

MOBIDATALAB

Labs for prototyping future mobility data sharing solutions in the cloud

D4.3 Reference Data Catalogue (V1)

11/08/2022

Author(s): Thierry CHEVALLIER (AKKA), Didier DE-RYCK (KISIO), Hiba MECHYAKHA (KISIO), Huy Minh NGUYEN (HERE), Renée OBREGON-GONZALES (AKKA), Sorel SIGHOKO (AKKA)



MobiDataLab is funded by the EU under the H2020 Research and Innovation Programme (grant agreement No 101006879).

Summary sheet

Deliverable Number	D4.3
Deliverable Name	Reference Data Catalogue (V1)
Full Project Title	MobiDataLab, Labs for prototyping future Mobility Data sharing cloud solutions
Responsible Author(s)	Thierry CHEVALLIER (AKKA), Didier DE-RYCK (KISIO), Huy Minh NGUYEN (HERE)
Contributing Partner(s)	AKKA, HERE, KISIO
Peer Review	POLIS, URV
Contractual Delivery Date	31-07-2022
Actual Delivery Date	29-07-2022
Status	Final
Dissemination level	Public
Version	1.0
No. of Pages	53
WP/Task related to the deliverable	WP4/T4.2
WP/Task responsible	AKKA/AKKA
Document ID	MobiDataLab-D4.3-ReferenceDataCatalogueV1_v1.0.docx
Abstract	This deliverable is a report to provide an overview of the Task 4.2 demonstrator

Legal Disclaimer

MOBIDATALAB (Grant Agreement No 101006879) is a Research and Innovation Actions project funded by the EU Framework Programme for Research and Innovation Horizon 2020. This document contains information on MOBIDATALAB core activities, findings, and outcomes. The content of this publication is the sole responsibility of the MOBIDATALAB consortium and cannot be considered to reflect the views of the European Commission.

Project partners

Organisation	Country	Abbreviation
AKKA I&S	France	AKKA
CONSORZIO INTERUNIVERSITARIO PER L'OTTIMIZZAZIONE E LA RICERCA OPERATIVA	Italy	ICOOR
AETHON SYMVOULI MICHANIKI MONOPROSOPI IKE	Greece	AETHON
CONSIGLIO NAZIONALE DELLE RICERCHE	Italy	CNR
KISIO DIGITAL	France	KISIO
HERE GLOBAL B.V.	Netherlands	HERE
KATHOLIEKE UNIVERSITEIT LEUVEN	Belgium	KUL
UNIVERSITAT ROVIRA I VIRGILI	Spain	URV
POLIS - PROMOTION OF OPERATIONAL LINKS WITH INTEGRATED SERVICES	Belgium	POLIS
F6S NETWORK IRELAND LIMITED	Ireland	F6S

Document history

Version	Date	Organisation	Main area of changes	Comments
0.1	02/06/2022	AKKA	outline	
0.2	15/06/2022	HERE	All	GeoNetwork contribution
0.3	10/07/2022	AKKA	2.1, 3.1	CKAN contribution
0.4	15/07/2022	KISIO	2.2	OpenDataSoft contribution
0.5	22/07/2022	AKKA	1.x, 2.x, 3.x	Introduction
0.6	25/07/2022	URV, POLIS	All	Review
0.7	27/07/2022	AKKA	All	Rework
0.8	29/07/2022	AKKA	All	Quality Check
1.0	29/07/2022	AKKA	All	Submission

Executive Summary

The deliverable D4.3 is a report providing an overview on the version 1 of the MobiDataLab data catalogue. This integrated catalogue built on CKAN, GeoNetwork and OpenDataSoft solutions, references all open transport datasets and corresponding metadata in the territorial context and specific domains of the “Reference Group” of MobiDataLab stakeholders.

Table of contents

1. INTRODUCTION.....	9
1.1. PURPOSE OF THE DELIVERABLE.....	9
1.2. THE REFERENCE GROUP OF MOBILITY STAKEHOLDERS.....	10
1.2.1. Reference group of local organisations.....	10
1.2.2. Reference group of international organisations.....	10
1.2.3. Data sources from the reference group.....	10
1.3. DATA PROVIDERS AND DATA CONSUMERS.....	12
1.4. CATALOGUE SOFTWARE SYSTEMS.....	12
2. USING SOFTWARE CATALOGUE SYSTEMS FOR MOBILITY DATA DISCOVERY.....	14
2.1. USING CKAN FOR MOBILITY DATA DISCOVERY.....	15
2.1.1. Discovering data in CKAN.....	15
2.1.2. Managing organisations in CKAN.....	15
2.1.3. Managing groups in CKAN.....	16
2.1.4. Managing datasets in CKAN.....	17
2.1.5. Data and metadata storage.....	18
2.1.6. Visualising data in CKAN.....	18
2.1.7. Harvesting.....	19
2.1.8. Multilingual management.....	21
2.1.9. Interoperability.....	21
2.2. USING OPENDATASOFT FOR MOBILITY DATA DISCOVERY.....	22
2.2.1. Customizing the catalogue: Back-office.....	23
2.2.2. Data catalogue.....	24
2.2.3. Data analytics.....	27
2.2.4. OpenDataSoft APIs.....	28
2.2.5. Integrating reference data into KISIO OpenDataSoft instance.....	28
2.3. USING GEONETWORK FOR MOBILITY (GEO-REFERENCED) DATA DISCOVERY.....	29
2.3.1. Web interface.....	29
2.3.2. Managing metadata in GeoNetwork.....	30
2.3.3. Data discovery.....	37
2.3.4. Visualising data in GeoNetwork.....	38
2.3.5. Interoperability.....	39
2.3.6. Harvesting.....	39
2.3.7. Multilingual management.....	40
2.3.8. Permission management.....	41
2.3.9. Admin console.....	42
3. INTEGRATING CATALOGUE SOFTWARE SYSTEMS INTO THE MOBIDATALAB TRANSPORT CLOUD.....	44
3.1. INTEGRATING CKAN IN THE TRANSPORT CLOUD.....	45
3.1.1. Installing CKAN packages for Ubuntu 20.04.....	45

3.1.2. Installing PostgreSQL	46
3.1.3. Installing Solr	46
3.1.4. Initialise the CKAN database	47
3.2. INTEGRATING OPENDATASOFT IN THE TRANSPORT CLOUD	48
3.3. INTEGRATING GEONETWORK IN THE TRANSPORT CLOUD	48
3.3.1. Installing GeoNetwork.....	48
3.3.2. Configuring GeoNetwork	48
4. CONCLUSIONS	52

List of figures

Figure 1 Excel file of Reference group metadata	11
Figure 2 Open data catalogue systems used by Reference Group of public authorities	13
Figure 3 Organisations in the MobiDataLab CKAN instance	16
Figure 4 CKAN dataset of public transport GTFS data in Rome	17
Figure 5 Managing views on a CKAN resource page	18
Figure 6 Harvest Source creation page in CKAN	19
Figure 7 Harvesting frequency configuration in CKAN	20
Figure 8 Management of harvest jobs in CKAN	20
Figure 9: OpenDataSoft web interface / portal	22
Figure 10: Menu of OpenDataSoft Backoffice: user management view	23
Figure 11: Creating a new page in OpenDataSoft	24
Figure 12: Creating a new dataset in OpenDataSoft	25
Figure 13: OpenDataSoft processors	25
Figure 14: OpenDataSoft data harvesters	26
Figure 15: ODS datasets monitoring view	28
Figure 16 GeoNetwork home page for User Administrator	30
Figure 17 Metadata creating and editing in GeoNetwork.....	31
Figure 18 Editor board in GeoNetwork	32
Figure 19 Metadata details page with ratings in GeoNetwork.....	33
Figure 20 Review and rating metadata in GeoNetwork	33
Figure 21 Publishing metadata in GeoNetwork	34
Figure 22 Metadata privileges in GeoNetwork.....	35
Figure 23 Import records in GeoNetwork	35
Figure 24 Imported metadata are filtered in GeoNetwork	37
Figure 25 Search in GeoNetwork	38
Figure 26 CSW Harvester in GeoNetwork.....	40
Figure 27 Multilingual support in GeoNetwork.....	40
Figure 28 User and profile management in GeoNetwork.....	41
Figure 29 Admin console viewed by User Administrator in GeoNetwork	42
Figure 30 Admin console viewed by Administrator in GeoNetwork.....	42
Figure 31 Templates in GeoNetwork.....	43
Figure 32 Microsoft Azure subscription for MobiDataLab	44

List of tables

Table 1 Commonly used metadata catalogues	12
Table 2 Data indicators for analytics in OpenDataSoft	27
Table 3 Mapping between Excel column and Dublin Core property	36
Table 4 GeoNetwork configurations for PostgreSQL and Elasticsearch	49
Table 5 Basic settings in GeoNetwork.....	50

Abbreviations and acronyms

Abbreviation	Meaning
API	Application Programming Interface
CKAN	Comprehensive Knowledge Archive Network
CSV	Comma-separated values
CSW	Catalogue Service for the Web
DCAT	Data Catalogue Vocabulary
EC	European Commission
ETA	Estimated Time of Arrival
FAIR	Findable, Accessible, Interoperable and Reusable
FTP	File Transfer Protocol
GIS	Geographic Information System
GTFS	General Transit Feed Specification
H2020	Horizon 2020
KPI	Key Primary Indicator
ODS	OpenDataSoft

OGC	Open Geospatial Consortium
SaaS	Software as a Service
WFS	Web Feature Service
WMS	Web Map Service
WP	Work Package
XML	eXtensible Markup Language

1. Introduction

The MobiDataLab Transport Cloud is a cloud-based prototype platform for sharing transport data, accessible to interested mobility actors. This platform, technically designed according to federated cloud principles, shows how to facilitate access to mobility data in an open, interoperable and privacy preserving way, using open tools. In particular, the aim of the MobiDataLab Transport Cloud is to make mobility data FAIR, i.e.:

- Findable – allowing the discovery of data, either static or dynamic
- Accessible – providing access to mobility data
- Interoperable – prototyping data processors for adding value to the data
- Reusable – demonstrating anonymisation and privacy-preserving tools

To showcase that the project meets these four objectives, four demonstrators are planned as part of the prototype implementation (WP4), each one of them in two versions. In the context of the H2020 programme, a demonstrator (or pilot, or prototype) is a specific type of deliverable, which differs from other deliverables in that it is not primarily in written form, although it is accompanied by a report such as the present document.

In the MobiDataLab context, where the open-source approach is favoured, demonstrators are available as a web server, a database, an open data portal, source code on a repository, etc.

1.1. Purpose of the deliverable

The following deliverable describes the first of the abovementioned Transport Cloud demonstrators, namely the reference data catalogue (version 1), which aims to improve the Findability of transport datasets in the territorial context and specific domains of the “Reference Group” of MobiDataLab stakeholders, using common catalogue software systems. This catalogue meets a double challenge:

- cataloguing transport data in the local context of the project stakeholders (that can be reused by mobility digital services like journey planners)
- cataloguing the use case data that can be used to enrich stakeholder transport datasets

Since several catalogue solutions are available in the open data ecosystem, it is often difficult for data publishers to know which solution to turn to. MobiDataLab reviewed the most used on their features, interoperability capabilities, delivery methods SaaS or on-premises, etc.

As a result, this demonstrator also shows the differences between these cataloguing solutions, the possible reasons for switching from one to the other (portability), and the relevance of combining them.

1.2. The Reference Group of mobility stakeholders

The MobiDataLab catalogue references all open transport datasets and corresponding metadata provided by the Reference Group of MobiDataLab stakeholders. Therefore, it is important to recall which organisations are involved, whether they are public authorities, operators, or international transport actors. The coordination of this reference group is the subject of the ongoing task T6.4 “multi-stakeholder group creation and coordination”.

1.2.1. Reference group of local organisations

- Comune di Roma, RSM, ATAC (Italy)
- Comune di Milano, AMAT (Italy)
- City of Leuven (Belgium)
- Brussels MIVB (Belgium)
- City of Eindhoven (Netherlands)
- Municipality of Malaga (Spain)
- Municipality of Trikala (Greece)
- Primaria Timisoara (Romania)
- Baden-Wurtemberg (Germany)
- Nouvelle-Aquitaine Mobilités (France)
- New York State (United States)

1.2.2. Reference group of international organisations

- Cubic Transportation Systems
- Tier Mobility
- Mobility Data

1.2.3. Data sources from the reference group

1.2.3.1. Initial inventory

The MobiDataLab consortium partners identified relevant datasets and organisations in the respective areas of the Reference Group. This resulted in an excel file of “metadata” in which it is possible to filter according to city/municipality/region and other criteria such as country, publisher, themes, keywords, data format, licenses, etc.

A	B	C	D	E	F	G	H	I
Data set description	Themes	Key words	Country	Producer/Publisher	Link to the source website	Type of data being represented (data set)	City/Municipality/Region	SI
The dataset contains the list of Shopping Centres on the territory of the city of Rome. In the file lists the name of the Shopping Center and the types of Exercises within the same divided in: • Neighborhood businesses (sales area of less than 250,00 Sqm) • Medium-sized Facilities (surface of sale between 250,01 Sqm 2.500,00 Sqm) • Large Structures (sales area of more than 2.500,00 Sqm)	Points of Interest, Tourism/Business	Shopping Centre	Italy		https://dati.comune.roma.it/catalogo/dataset/ds89	The shopping centers in the city of Rome.	Rome	
Stations data of the city of Stuttgart. Errors or questions directly to the city of Stuttgart.	Transportation	stations	Germany		https://www.mobidata-bw.de/dataset/stadt-mobil-stuttgart-stationsdaten	stadtmobil Stuttgart (station data)	(Bade-Wurtemberg)	
Estado de estaciones del Núcleo de Cercanías de Málaga Areas of the stop, i.e. areas in which it is made a stop on the road to be used by residents or for a fee, according to specific guidelines for each area. Each part is identified by a number that appears in the signals to stop unregulated, and in the pass that is given to the residents inside. Dataset fields name (Name): the Name of the Part scope (Scope): numerical Code identifying the Scope geom (Geometry): Geometry (MultiPolygon)	Transportation	stations	Spain	Renfe-Operadora (MINISTERIO DE TRANSPORTES, MOVILIDAD Y AGENDA URBANA)	https://datos.mib.es/es/catalogo/ea0003337-estaciones-cercanias-malaga	Estaciones Cercanías Málaga	Málaga	
The sequence of the stops that make up the locations of underground lines. Dataset fields path (path): Code number of the route id_ferm (Stop): numeric ID of the stop num (Number): sequence Number of the stop along the line L'opération renvoie une liste des arrêts successifs d'une ligne spécifique où le véhicule passe pendant son parcours. Cette information est fournie dans les deux sens (aller-retour) et à un moment spécifique. Ce jeu de données comprend : o le numéro de la ligne ; o les destinations ; o la liste de tous les arrêts desservi	Transportation, Mobility	stop, route	Italy	municipality of Milan	https://dati.comune.milano.it/dataset/ds83-infogeo-sosta-ambiti-localizzazione	Areas of stop	Milan	
	Transportation, Mobility	stop, route	Italy	municipality of Milan	https://dati.comune.milano.it/dataset/ds533-atm-composizione-percorsi-linee-metropolitane	ATM - Composition paths metropolitan lines	Milan	
	Transportation, Mobility	stops, lines	Belgium	STIB/MIVB	https://www.stib-mivb.be/fr/jgo/km/docs/resource/OpenData/be7b0f46-6608-44e6-9149-e40ab0c0ec7.pdf	Arrêts par ligne	Brussels	

Figure 1 Excel file of Reference group metadata

Based on this file, we deployed a solution built on catalogue software systems (see below), that is demonstrated in this task and that will be integrated into our "virtual lab" for data analysis hackathons. This list is constantly being updated with the help of reference group members during dedicated workshops. For example, on April 26th, 2022, a workshop was organised in this respect.

