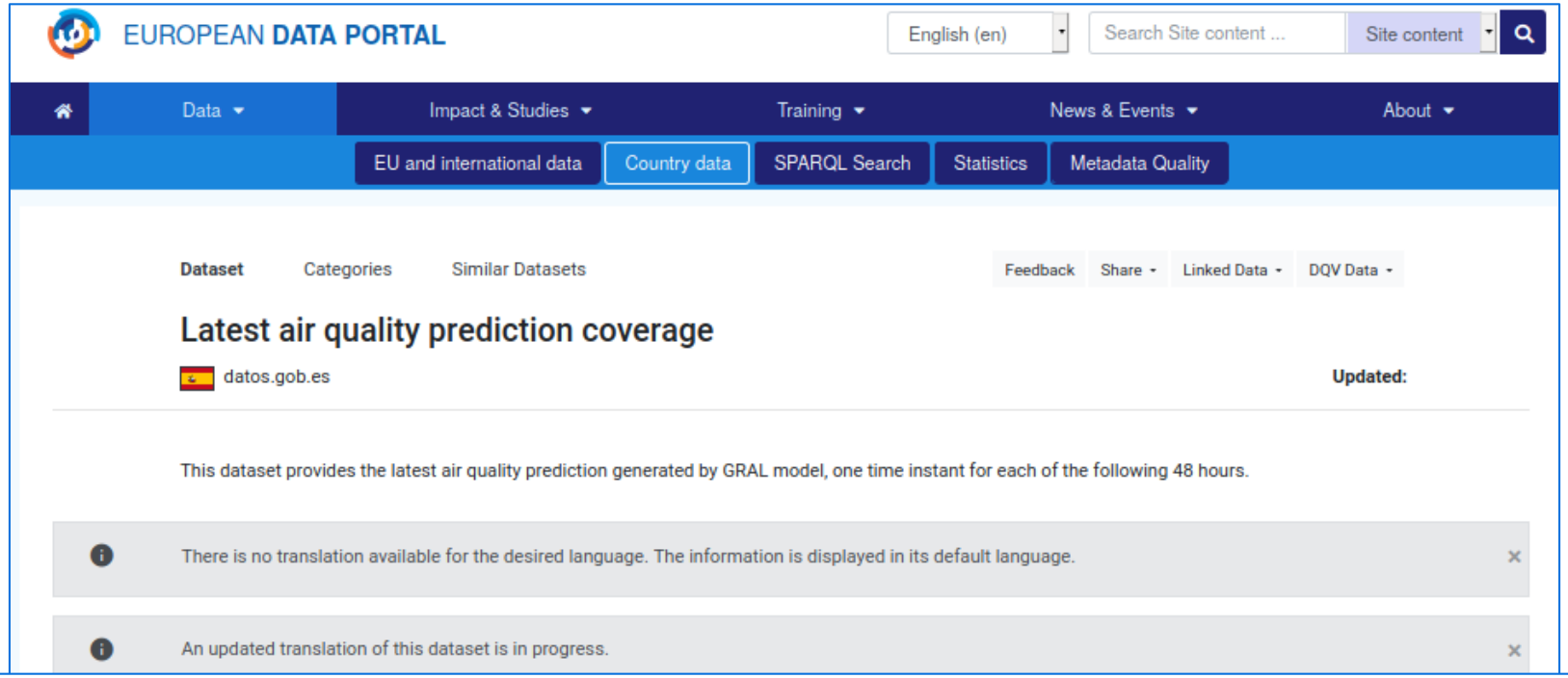
An example of the harvesting process (IV)

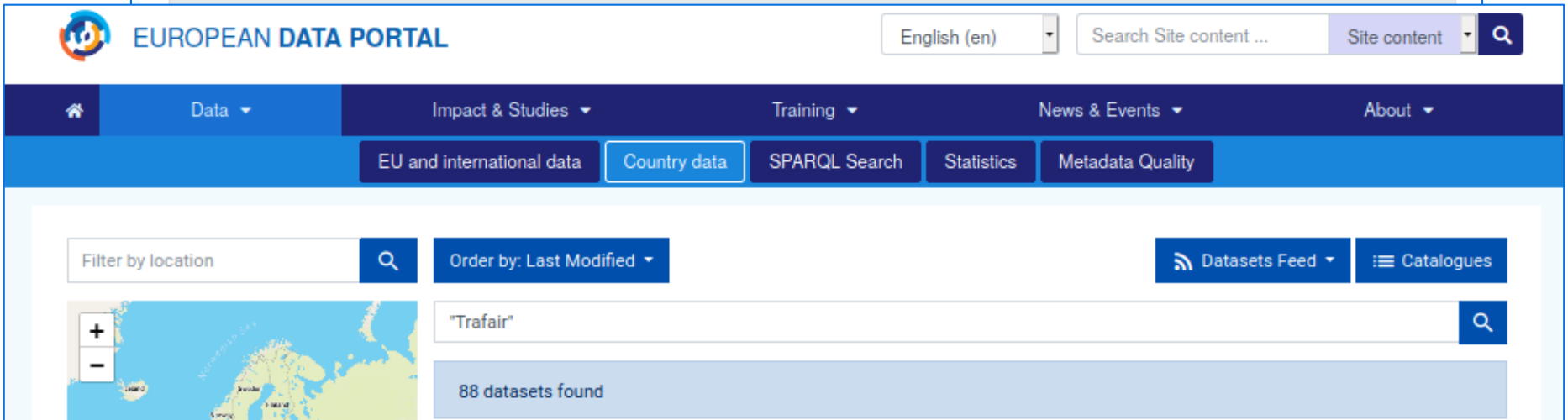
TRAFAIR
www.trafair.eu



An example of the harvesting process (V)

TRAFAIR
www.trafair.eu





5. Conclusion and future work (I)

TRAFAIR
www.trafair.eu

- We have proposed the workflow for the publication of Open Spatial Data that solves jointly 3 challenges:
  - the automatic generation of metadata from the *GetCapabilities* responses of OGC services
  - the generation of DCAT-based metadata
  - the synchronized publication of data and metadata
- 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
- This workflow can be customized to other projects dealing with spatial data that must be publicly accessible

5. Conclusion and future work (II)

- Lessons learned
  - Not all GeoDCAT-AP guidelines to fill metadata elements could be followed by local Open Data portals because they must comply with constrained profiles of DCAT-AP imposed by national governments
    - Instead of INSPIRE data themes (i.e. “atmosphere” and ”environment facilities”), local portals must choose a theme from a national controlled vocabulary
    - Some elements are not allowed in national profiles (e.g. *dct:provenance*)
- Future work
  - Integrate software as a new plugin of CKAN, or as an extension of *ckanext-spatial*
  - Evaluate quality of generated metadata through different methods (e.g. EDP Metadata Quality Assessment methodology)

TRAFAIR

www.trafair.eu

TRAFAIR

Understanding traffic flows to improve air quality

http://trafair.eu/

@trafairproject

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

UNIVERSITÀ DEGLI STUDI FIRENZE

DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DISIT DISTRIBUTED SYSTEMS AND INTERNET TECHNOLOGIES LAB

Universidad Zaragoza

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Consiglio Nazionale delle Ricerche
Istituto per la BioEconomia

Comune di Modena

CONCELLO DE SANTIAGO

REGIONE TOSCANA

lepidaspa

CESGA

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