1.2.3.2. Challenges from the reference group and use case data

The scope of use case data is slightly different from the reference group data in the sense that they combine transport data and other kind of data, for example environmental data. Ideally, facilitating the discovery of data should not depend on its domain and application. However, there can be more specific methods for improving the discovery of more specialised data. In the geospatial domain, specific standards can be applied, for example the Catalogue Service for the Web (CSW) improving the discovery of spatial data. Therefore, we decided to use GeoNetwork in combination to more generic open data portals.

Datasets from the Use cases / Groups

- Data for ETA computation (traffic real-time and historic data, static map data, weather, rest time regulations, planned events like road closures etc.)
- Operational data (telematics data of vehicles, location of vehicles, completed stops, tour plans, driver shift time).
- Public transport data (static data, transportation lines, schedules, stop points, stop areas, real-time / dynamic data, disruptions, traffic alerts, next arrivals and departures, vehicle occupancy, etc.)
- Geographical data (cartography, addresses, points of interests)
- Other transport data (free floating, ride sharing, road traffic)
- National Access points for public transport data
- Environmental data portal (e.g., INSPIRE geoportal)
- Tourism data (e.g., DataTourisme)

1.3. Data providers and data consumers

The aim of this demonstrator is to show how to make mobility data easier to find, discover and reuse. This will be particularly necessary in the context of the upcoming datathon, and hackathon organised as part of the project (see WP5 – Living and Virtual Labs).

However, this single objective must be approached very differently, depending on whether one is the owner of the data to be reused or a reuser oneself. We thus need to differentiate between two categories of data catalogue users:

- data providers (or producers, or publishers) – i.e., persons or groups responsible for generating and maintaining data.
- data consumers – i.e., persons or groups accessing, using and potentially post-processing data.

Data providers aim to share data either openly or with controlled access. Data consumers (who may also be producers themselves) want to be able to find, use and link to the data.

Datasets could be used by different groups of data consumers, with different interests – which data publishers cannot all know in advance. It is necessary that the catalogue provides certain information that can help the reuse, such as structural metadata, descriptive metadata, access information, data quality information, provenance information, licensing information and usage information.

1.4. Catalogue software systems

The following table lists a set of commonly used metadata catalogues.

Table 1 Commonly used metadata catalogues

Name	Purpose	Website
CKAN	Metadata management system for data hubs led by the Open Knowledge Foundation. CKAN is an open-source solution with an active community. Many transport authorities use it as their open data portal.	https://ckan.org/
GeoNetwork	Reference implementation for geospatial data, harvesting options, network-based system, only stores metadata	https://geonetwork-opensource.org/
OpenDataSoft	Used for many open data formats, option to store data and metadata.	https://www.opendatasoft.com/
Socrata	Popular open data solution in Northern America	https://dev.socrata.com/

These different metadata catalogues each have specific capabilities for searching and managing data.

These differences correspond to different typical use cases, and the support of different metadata standards.

Public authorities from the reference group actually use different open data catalogues:

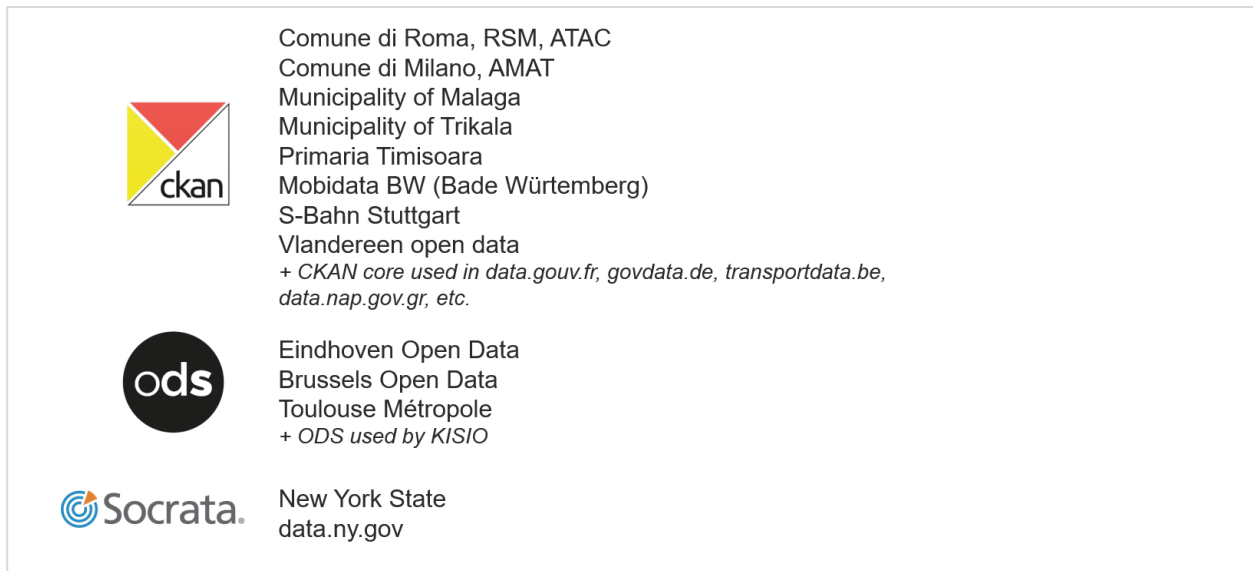


Figure 2 Open data catalogue systems used by Reference Group of public authorities

Support for as wide a range of data reuse and metadata standards as possible (e.g. DCAT, CSW, Dublin Core, etc.) is an objective of the MobiDataLab project, which will enable interoperability between catalogues and data users.

The MobiDataLab catalogue builds on several of these solutions, namely:

- CKAN
- OpenDataSoft
- GeoNetwork

These solutions follow different approaches which we propose to highlight in this demonstration.

2. Using software catalogue systems for mobility data discovery

Several catalogue software systems are available, each proposing a different approach, e.g., free open-source solution or proprietary SaaS solution. Data publishers like the reference group stakeholders have had to choose one of these solutions for their open data portal solution, after a thorough study of their needs, resources, and expectations. A general criterion for choosing between SaaS and on premises is related to human capacity for maintenance:

- Organisations that have skilled staff capable of maintaining an open data platform may find it attractive to host and maintain the solution. An open-source solution such as CKAN is often more suitable in this case.
- Organisations not having such a staff or that need a quick solution may prefer to opt for an all-in-one hosted solution such as OpenDataSoft which is more suitable in this case.

Choosing between these different approaches may also depend on the size of the organisation, the number of datasets and the frequency of updates (CKAN is interesting when there are many datasets with frequent updates).

Not only the technical and financial aspects, but also the use cases should be considered. When dealing exclusively with geo-referenced data, a GIS-specific solution like GeoNetwork could be preferable. Even though there is a degree of overlap between geographic and generic catalogues, they do not serve the exact same function. In some use cases CKAN or OpenDataSoft could be sufficient, but in a context where users would like to enrich mobility data with data from other sectors such as environmental data, it may be needed to provide also support for GIS standards (OGC, INSPIRE profiles, etc.). In this matter GeoNetwork offers serious advantages, in particular its large and well-established user base in the GIS community. For these specific users, well acquainted with geospatial Free and Open-Source Software, switching to more generic solutions like CKAN or OpenDataSoft may prove unnecessary.

Using a federation of catalogue services is also a possibility. For instance, CKAN can be used as a hub or aggregator, able to link different worlds (geo, statistics, etc.). This federated CKAN/GeoNetwork approach has, for example, been adopted by the OpenDataNetwork¹ Project which is a joint project promoted by a few Public Administration located in the Italian Region of Tuscany (Florence, Prato, Pistoia, Arno basin).

In any case, since a federated approach for the MobiDataLab Transport Cloud is required to demonstrate the portability of mobility digital services and to avoid any lock-in to a particular vendor or solution, an integrated catalogue is necessary.

¹ <http://www.opendatanetwork.it/>

2.1. Using CKAN for mobility data discovery

CKAN is an open-source open data platform used by most of the governments around the world to publish and manage their open data. The basis of CKAN is to make data more discoverable for users and reusers, for citizens, and for data consumers at large. CKAN helps governments to manage their data better and gives the ability to browse and discover data easily. CKAN is a highly extensible product, as will be shown in the following. Several useful external extensions that do not come packaged with CKAN (e.g., data storage, harvesting, CSW support, etc.) can be installed separately.

Several Reference Group members and public authorities already use CKAN to publish their transport datasets: Rome, Milan, Malaga, Trikala, Leuven, etc.

2.1.1. *Discovering data in CKAN*

CKAN is a metadata driven catalogue. It is precisely these metadata which provide a wealth of data exploration and discovery functionalities. Data can be discovered thanks to:

- Keyword search
- Tags and filters

2.1.2. *Managing organisations in CKAN*

CKAN uses organisations to represent different data publishers on the same data portal. In general, organisations correspond to public authorities, ministries, government departments, etc. In the MobiDataLab catalogue, organisations mainly correspond to reference group stakeholders.

Organisations are also part of CKAN's authorisation system and different user rights can be assigned to different organisations.

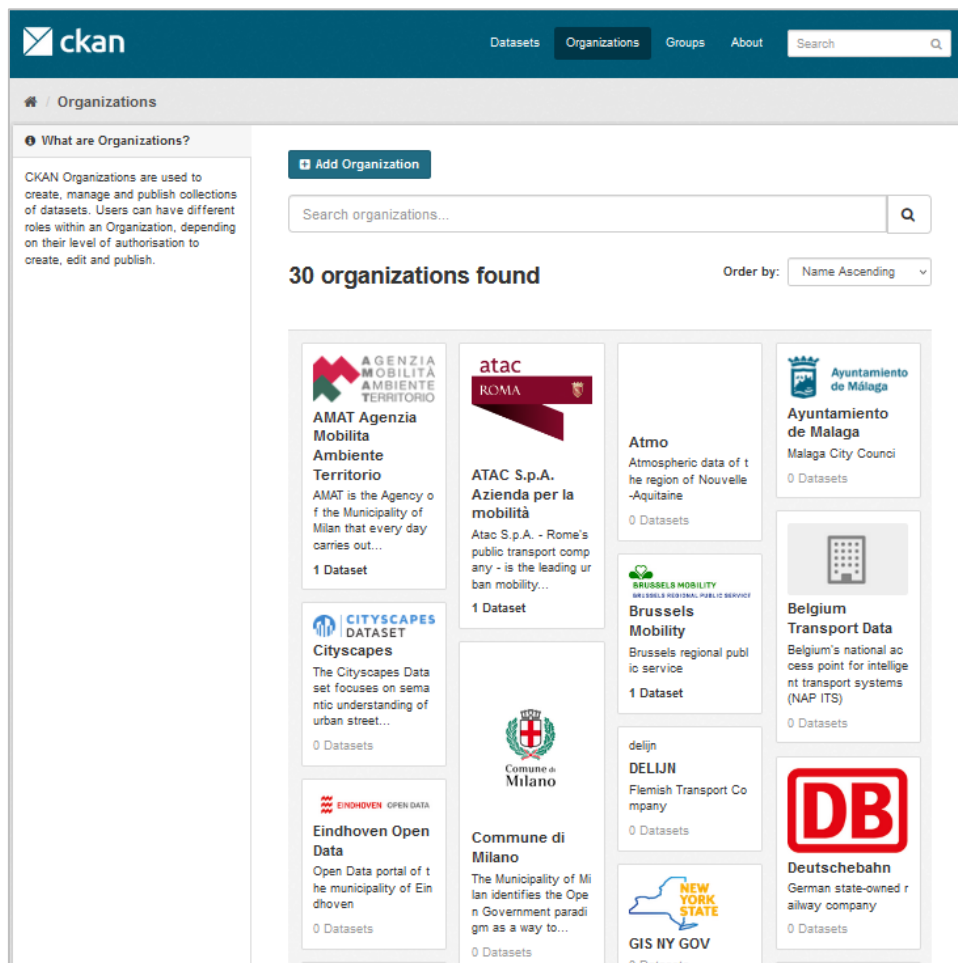


Figure 3 Organisations in the MobiDataLab CKAN instance

To create a new organization, the data publisher should log in to the CKAN portal, then go to **Organizations** and select **Add organization**. The data publisher enters its name (e.g., myCity) and a URL is automatically created.

2.1.3. Managing groups in CKAN

Another way of organizing data in CKAN are groups, which are most commonly used to categorize thematically related datasets. Groups are different from organizations, which usually represent a public authority, or any organization that runs that specific CKAN portal.

Groups are useful for organizing and curating datasets. While a dataset might belong to several groups, it can only be published by a single organization. Some typical group names might be transport, tourism, environment, or infrastructure, for example. In the MobiDataLab reference data catalogue, we choose to define groups corresponding to the use cases.

To create a new group, the data publisher should go to **Groups** and click **Add group** and enter a name (e.g., “transportation”). A URL is created automatically.

A short description helps users understand what the group is about. It is also recommended to provide an image, usually an icon, of something which visually represents the content of the data in that group, for example, a bus for transportation group. It is also possible to add custom fields to the group for refinement purposes.

2.1.4. Managing datasets in CKAN

In CKAN data is published in units called “datasets”. A dataset is a parcel of data, for example, the representation of public transport in a given municipality.

The screenshot shows the CKAN dataset page for 'Public Transport Data (GTFS) of the Municipality of Rome'. The page is divided into several sections:

- Left Sidebar:**
 - Followers:** 0, with a 'Follow' button.
 - Organization:** 'mobilità ROMA' with a logo and a 'read more' link.
 - Social:** Links to Twitter and Facebook.
 - License:** Creative Commons Attribution Share-Alike.
- Top Navigation:** Dataset, Groups, Activity Stream, and a Manage button.
- Main Content:**
 - Title:** Public Transport Data (GTFS) of the Municipality of Rome
 - Description:** Representation of Rome's public transport in GTFS format. Scheduled service and real-time data. Rappresentazione del trasporto pubblico di Roma in formato GTFS. Servizio programmato e dati in tempo reale.
 - Data and Resources:**
 - GTFS Static:** Representation of Rome's public transport in GTFS format. Scheduled service... (Explore button)
 - Trip updates:** GTFS Real Time, trip_updates.pb file with arrival forecasts for each vehicle... (Explore button)
 - Vehicle positions:** GTFS Real Time, file vehicle_positions.pb with the position of all vehicles... (Explore button)
 - Tags:** Atac, GTFS, RSM, Roma TPL, Trenitalia, mobilità
 - Additional Info Table:**

Field	Value
Source	https://dati.comune.roma.it/catalog/it/dataset/c_h501-d-9000
State	active
Last Updated	July 25, 2022, 10:49 AM (UTC+02:00)
Created	July 25, 2022, 10:41 AM (UTC+02:00)

Figure 4 CKAN dataset of public transport GTFS data in Rome

To upload a new dataset in CKAN the data publisher needs to log into the CKAN portal and go to **Datasets**. First, the data publisher should enter a descriptive title (e.g., “bus stations”) and a description. Some helpful tags such as “bus stops” or “bus stations”, separated with commas, could also be provided. A URL is then created automatically. The data publisher should choose the right license for the dataset and provide a source URL showing where the data comes from. Then she should click next **Add data**. On the second page it is possible to choose either a file from the local computer or provide a link to an external resource. Clicking on **Finish**, the dataset is uploaded.

2.1.5. Data and metadata storage

Besides metadata, CKAN can also store the data as well, in the same way as it stores metadata for data that is hosted on other platforms around the web.

The DataStore extension provides a database for storage of structured data from CKAN resources. Data can be extracted from resource files and stored in the DataStore. When a resource is added to the DataStore, the user can get automatic data previews on the resource page using the Data Explorer extension.

2.1.6. Visualising data in CKAN

The CKAN resource page can contain one or more visualizations of the resource data or file contents (a table, a bar chart, a map, etc). These are commonly referred to as resource views.

Users who are allowed to edit a particular dataset can also manage the views for its resources. To access the management interface, the data publisher should click on the **Manage** button on the resource page and then on the **Views** tab. From here she can create new views, update, or delete existing ones, and reorder them.

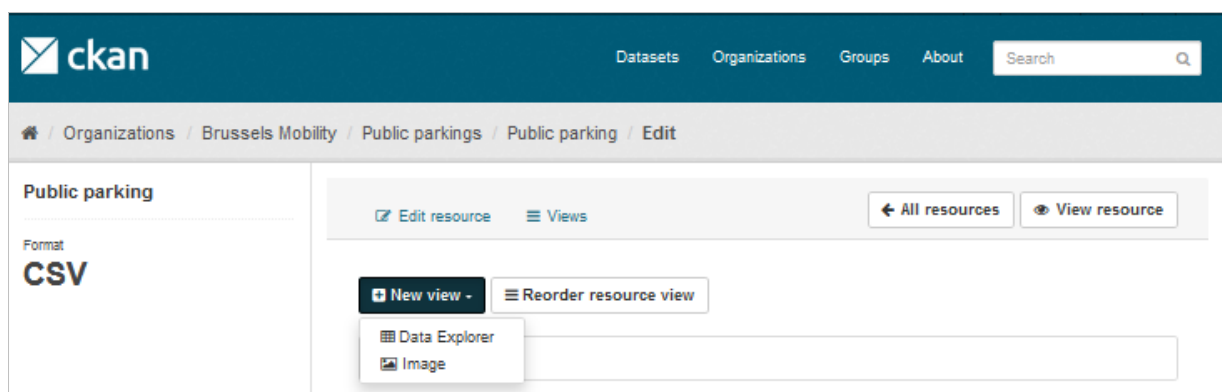


Figure 5 Managing views on a CKAN resource page

The “New view” dropdown will show the available view types for this particular resource.

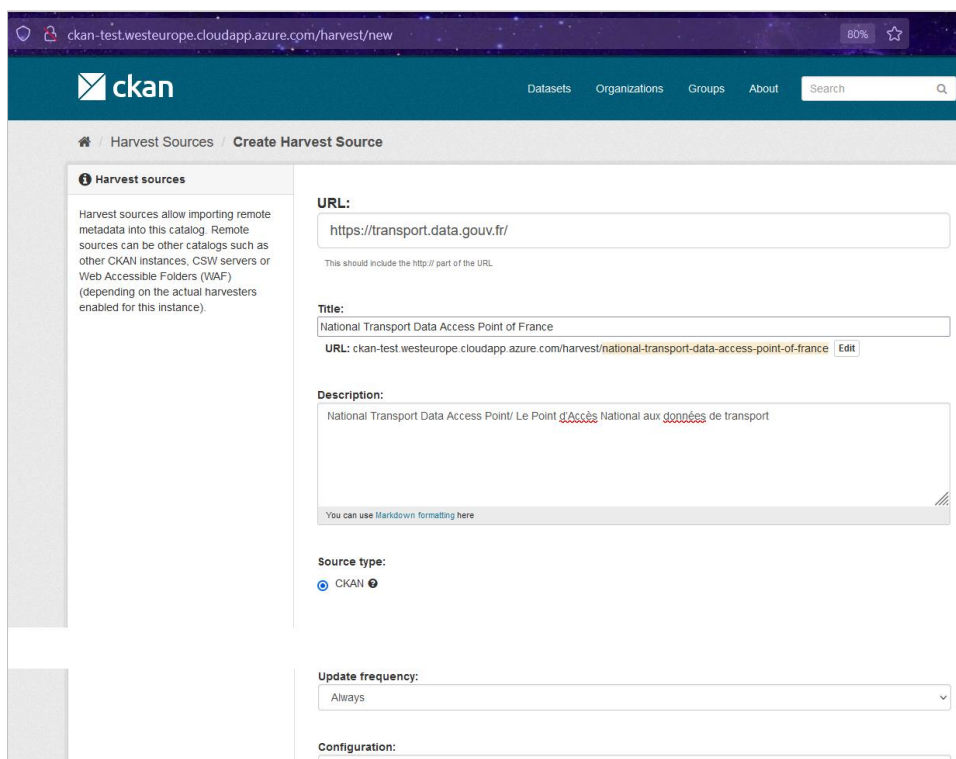
If the list is empty, you may need to add the relevant view plugins to the `ckan.plugins` setting on your configuration file (e.g. grid view, map view, etc.).

2.1.7. Harvesting

Another and more efficient way of adding datasets from open data portals is harvesting. A lot of the data in the reference data catalogue are being harvested from different data sources, namely catalogues from governments and government organizations. The MobiDataLab data portfolio can harvest them without too much effort from either side.

Let us try and harvest some data using our demo instance. What the user needs to do here is just add the URL of the source to be harvested.

In the example below we choose to harvest the data from the French National Access Point².



The screenshot shows the CKAN 'Create Harvest Source' page. The URL field contains 'https://transport.data.gouv.fr/'. The title is 'National Transport Data Access Point of France'. The description is 'National Transport Data Access Point/ Le Point d'Accès National aux données de transport'. The source type is set to 'CKAN'. The update frequency is set to 'Always'.

Figure 6 Harvest Source creation page in CKAN

In CKAN the frequency of harvesting can be set. There are a few choices: daily, biweekly, weekly or monthly.

² <https://transport.data.gouv.fr/>

ckan-test.westeurope.cloudapp.azure.com/harvest/edit/national-transport-data-access-point-of-france

You can use Markdown formatting here

Source type:
☒ CKAN ⓘ

Update frequency:
 Always
 Manual
 Monthly
 Weekly
 Biweekly
 Daily
 Always

Organization:
 le-point-d-acces-national-aux-donnees-de-transport

Delete + Save

Figure 7 Harvesting frequency configuration in CKAN

The harvesting module can even be set to constantly communicate with the data sources (always) or configured to be manually triggered.

ckan-test.westeurope.cloudapp.azure.com/harvest/admin/primaria-timisoara

ckan

Datasets Organizations Groups About Search

/ Organizations / Primaria Timisoara / Harvest Sources / Timisoa... / Admin

Timisoara City Hall - Open Data Portal
 Open data portal of Timisoara. In this portal can be found datasets from the following groups : Culture; Education; Infrastructure; Justice; Environment; Mobility; Population;... read more

Datasets
31

Dashboard Jobs Edit Reharvest Stop Clear View harvest source

Last Harvest Job

Id	9521789c-cbc6-4047-9e97-39aa0b760b8b
Created	July 22, 2022, 4:52 PM (UTC+02:00)
Started	
Finished	
Status	Running

View full job report

Figure 8 Management of harvest jobs in CKAN

2.1.8. Multilingual management

Data publishers are advised to provide human-readable metadata in multiple languages and, where possible, to provide the information in the language(s) that the intended users will understand. In the MobiDataLab project, the reference data catalogue aims to be used in the context of international living and virtual labs in the form of hackathons, datathons, innovation sessions. Therefore, the metadata needs to be made available, besides original language, at least in English.

The Multilingual CKAN extension allows the administrator to enter translations into CKAN's database. When a user is viewing the CKAN site, if the translation terms database contains a translation in the user's language for the name or description of a dataset or resource, the name of a tag or group, etc. then the translated term will be shown to the user in place of the original.

2.1.9. Interoperability

- CKAN API

The CKAN API³ exposes all the main CKAN functionalities to API clients. All the core functionality of a CKAN website (everything that can be done with the web interface and more) can be used by external code that calls the CKAN API. This is particularly useful for developers who want to write code that interacts with CKAN sites and their data.

- DCAT

The Data Catalogue Vocabulary (DCAT) is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web⁴. The CKAN DCAT extension allows CKAN to expose and consume metadata from other catalogues using RDF documents serialised using DCAT.

- CSW

CSW (Catalogue Service for the Web) is a specification from the Open Geospatial Consortium for exposing geospatial catalogues over the web. The CKAN CSW extension provides the support for the CSW standard, with the ability to import records from CSW servers with the CSW harvester, and a fully compliant CSW interface for harvested records.

³ <https://docs.ckan.org/en/2.9/api/>

⁴ <https://www.w3.org/TR/vocab-dcat/>

2.2. Using OpenDataSoft for mobility data discovery

OpenDataSoft (ODS) is a private cloud-based SaaS (Software as a Service) contrary to the CKAN solution (and GeoNetwork, see after). ODS allows the user to deliver a data catalogue of different data formats and sources, in which she can organize, share, and visualize data using some integrated tools (such as tables and graphs), as well as make data available via APIs.

Users of the software can share their data publicly as open data, but it can also be published in a limited way within an organization or a defined group for use by partners or employees for example. In this regard, ODS was chosen to create the open data portals of the City of Paris (Paris Data), the Occitania Region, Mexico City, Newark, Vancouver, and more.

Also, some private companies use ODS as their main data catalogue, as it is the case for Kisio Digital (Hove), because it allows its users to:

- Customize the catalogue to suit their needs using a user-friendly management interface/ back office.
- Provide a storage along with the catalogue, which allows its users to re-distribute their own data and make it available for public use.
- Provide an analysis chart / dashboard to broaden its data's exploitation perimeter.
- Make use of the open data provided by ODS in their Data Hub⁵ which is also accessible easily in its clients' private instances.

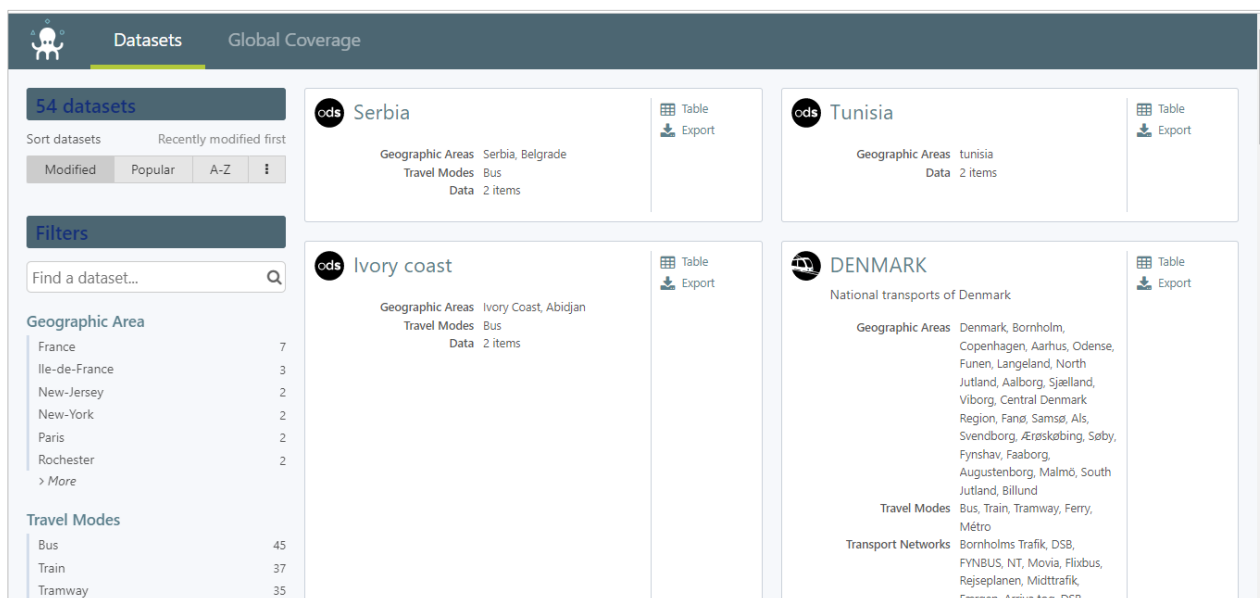


Figure 9: OpenDataSoft web interface / portal

⁵ <https://data.opendatasoft.com/explore/>

2.2.1. Customizing the catalogue: Back-office

2.2.1.1. Managing users & groups

ODS has a back-office allowing its users to administrate their portal, customize it, publish data and so on of the available features accessible through the menu.

An admin called “domain administrator” can define different levels of permissions and assign them to every user. She also can create groups to cluster users belonging, for example, to a specific organization, or customize a user’s profile.

The default users’ profiles are:

- domain administrators : Super admins, they have all the permissions.
- data publishers: they can create, publish, and manage the datasets.
- data users: they have no permissions; they can only view the datasets.
- service creators: they can only create and edit pages
- content designers: service creators with the permission to manage the theme of the portal.

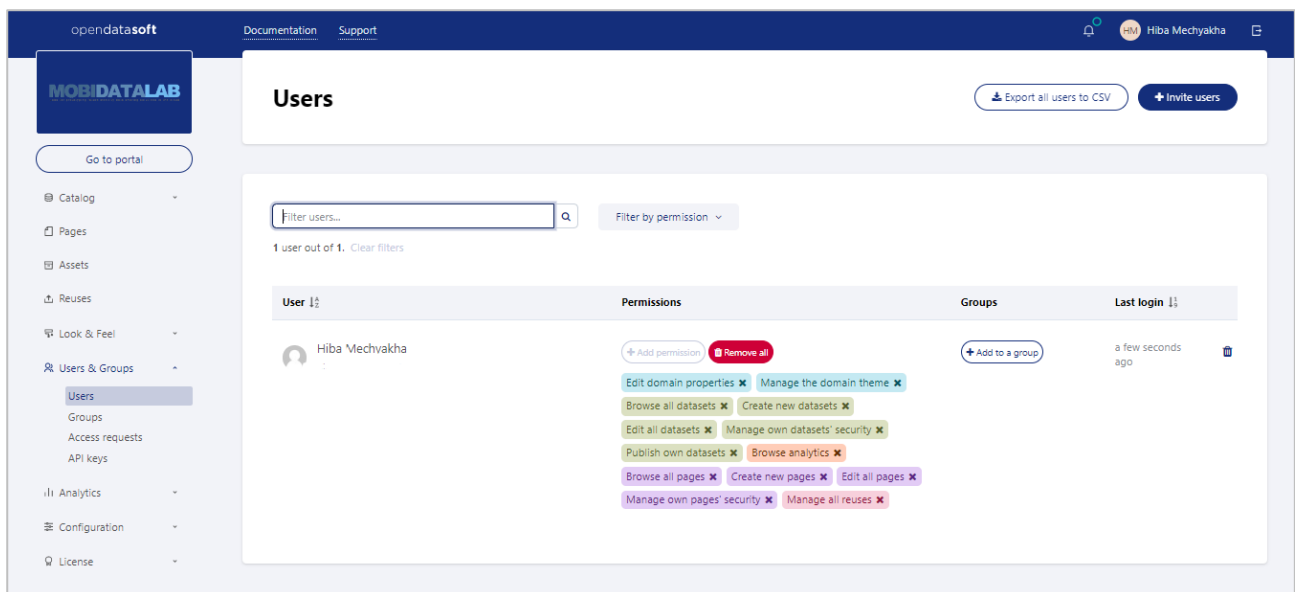


Figure 10: Menu of OpenDataSoft Backoffice: user management view

2.2.1.2. Creating pages

The content designers can either create an open or private pages using the studio provided by ODS which has text, images, KPIs and charts blocks ready to use, or do it using the classic mode which compiles HTML and CSS code to generate the new page.

There is also a code library⁶ with code snippets defining multiple components that the designers can use.

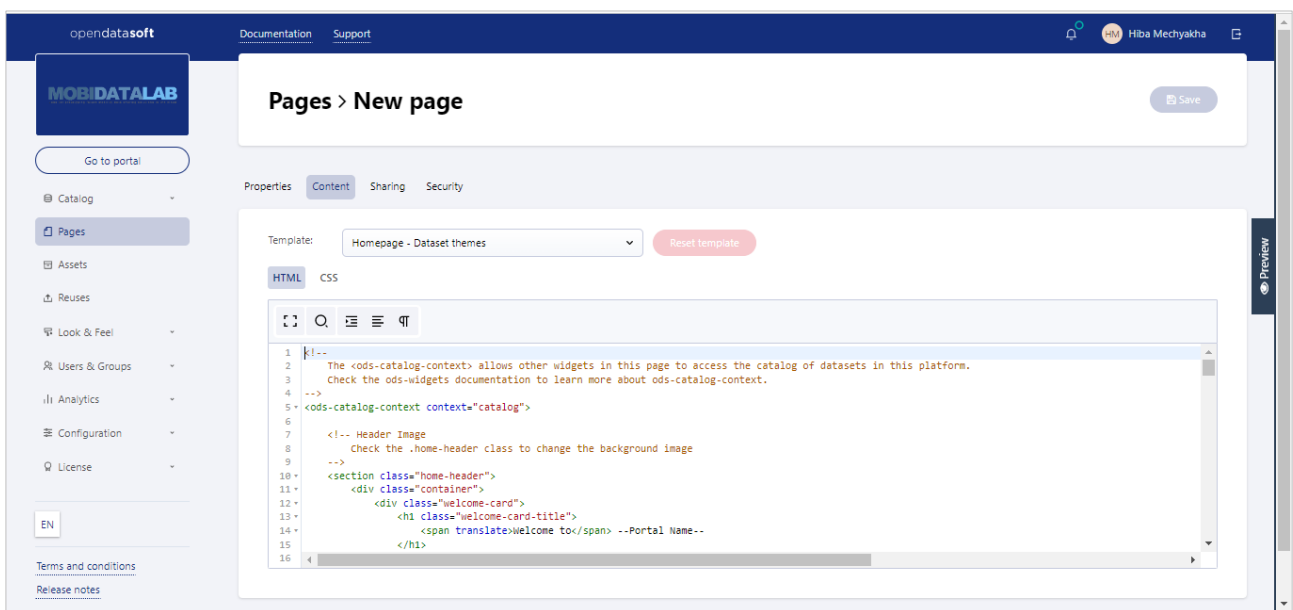


Figure 11: Creating a new page in OpenDataSoft

2.2.2. Data catalogue

2.2.2.1. Creating a new dataset

Users with data publisher permissions can create their own dataset and store it in ODS. It can also be retrieved periodically from various sources as shown in Figure 12: Creating a new dataset in OpenDataSoft.

⁶ <https://codelibrary.opendatasoft.com/components/>

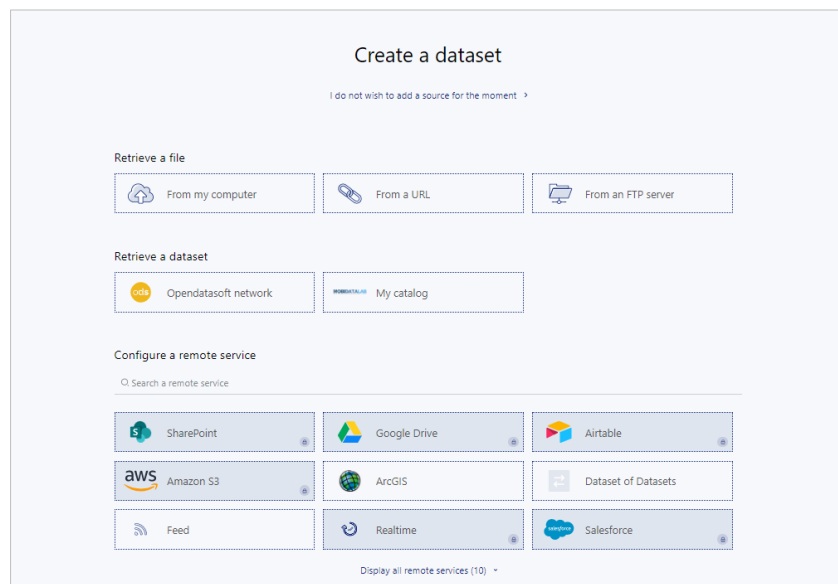


Figure 12: Creating a new dataset in OpenDataSoft

The user can apply a processor⁷ to the new dataset (like a geocoder in order to convert a human-readable address into a geo point or a split text field value and extract part of it in a new field), generate visualizations, edit the meta-data as well as manage its accessibility.

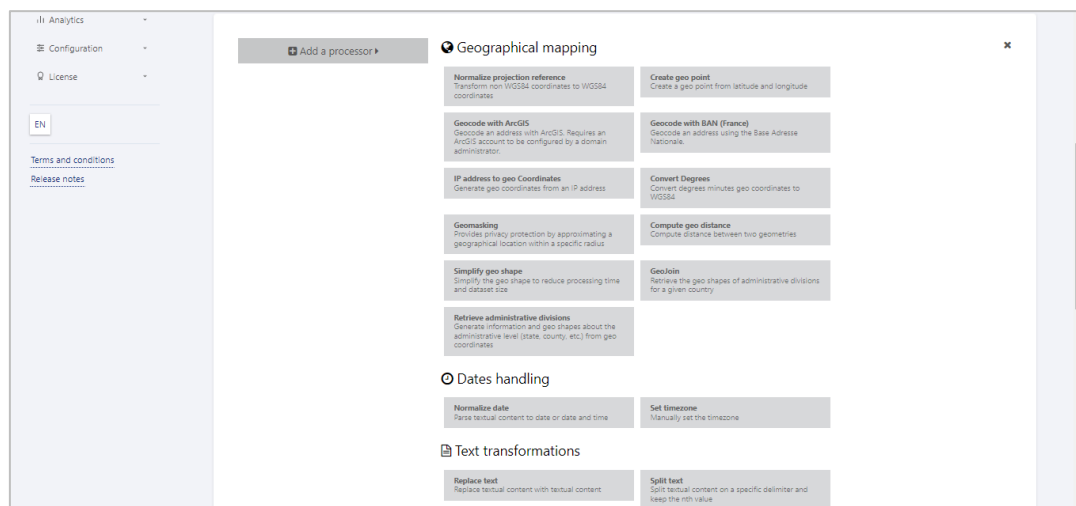


Figure 13: OpenDataSoft processors

⁷

https://help.opendatasoft.com/platform/en/publishing_data/05_processing_data/adding_processors_to_a_dataset.html

2.2.2.2. Harvesters in OpenDataSoft

To create a harvester in ODS, click on the **harvester's** menu in your back office and then on **Add harvester**.

You will be asked to choose a name and the type of the portal you want to harvest. Some options are available:

- ArcGis: it retrieves datasets from an ArcGIS server.
- ArcGIS Hub Portal: it retrieves datasets from portals listed on ArcGIS Hub.
- CKAN: it retrieves datasets from a CKAN portal.
- Data.json: it retrieves datasets from /data.json file at the root of a National Access Point (e.g., data.gouv.fr portal).
- FTP with metadata CSV: it retrieves datasets and their metadata from an FTP server.
- Opendatasoft: it harvests all the datasets or some using filters from another OpenDataSoft domain
- Socrata: like the CKAN harvester. This one retrieve all the datasets from a Socrata portal.
-

Figure 14: OpenDataSoft data harvesters

2.2.3. Data analytics

Users having access to the back-office can check data analytics of the OSD portal usage. They allow the domain administrators to monitor domain configuration activity, API calls and other KPIs.

These indicators can be refined on a specific time interval too. The table below presents a sample of some of the default ones.

Table 2 Data indicators for analytics in OpenDataSoft

Theme	Indicators
Users	<ul style="list-style-type: none"> • A timeplot of API calls, number of users • Average of API calls per user • Number and percentage of anonymous users and mobile users • TOP 5 users (API calls), map of API calls
Data	<ul style="list-style-type: none"> • Number of datasets, records, API calls and downloads • Bar chart for the TOP 5 most popular and least popular datasets using the popularity score. It is the result of a calculation that uses the number of downloads, reuses, and API calls of a dataset. The higher the score, the more the dataset is being used • TOP 5 of datasets based on downloads and API calls • Distribution chart per theme with an indication to the average popularity score
Actions	<ul style="list-style-type: none"> • Timeplot of the actions (analyze, geo and search) based on API calls • Number of text searches with no results
Quality	<ul style="list-style-type: none"> • Pie charts of publishers, licenses, and themes

OpenDataSoft datasets monitoring features generates an aggregated metadata file for all the published and unpublished datasets.

It includes: dataset_id, title, domain_id, if it was modified (processed), publisher, license, keywords, theme, reuse count, API call count, download count, attachments download count, file fields download count, popularity score, records count, records size, security (private or open), etc.

20 records

ods-datasets-monitoring

Active filters: visibility, domain_id: mobidalab

Filters: Search records... Q

dataset_id	title	domain_id	modified	publisher	license	keyword	theme
1	us-ca	USA : California State	mobidalab	September 11, 2015			Transports, Déplacements
2	cz	Czech Republic	mobidalab	February 13, 2017			Transports, Déplacements
3	us-wa	USA : Washington State	mobidalab	September 15, 2015			Transports, Déplacements
4	fr-nw	FRANCE : North-West Quarter	mobidalab	September 16, 2015			Transports, Déplacements
5	ie	IRELAND	mobidalab	December 11, 2015			Transports, Déplacements
6	lu	LUXEMBOURG	mobidalab	July 22, 2016			Transports, Déplacements
7	pl	Poland	mobidalab	July 18, 2022		transports:GTFS:Poland:Pologne:Pois...	Transport, Movements
8	us-tx	USA : Texas State	mobidalab	September 15, 2015			Transports, Déplacements
9	fr-se	FRANCE : South-East Quarter	mobidalab	September 16, 2015			Transports, Déplacements
10	ca-on	CANADA : Ontario State	mobidalab	February 14, 2018		Canada:Ontario	Transports, Déplacements
11	us-mi	USA : Michigan State	mobidalab	March 28, 2022			Transports, Déplacements
12	iceland	Iceland	mobidalab	July 18, 2022		Open Database License (ODbL)	Transports, Déplacements
13	gtfs-lille-20140801_v3	GTFS-Lille-20140801_v3	mobidalab	July 18, 2022			
14	uk	UNITED-KINGDOM	mobidalab	September 16, 2015			Transports, Déplacements:TrainSub.
15	se	SWEDEN	mobidalab	September 16, 2015			Transports, Déplacements
16	hu	HUNGARY	mobidalab	October 12, 2015		Hungary:Hongrie:Magyarország:Bud...	Transports, Déplacements
17	au	AUSTRALIA	mobidalab	November 19, 2015			Transports, Déplacements
18	ca-mb	CANADA : Manitoba state	mobidalab	February 14, 2018		Canada:Manitoba	Transports, Déplacements
19	ghana	Ghana	mobidalab	March 13, 2018			Transports, Déplacements
20	ci	Ivory coast	mobidalab	July 18, 2022			Transports, Déplacements

Figure 15: ODS datasets monitoring view

2.2.4. OpenDataSoft APIs

OpenDataSoft provides access to 5 APIs:

- [ODS Explore API V2](#): main OpenDataSoft explore API, used to explore catalogues and datasets with a custom SQL-like query language: ODSQL.
- [Triple Pattern Fragments API](#): ODS API for triple pattern querying over datasets from OpenDataSoft portals in Resource Description Framework (RDF) format.
- [OData](#), [WFS](#), and [CSW](#): 3 standard protocols supported and provided by OpenDataSoft. OData is a standard for REST APIs that provides a common language to be used across APIs to perform requests, whereas WFS and CSW are standards focusing on geographic data. They are especially relevant, for example, to interface the platform with other GIS software.

All these APIs provide access to any data pushed to the platform, no matter their source or format, if the security rules defined by the data owner allow that access.

These APIs can be used, for example, to search for datasets and data, to compute analysis, or to perform geographic aggregations.

2.2.5. Integrating reference data into KISIO OpenDataSoft instance

KISIO integrated seven municipalities' transport datasets from the reference data catalogue (Rome, Milan, Eindhoven, New York, Malaga, Leuven, and Hamburg) into its OpenDataSoft instance.

The datasets (GTFS files) were exported to a private FTP server with a csv file containing the metadata:

- **Description** of the dataset
- **Download** as in name of the file
- Format
- ID
- License
- License link
- Size
- Update date
- Validity end date
- Validity start date

Using an FTP harvester, these datasets were added to the catalogue.

2.3. Using GeoNetwork for mobility (geo-referenced) data discovery

The world of geo-data is highly standardised, including regarding discoverability. For example, the European INSPIRE directive recommends that each producer of geographical data must publish it on the Internet, the consistency of the system being ensured by the cataloguing of metadata (i.e., information describing the data, and facilitating their inventory, search and use). The INSPIRE directive requires that each spatial dataset within its scope be described by metadata, and that these metadata be kept up to date and, like the data, published on the Internet. These geo-data and their metadata follow specific standards, like the OGC (Open Geospatial Consortium) standards, which makes the interoperability of geographical data services possible.

GeoNetwork is a metadata catalogue specifically designed for geographical information (GIS). It is based on open standards, offers powerful metadata editing and search functions, and integrates an interactive web map viewer. The following sections aim to demonstrate the capabilities of the GeoNetwork catalogue.

2.3.1. Web interface

GeoNetwork web interface is available under the path `/srv/eng/catalog.search#/home`. The webapp folder name in Jetty or Tomcat webserver normally needs to be included in the complete URL.

The web interface has five main pages: Home, Search, Map, Contribute and Admin.

1. Home page shows an overview of the metadata in GeoNetwork by Topic and Category.

2. Search page (section 2.3.3) shows extensive search and filter functionalities and search results.
3. Map page shows an interactive world map to visualize supported data source, for instance WFS, WMS data.
4. Contribute page contains several subpages, notably: Editor board, Add new record, Import records.
5. Admin console (section 2.3.9) contains several subpages, notably: Metadata & templates, Users and groups (section 2.3.8), Harvesting, Settings, Tools.

Unauthorized users are able to see only the Home page, Search page and Map page, which can also be adjusted in Settings. The Contribute page is accessible for users with Editor profile or above. The Admin console is reserved for users with User Administrator profile and above. Details on User and groups and their permission management are discussed in section 2.3.8.

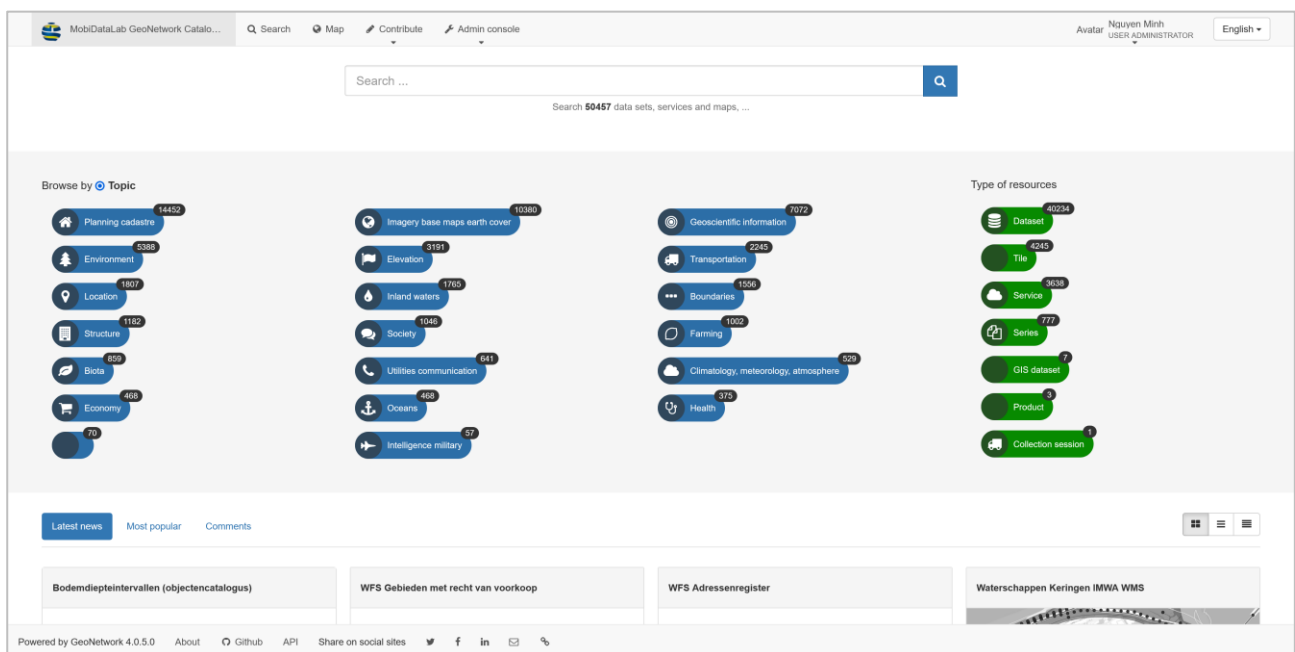


Figure 16 GeoNetwork home page for User Administrator

2.3.2. Managing metadata in GeoNetwork

Metadata management is done in Contribute page, which is only accessible to authorized users with Editor profile or above. Once eligible users are logged in, the tab menu **Contribute** will appear.

2.3.2.1. Create metadata

Prior to creating metadata, the metadata templates need to be imported by an Administrator (section 2.3.9).

To create metadata, the user should navigate to menu **Contribute > Add new records**, then select the appropriate template, click **Create** and proceed with filling in the required information of the dataset. The metadata of the dataset can be validated against the metadata schema. Once done, the user should click on the **Save** button, then the metadata is created and only accessible to the owner. If the user wishes to make the created metadata accessible to everyone, or another specific user group, the user needs to publish the metadata (section 2.3.2.4).

Figure 17 Metadata creating and editing in GeoNetwork

2.3.2.2. Edit metadata

To edit an existing metadata, the user should navigate to the menu **Contribute > Editor board**. The editor board contains a search box on top, filtering panel on the left and search results panel on the right.

In addition, an editing toolbar is placed on the right of editable metadata. The user can search for, filter and select a specific metadata, then click on the **Edit** button on the right (Figure 18). The editing page (Figure 17) shows up, allows the user to make adjustments, to validate and to save the changes.

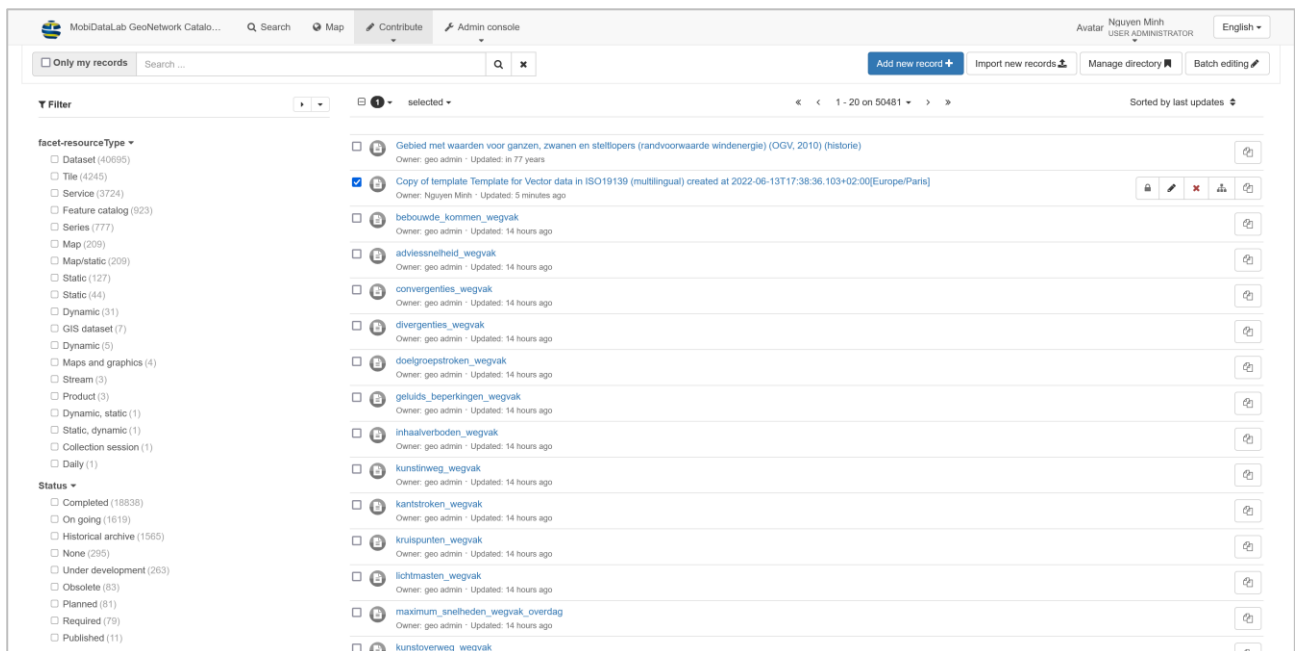


Figure 18 Editor board in GeoNetwork

2.3.2.3. Review metadata

From the Editor board, the user can view a selected metadata without editing by clicking on the metadata name. The metadata details page shows all the information of the dataset input by the editor. On the right panel, the ratings given by users are shown together with the plus button for adding new rating. Clicking this plus button opens the review page for the user to input their ratings and reviews (Figure 20).

Once the reviews are saved, the overall ratings will be reflected in the metadata details page (Figure 19). Reviews can also be removed once they are no longer relevant.

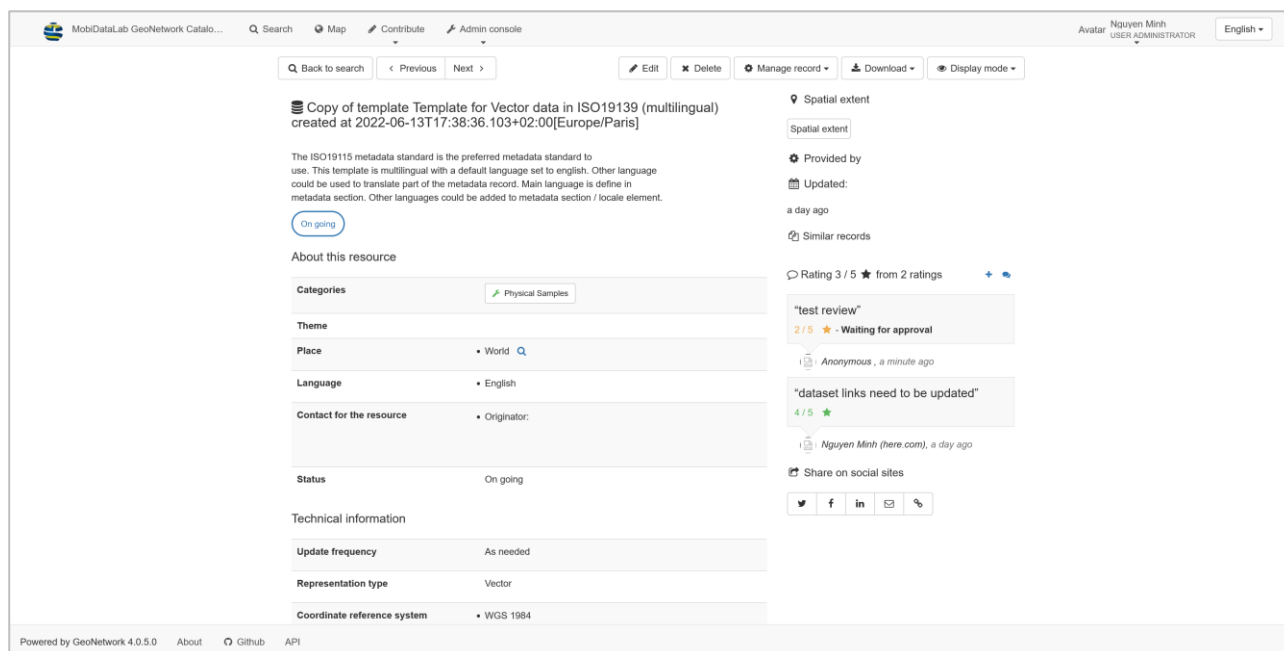


Figure 19 Metadata details page with ratings in GeoNetwork

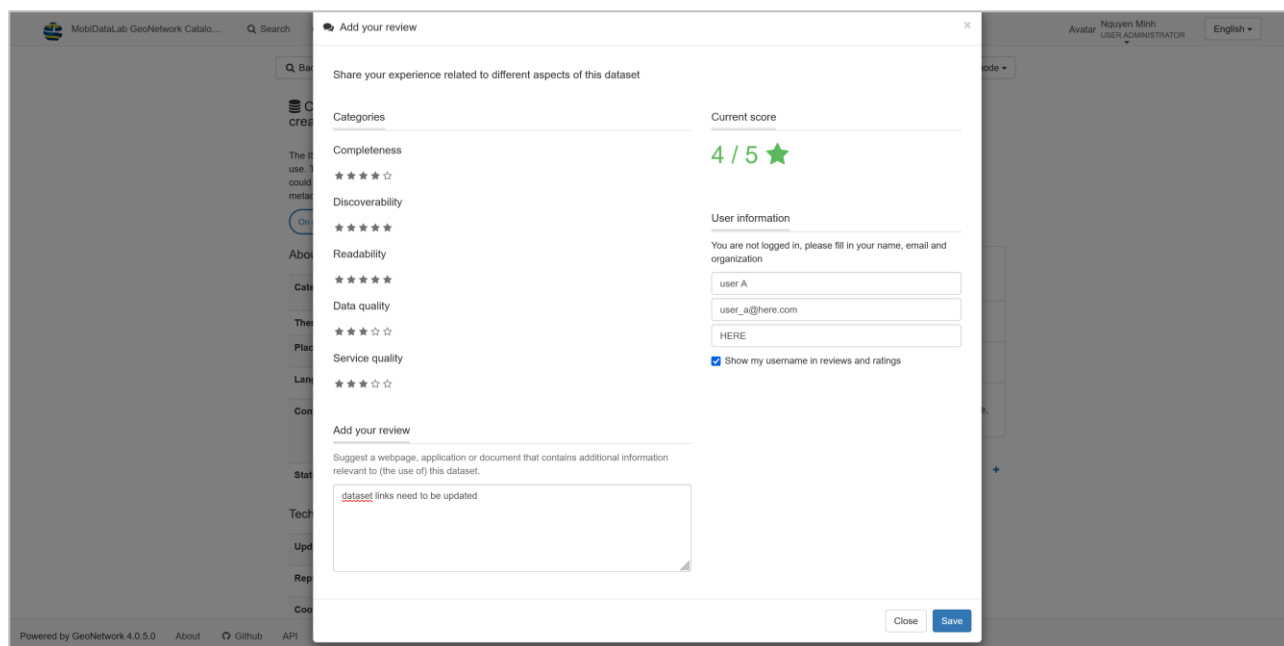


Figure 20 Review and rating metadata in GeoNetwork

2.3.2.4. Publish metadata

From the Editor board, the user should navigate to the button selected > **Publish**, to publish all the selected metadata (Figure 21).

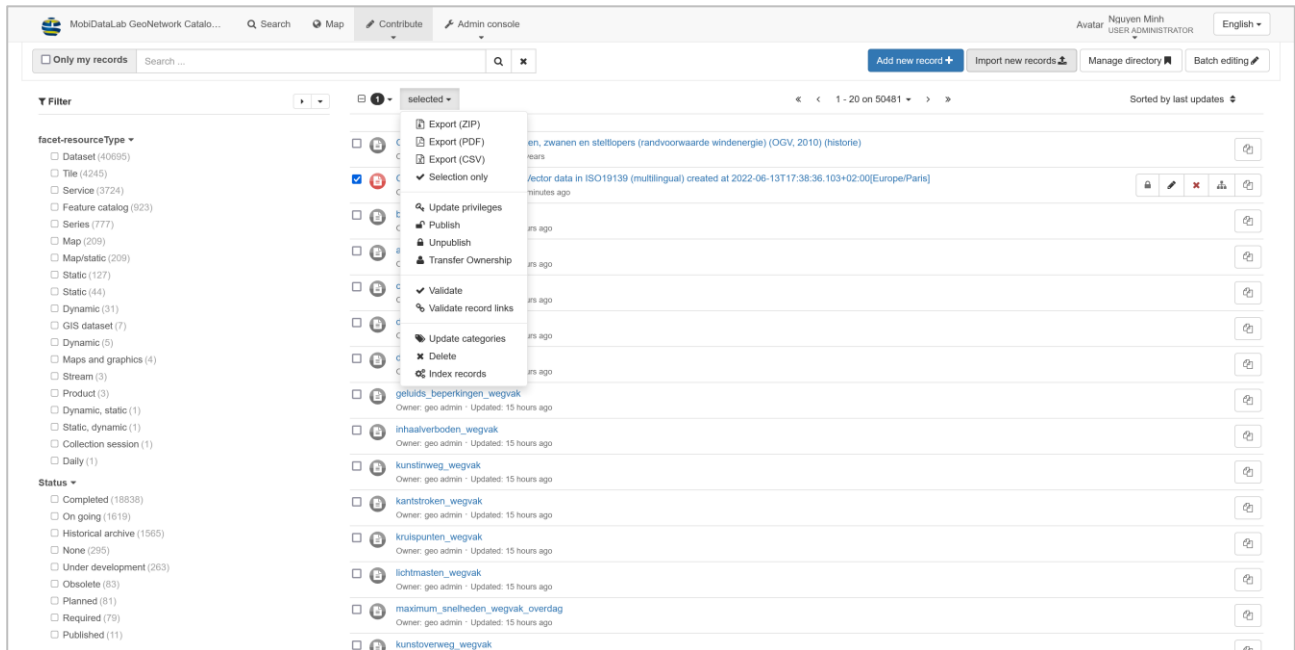


Figure 21 Publishing metadata in GeoNetwork

Published metadata are accessible by all users, including unauthorized ones. If public access is not desired, the user can instead navigate to the button selected > **Update privileges** to open the privileges popup.

It shows different user group and their privileges on the selected metadata (Figure 22).

The user should check the appropriate box to give certain user groups access to the selected metadata. Then the user can choose either to replace or to merge the privileges.

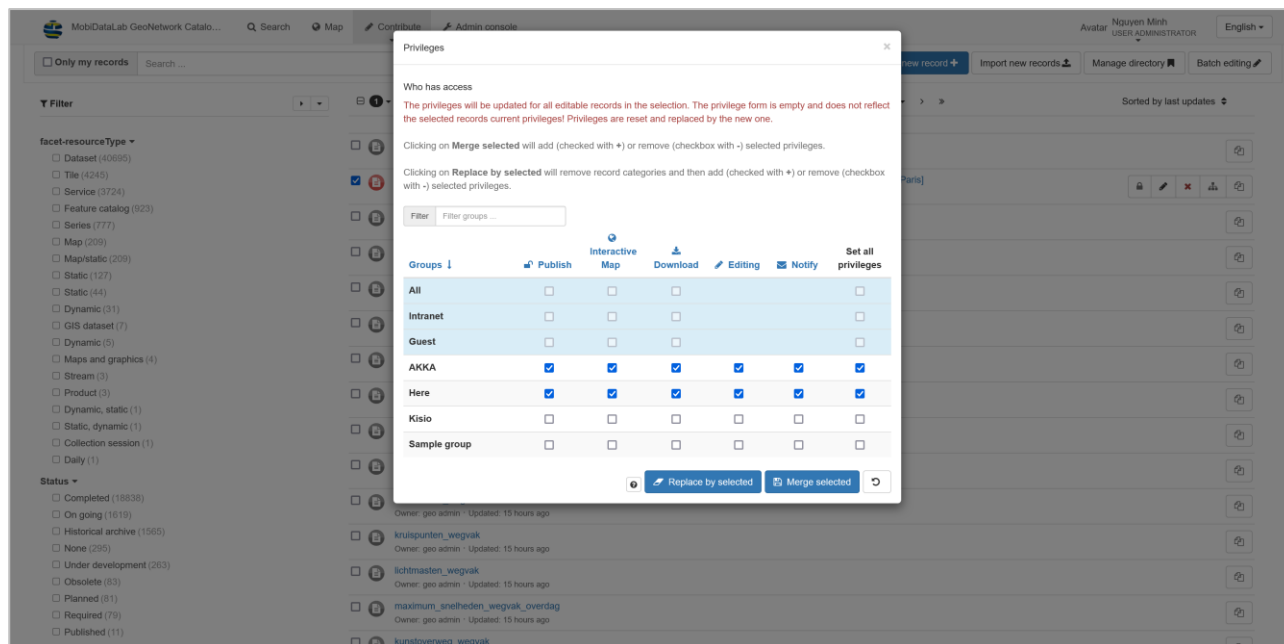


Figure 22 Metadata privileges in GeoNetwork

2.3.2.5. Import metadata

From the menu **Contribute > Import new records**, the user can choose to import metadata from a file, from the local machine or from a remote location (Figure 23). The metadata needs to be an XML file with supported schema format, namely ISO 19139, ISO 19115, ISO 19110 or Dublin Core. If the source metadata is not supported, they need to be converted to one of the supported formats above.

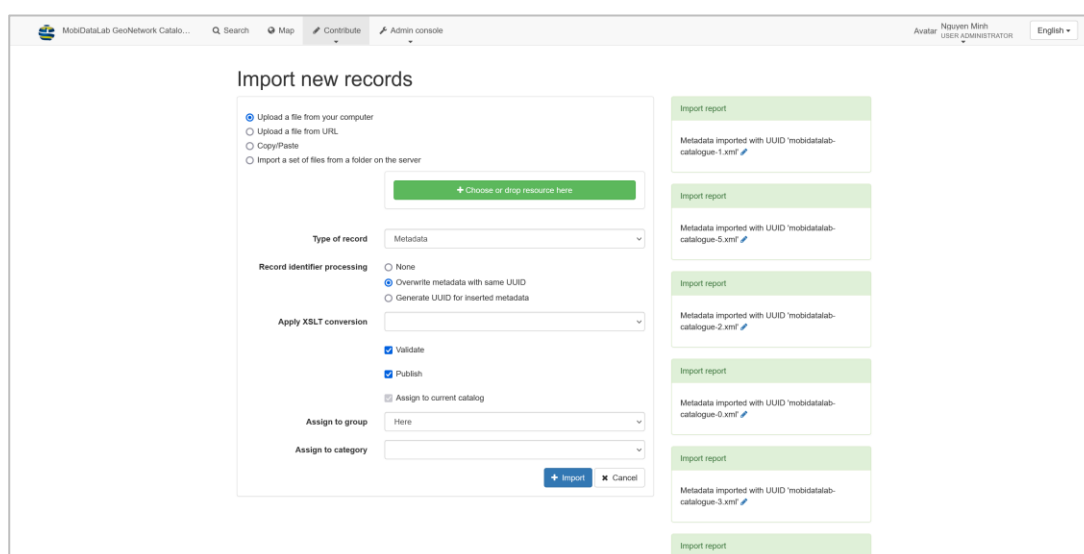


Figure 23 Import records in GeoNetwork

Table 3 shows a simple mapping used in the conversion from the reference data catalogue in Excel format to the Dublin Core format.

As a result, 439 rows in Excel sheet have been converted to 439 Dublin Core XML metadata file.

Table 3 Mapping between Excel column and Dublin Core property

Column name in Excel	Property name in Dublin Core
Contributor	dc:contributor
City/Municipality/Region	dc:coverage
Creator	dc:creator
Data set description	dc:description
Data set description	dct:abstract
Data format	dc:format
Country	dc:language, dc:subject
Producer/Publisher	dc:publisher
License	dc:rights
Key word	dc:subject
Themes	dc:subject
Link to the source website	dct:references

These imported records can be filtered and further edited if required (Figure 24).

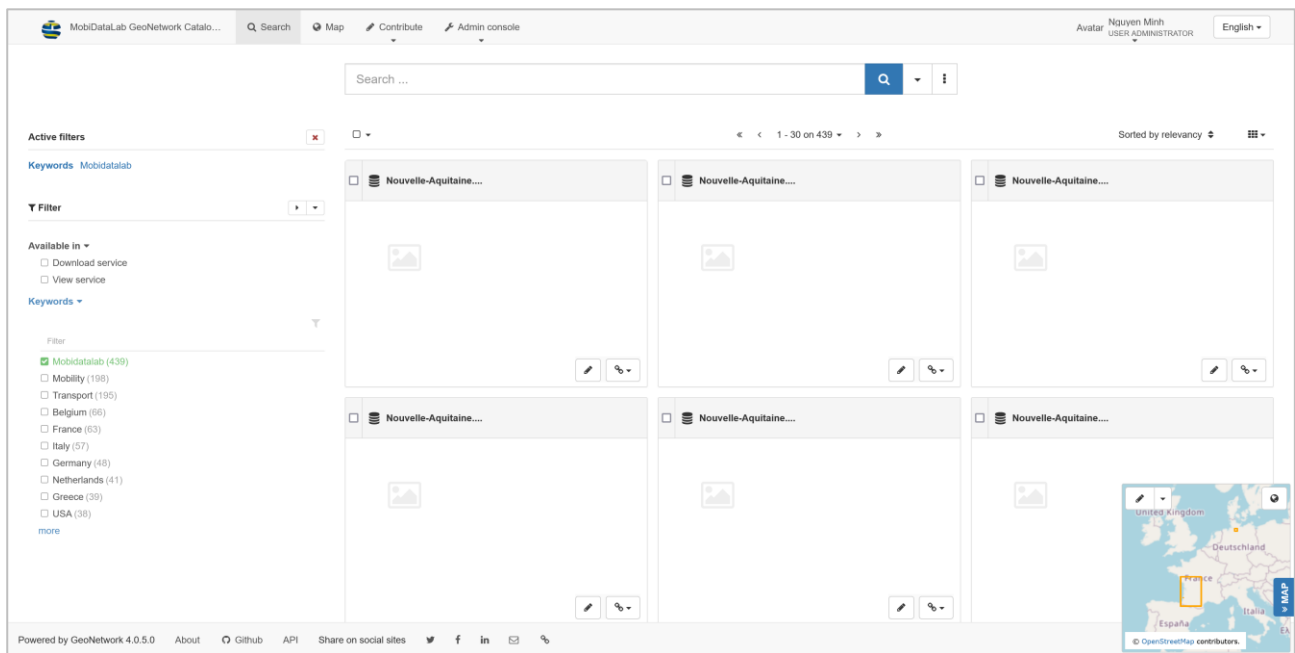


Figure 24 Imported metadata are filtered in GeoNetwork

2.3.2.6. Export metadata

Selected metadata can be exported in their original schema format as a ZIP file via the menu selected > **Export (ZIP)**.

2.3.3. Data discovery

The Search page contains a search box on top for text search and temporal search, a mini map for spatial search, filtering panel on the left and search results panel on the right. In addition, an editing toolbar is placed on the right of editable metadata.

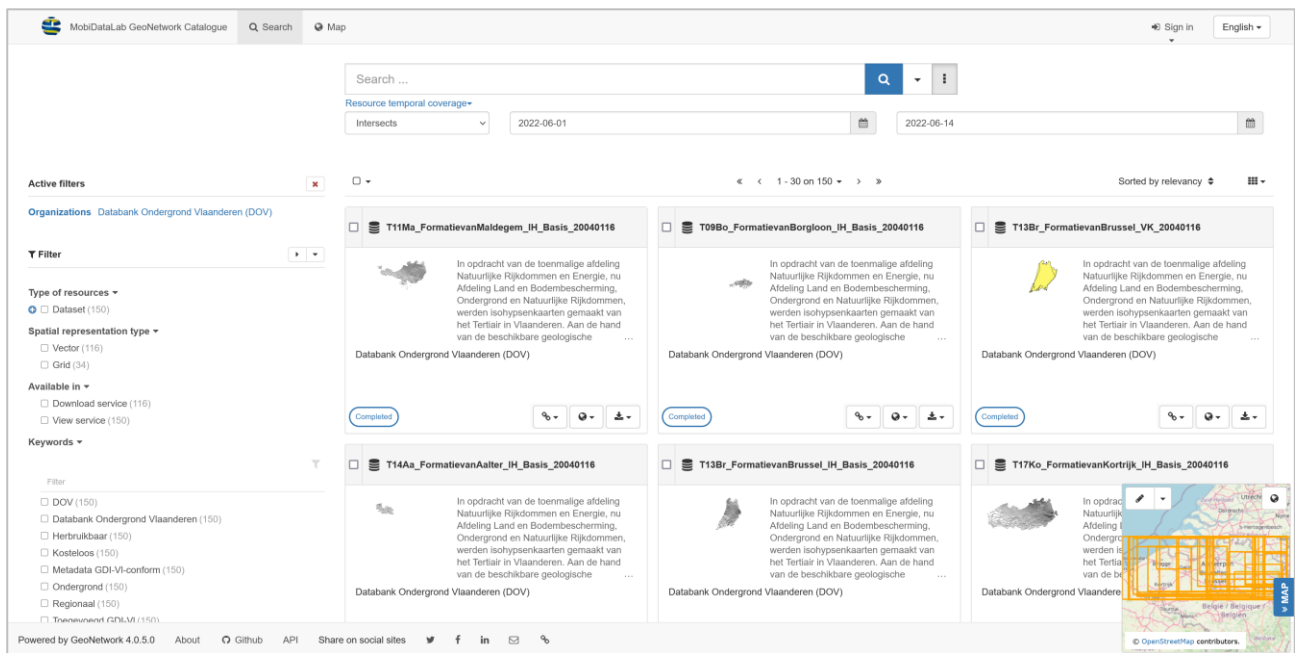


Figure 25 Search in GeoNetwork

- Text search and filter

The user can use the search box on top to search for a given query string in all accessible metadata in GeoNetwork. The filtering options on the left panel allows the user to filter only metadata with specified attributes, such as keywords, organizations, data types, etc. Which attributes appear in the filter panel can be defined in Settings.

- Spatial search

The user can use the pen tool on the mini map to draw the extent for spatial search. The metadata whose extent intersects with the query extent are returned.

- Temporal search

The three-dot button next to the search box enables temporal search capability. The user can then specify the time range as query.

2.3.4. Visualising data in GeoNetwork

GeoNetwork allows the users to visualize standardized external data source directly on the map viewer. Supported data sources are WMS, WFS, WMTS, KML services and KMZ file.

2.3.5. Interoperability

- CSW support

The Catalog Service for the Web (CSW) end point exposes the metadata records in your catalogue in XML format under the path /srv/eng/csw? using the OGC CSW protocol (version 2.0.2), specifically the CSW and CSW-T protocols

1. CSW: Provides the ability to search and publish metadata for data, services and related information.
2. CSW-T: Provides an interface for creating, modifying and deleting catalogue records via the CSW protocol.

A typical CSW request sent to GeoNetwork takes the following form⁸:

<http://localhost:8080/geonetwork/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities>

- GeoNetwork API

In version 4, GeoNetwork API description uses OpenAPI specification and is served under /srv/api/ path. GeoNetwork Swagger page is found under path /doc/api

- DCAT-AP support

DCAT-AP is only supported till GeoNetwork version 3, it is not supported in version 4.0.x yet.

- Opensearch

Opensearch is only supported till GeoNetwork version 3, it is not supported in version 4.0.x yet.

2.3.6. Harvesting

The user needs to have at least a User Administration profile to manage catalogue harvesters. From the menu **Admin console > Harvesting > Catalog harvesters**, users can add and manage catalogue harvesters (Figure 26). Some supported protocols for harvesting are OGC CSW 2.0.2,

⁸ More details about CSW support in GeoNetwork can be found in <https://geonetwork-opensource.org/manuals/4.0.x/en/api/csw.html>

OGC WFS, Geonetwork, Geoportal, ArcSDE. The harvesters shall be scheduled to run during off-hours to avoid degrading performance.

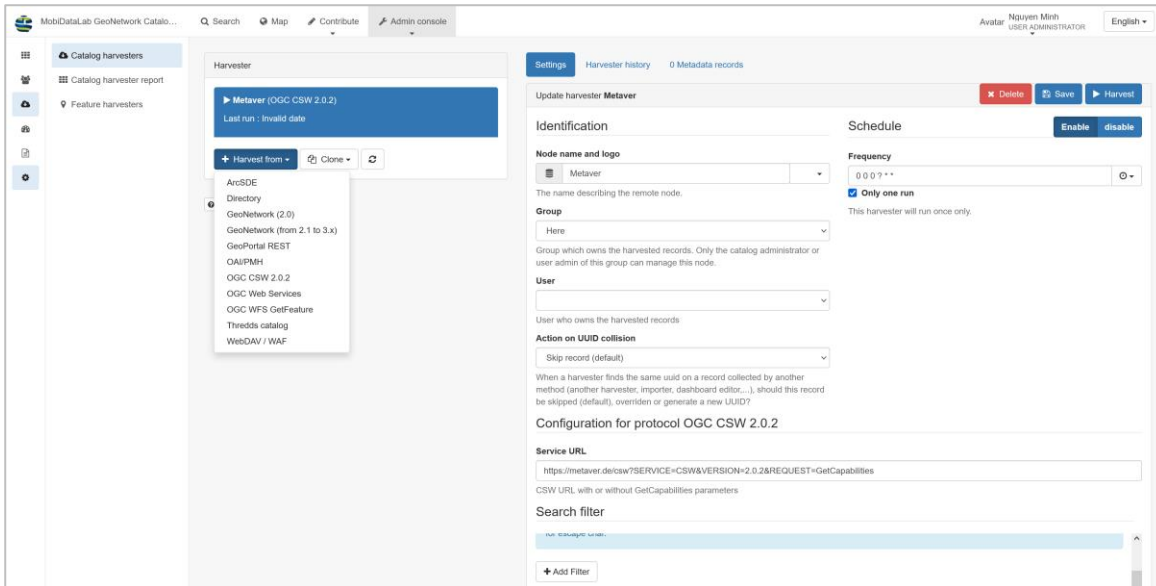


Figure 26 CSW Harvester in GeoNetwork

2.3.7. Multilingual management

GeoNetwork web interface supports 16 languages. Multilingual support in metadata relies on the schema definition, for instance, in ISO 19139 schema, property “gmd:locale” is used to specify translation of the metadata in different languages.

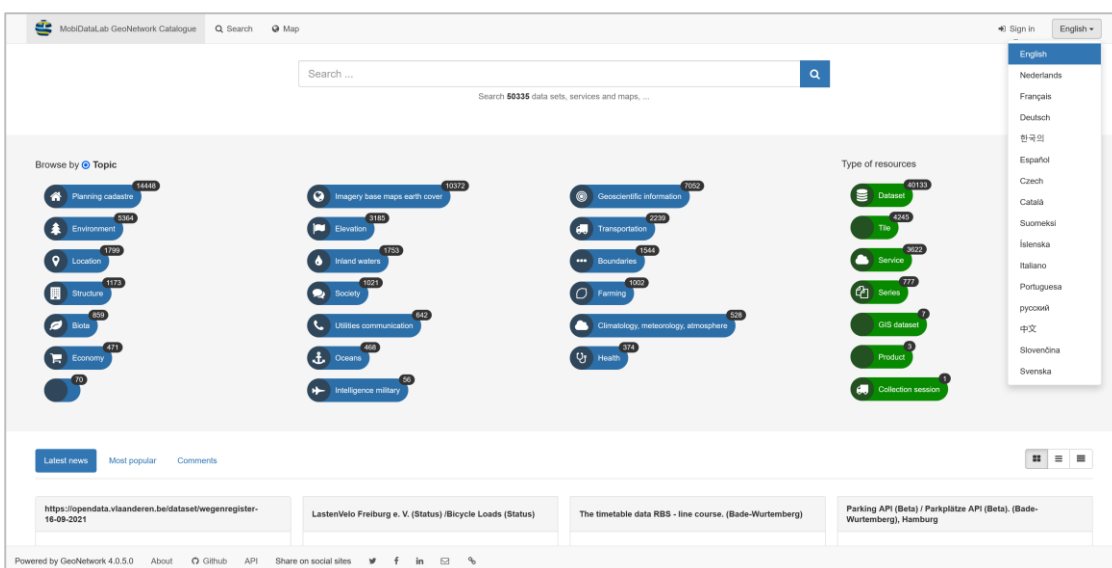


Figure 27 Multilingual support in GeoNetwork

2.3.8. Permission management

Permissions in GeoNetwork are managed via Users and groups, which can be found in Admin console (Figure 28).

There are 4 user profiles in GeoNetwork, representing 4 privileges level:

1. Registered user: has read permission
2. Editor: has read and write permission
3. Reviewer: has read, write and publish permission
4. User administrator: has read, write, publish and harvester configuring permission

Each user is assigned with a user profile that grants them permission to metadata belonging to a user group. Matrix-based permission enable users to play different roles in different groups.

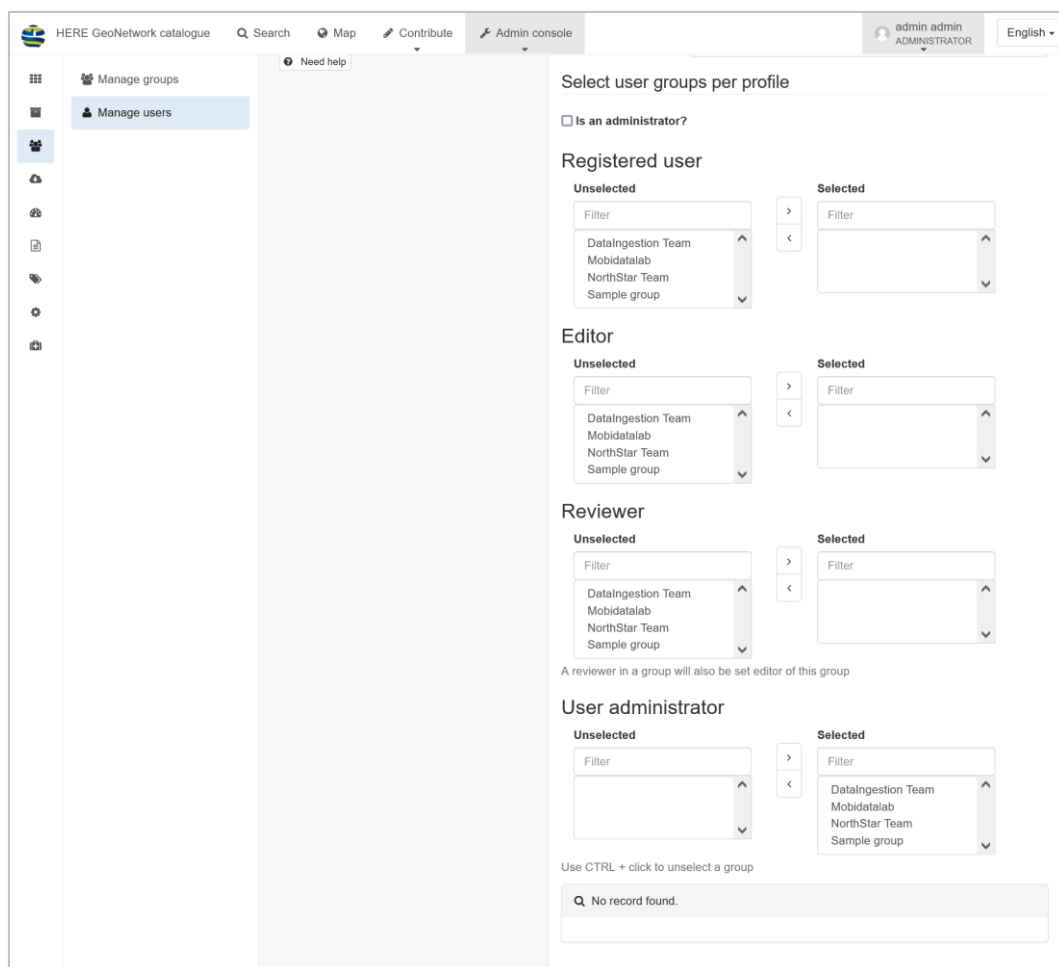


Figure 28 User and profile management in GeoNetwork

2.3.9. Admin console

The Admin console can be viewed by any user having User Administrator (Figure 29) or Administrator (Figure 30) profile. To enable the capability to Create metadata in GeoNetwork, default schema templates need to be loaded in **Admin console > Metadata** and templates (Figure 31).

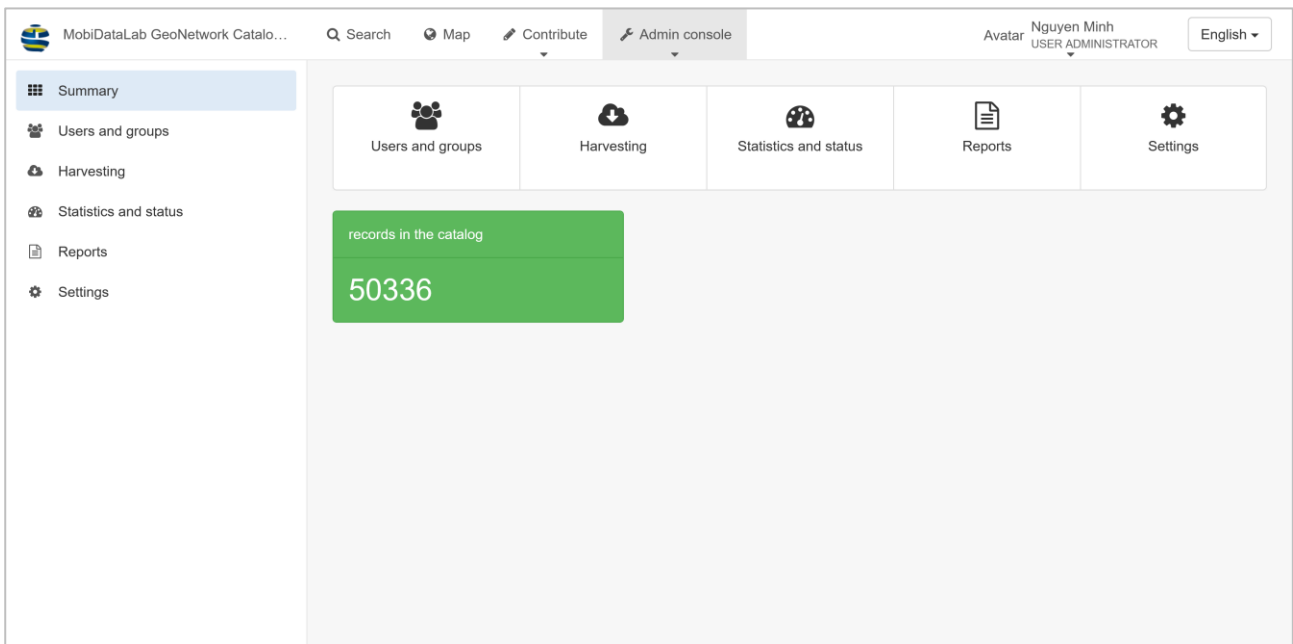


Figure 29 Admin console viewed by User Administrator in GeoNetwork

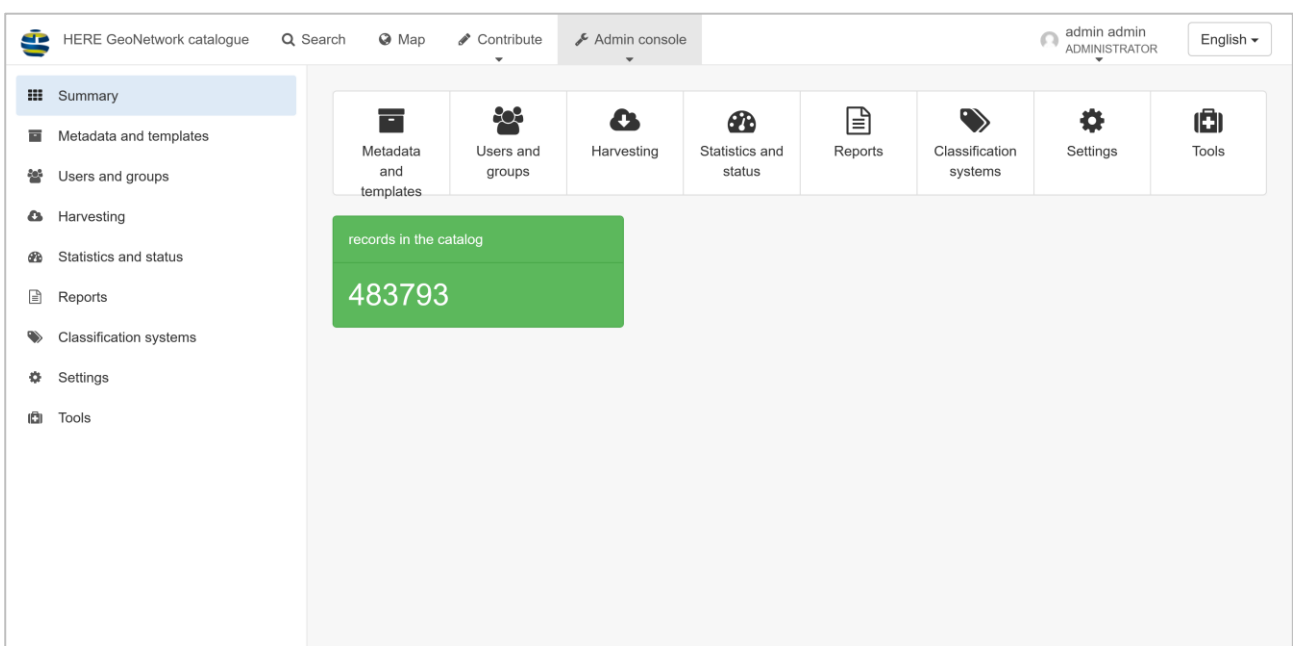


Figure 30 Admin console viewed by Administrator in GeoNetwork

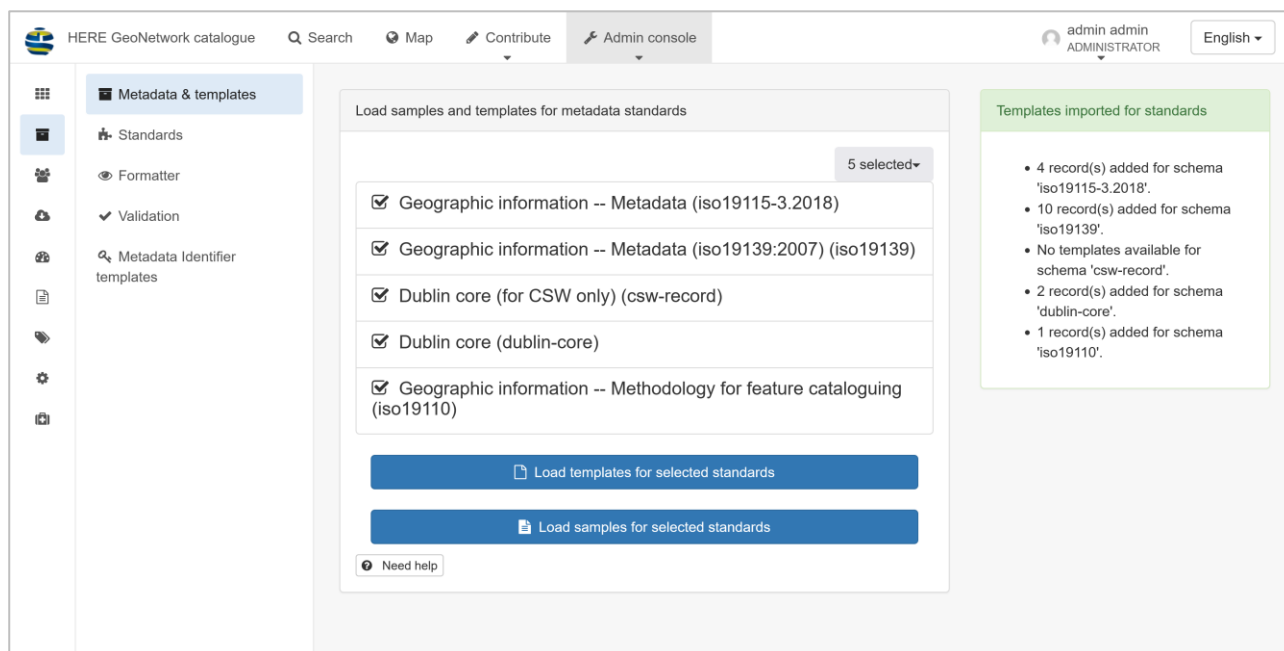


Figure 31 Templates in GeoNetwork

3. Integrating catalogue software systems into the MobiDataLab Transport Cloud

The Transport Cloud prototype and its catalogues have been deployed on a Microsoft Azure cloud infrastructure, provided by AKKA IT services. Even though eliciting a specific cloud provider may lead to vendor dependency, cloud agnosticism has remained a strong requirement for the MobiDataLab Transport Cloud, meaning that only cloud technologies which are portable (i.e., commonly available regardless of the cloud provider), must be used.

Portability means that the integrated catalogue system must not tie its usability to one particular tool, version, flavour, or vendor, which would put at risk any upgrade or evolution and jeopardize the whole platform.

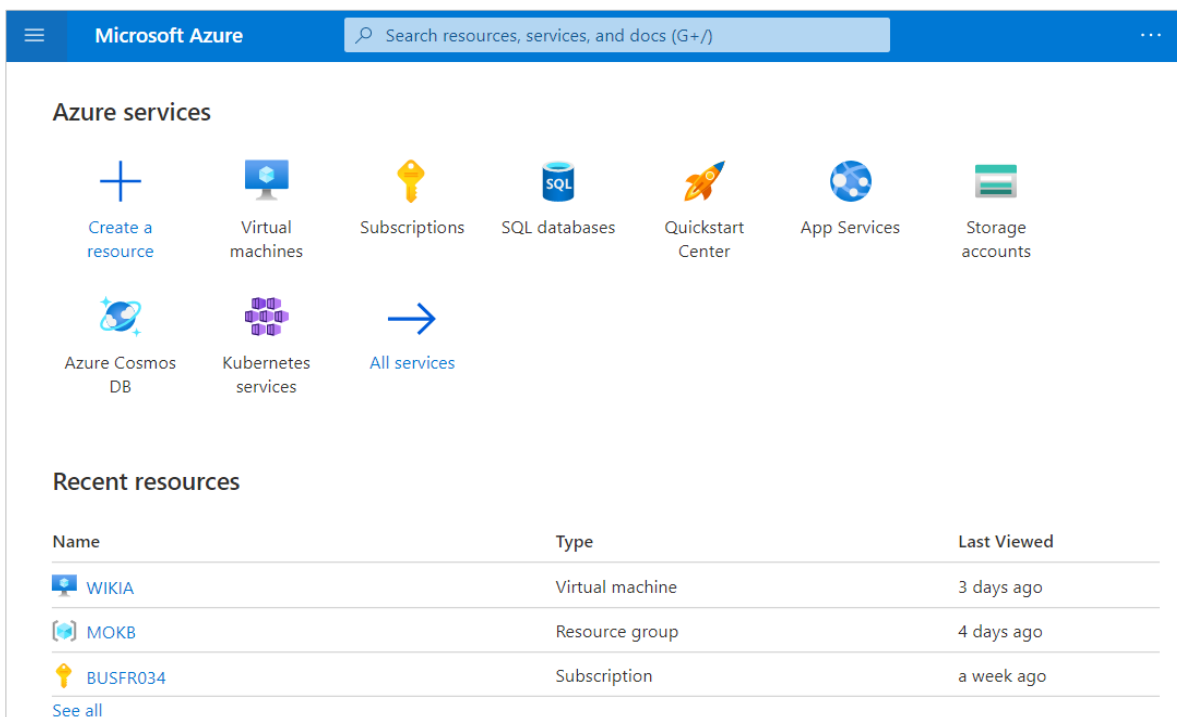


Figure 32 Microsoft Azure subscription for MobiDataLab

Once the cloud infrastructure was set up, we checked that the catalogue solutions were mature enough for being used reliably in such a cloud context. Since OpenDataSoft is only available as a SaaS solution, we rather focused on benchmarking and testing the following solutions:

- CKAN, and its ecosystem – namely Solr and PostgreSQL
- GeoNetwork, and its ecosystem – namely Java, Tomcat, ElasticSearch (an alternative to Solr) and PostGIS (the spatial extension of PostgreSQL)

These ecosystems have an impact on the infrastructure, as their components need to be integrated and even be shared between several services, which can lead to CPU-related requirements. They also need to be taken into account when considering the process of packing, deploying, and running. For example, GeoNetwork, ElasticSearch and PostGIS should not be included in the same package (e.g., the same Docker⁹ file).

Considering the above, the initial deployment of these solutions was done on a sandbox server for testing and validation purposes, benchmarking, and for writing the installation guide. Then the cloud services were evaluated in several configurations (virtual machines, containers, SaaS). The following sections correspond to the catalogue installation guides, first done on the sandbox and validated after deployment in the cloud (not exactly the same as in the sandbox).

3.1. Integrating CKAN in the Transport Cloud

3.1.1. *Installing CKAN packages for Ubuntu 20.04*

- Update ubuntu package index

```
sudo apt update
```

- Install Ubuntu required packages for CKAN

```
sudo apt install -y libpq5 redis-server nginx supervisor
```

- Download CKAN package (for Python 3)

```
wget https://packaging.ckan.org/python-ckan_2.9-py3-focal_amd64.deb
```

- Install additional needed packages

```
sudo apt-get install python3.8-distutils
```

- Install CKAN package

```
sudo dpkg -i python-ckan_2.9-py3-focal_amd64.deb
```

⁹ Docker (<https://www.docker.com/>) is an open-source tool that allows to ship an application with all the necessary functionalities as one package.

3.1.2. Installing PostgreSQL

- Install PostgreSQL

```
sudo apt install -y postgresql
```

- Check that it installed correctly by running the following command and ensuring that the encoding of databases is UTF8

```
sudo -u postgres psql -l
```

- Create a database user and create a password for the new user when prompted, replace username with the username of your choice

```
sudo -u postgres createuser -S -D -R -P username
```

- Create a new PostgreSQL database owned by the new user, replace ckan_default with a database name of your choice and username with the username of the user just created

```
sudo -u postgres createdb -O username ckan_default -E utf-8
```

- Edit the CKAN configuration file and fill in the password, database, and database user for the database you've created

```
vim /etc/ckan/default/ckan.ini
```

In this example our information is as follows

Username: ckanuser

Password: password

Database Name: ckan_default

3.1.3. Installing Solr

- Install Solr

```
sudo apt install -y solr-tomcat
```

- Change the default port Tomcat runs on to the one expected by CKAN

```
vim /etc/tomcat9/server.xml
```

Edit the following line

```
<Connector port="8080" protocol="HTTP/1.1"
```

to

```
<Connector port="8983" protocol="HTTP/1.1"
```

- Replace the default schema.xml file with a symlink to the CKAN schema file.

```
sudo mv /etc/solr/conf/schema.xml /etc/solr/conf/schema.xml.bak
```

```
sudo ln -s /usr/lib/ckan/default/src/ckan/ckan/config/solr/schema.xml /etc/solr/conf/schema.xml
```

- Restart Solr

```
sudo service tomcat9 restart
```

- Check that Solr is running by entering the following into your browser

```
http://localhost:8983/solr/
```

- Edit the solr_url line in the CKAN configuration file to go to your Solr server

```
vim /etc/ckan/default/ckan.ini
```

- Also edit the following options in the CKAN configuration File

```
site_id
```

```
site_url
```

3.1.4. Initialise the CKAN database

- Initialize CKAN database by running the following command

```
sudo ckan db init
```

- Reload the Supervisor daemon so the new processes are picked up

```
sudo supervisorctl reload
```

- Check the status of the processes

```
sudo supervisorctl status
```

The following should appear with no errors

```
ckan-datapusher:ckan-datapusher-00  RUNNING  pid 1963, uptime 0:00:12
```

```
ckan-uwsgi:ckan-uwsgi-00          RUNNING  pid 1964, uptime 0:00:12
```

```
ckan-worker:ckan-worker-00       RUNNING  pid 1965, uptime 0:00:12
```

- Restart Nginx

```
sudo service nginx restart
```

3.2. Integrating OpenDataSoft in the Transport Cloud

As mentioned before (2.2), OpenDataSoft is not an open-source catalogue. However, for MobiDataLab, we are using KISIO private instance and a demo that can be requested in the official website. And since it is a SaaS, it is almost ready to use after the initial configurations.

3.3. Integrating GeoNetwork in the Transport Cloud

3.3.1. Installing GeoNetwork

GeoNetwork version 4.0.5 is considered for the reference data catalogue demonstrator. GeoNetwork 4.0.5 is Java web application, assembled in a WAR artifact and requires at least Java 8 and the following minimal dependencies:

1. Java web server: Jetty 9 (alternative Tomcat 8.5)
2. Database: PostgreSQL 13.4 (alternative: H2 database)
3. ElasticSearch 7.6.2

3.3.2. Configuring GeoNetwork

3.3.2.1. Pre-deployment configuration

Many of GeoNetwork configuration pre-deployment are done via Java system properties or environment variables.

The configurations related to PostgreSQL and ElasticSearch are listed in Table 4. Java system properties shall be given in corresponding config file, or via Java CLI options in web server initialization, namely:

- For Jetty:
 - `export JAVA_OPTS="-Dgeonetwork.db.type=postgres"`
 - `java -Dgeonetwork.db.type=postgres $JETTY_HOME/start.jar`
- For Tomcat:
 - `export CATALINA_OPTS="-Dgeonetwork.db.type=postgres"`

Table 4 GeoNetwork configurations for PostgreSQL and Elasticsearch

Configuration description	Value	Java system property (or Environment variable)	Config file in WEB-INF folder
Database type	postgres	geonetwork.db.type GEONETWORK_DB_TYPE	config.properties
Database host	localhost	jdbc.host GEONETWORK_DB_HOST	config-db/jdbc.properties
Database port	5432	jdbc.port GEONETWORK_DB_PORT	config-db/jdbc.properties
Database name		jdbc.name GEONETWORK_DB_NAME	config-db/jdbc.properties
Database username		jdbc.username GEONETWORK_DB_USERNAME	config-db/jdbc.properties
Database password		jdbc.password GEONETWORK_DB_PASSWORD	config-db/jdbc.properties
Database: The maximum number of active connections that can be allocated from this pool at the same time	33	jdbc.basic.maxActive	config-db/jdbc.properties
Database: The maximum number of connections that should be kept in the pool at all times	33	jdbc.basic.maxIdle	config-db/jdbc.properties
Database: The initial number of connections that are created when the pool is started	33	jdbc.basic.initialSize	config-db/jdbc.properties

Database: The maximum number of milliseconds that the pool will wait (when there are no available connections) for a connection to be returned before throwing an exception				
	200	jdbc.maxWait		config-db/jdbc.properties
ElasticSearch complete URL		es.url		config.properties
ElasticSearch protocol	http	es.protocol		config.properties
ElasticSearch host	localhost	es.host		config.properties
ElasticSearch port	9200	es.port		config.properties
ElasticSearch username		es.username		config.properties
ElasticSearch password		es.password		config.properties
ElasticSearch (records)	index	gn-records	es.index.records	config.properties

3.3.2.1. Post-deployment configuration

After GeoNetwork is configured and deployed, users are able to access its homepage under `/srv/eng/catalog.search#/home`. For the first deployment, the user needs to log in as admin, to enable some GeoNetwork features as well as to customize the UI. The changes made in admin settings are made persistent (i.e., recorded) in the database, so configuring is only required for uninitialized database. These changes can be done either manually in the web interface, in the database directly or in the custom initialization SQL script, namely `WEB-INF/classes/setup/sql/ data/custom-data-db-default.sql`.

Table 5 Basic settings in GeoNetwork

Usage	Setting name in SQL	Setting in Web interface	name
To change GeoNetwork catalogue name	system/site/name	Catalog description > Catalog Name	
To change organization name	system/site/organization	Catalog description > Organization	

To change time zone	system/server/timeZone	Catalog server > Timezone
To correct host and metadata reference	system/server/host	Catalog server > host
	system/server/protocol	Catalog server > Preferred Protocol
	system/server/port	Catalog server > Port
To enable CSW	system/csw/enable	Catalog Service for the Web (CSW) > CSW enabled
To enable email feedback (optional)	system/feedback/email	Feedback > Email
	system/feedback/mailServer/host	Feedback > SMTP host
	system/feedback/mailServer/port	Feedback > SMTP port
	system/feedback/mailServer/username	Feedback > User name
	system/feedback/mailServer/password	Feedback > Password

3.3.2.1. Limitations

When the user tries to import sizeable amount of metadata XML files, error responses are returned. This occurs when GeoNetwork tries to request more database connections than the limit by configuration or the limit of the database instance. JDBC connection pool configuration and database connection limit need to be configured accordingly to actual usage.

When the user tries to import a MEF/ZIP file that contains sizeable amount of metadata XML files, gateway error is returned, either “504 Gateway Timeout Error” or “502 Bad Gateway Error”. The error is caused either by the timeout configuration of the web server or the load balancer in the deployed network. Although timeout can be increased to actual needs, it may cause unexpected errors in other services that share the web server or load balancer.

When the user tries to export metadata as ZIP file, error “No space left on device” is returned. This could be caused by excessive logs or internal indices created by GeoNetwork during runtime. As no detailed investigation was done, increasing the server physical storage is realized as simple workaround.

4. Conclusions

This report documents the experience of demonstrating the reference data catalogue of the MobiDataLab project. We have shown how, starting from the mobility data provided by the so-called Reference Group of transport stakeholders (local authorities and international organisations) we could facilitate the discovery of these data for potential consumers and reusers.

Data discoverability is improved thanks to different cataloguing solutions, generalist or thematic, that we evaluated in a cloud context. This demonstrator aimed at making clear the differences between these cataloguing solutions and evaluating several of them in the context of a Transport Cloud (i.e., federation of cloud services), for which portability is a strong requirement. Three solutions have been explored – CKAN, OpenDataSoft and GeoNetwork. All three solutions could be offered to Labs participants to help them discover mobility data, as it is not possible to know in advance their individual tool preferences.

This demonstrator should be seen as part of a set of demonstrators corresponding to the other tasks of WP4, namely in relation to data access services, data processors, and data anonymisation. These 4 demonstrators are the software counterpart of the FAIR principles, the data catalogue corresponding to the first letter of this acronym, the F of findability. Findability is a prerequisite to facilitate data access (A), data interoperability (I) and data reusability (R), and that is why this catalogue will be useful for the other tasks.

However, for this link between catalogues, services, processors and anonymisation tools to be tighter and therefore more effective, a more advanced integration is needed, and this is what the MobiDataLab WP4 partners will work on for the next version 2.

| MobiDataLab consortium

The consortium of MobiDataLab consists of 10 partners with multidisciplinary and complementary competencies. This includes leading universities, networks and industry sector specialists.



[@MobiDataLab](https://twitter.com/MobiDataLab)
[#MobiDataLab](https://twitter.com/MobiDataLab)



<https://www.linkedin.com/company/mobidatalab>

For further information please visit www.mobidatalab.eu



MobiDataLab is co-funded by the EU under the H2020 Research and Innovation Programme (grant agreement No 101006879).

The content of this document reflects solely the views of its authors. The European Commission is not liable for any use that may be made of the information contained therein. The MobiDataLab consortium members shall have no liability for damages of any kind that may result from the use of these materials.