DIGICIRC

WP3 – ENABLE: Creating LSD Enabler Tools

D3.5: Circular economy data hub

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 873468.

Document Information

Grant Agreement Number	873468	Acronym		DigiCirc
Full Title	European cluster-led accelerator for digitization of the circular economy across key emerging sectors			
Start Date	1st May 2020	Duration		32 months
Project URL	https://digicirc.eu/			
Deliverable	D 3.5 - Circular economy data hub			
Work Package	WP 3 - ENABLE: Creating LSD Enabler Tools			
Date of Delivery	Contractual	31st January 2021	Actual	31st January 2021
Nature	Other	Dissemination Level		Public
Lead Beneficiary	DRAXIS			
Responsible Author	Eleni Ntzioni, Panagiota Syropoulou, George Letsos			
Contributions from	Melanie Pellen, Igor Milosavljevic			

Document History

Version	Issue Date	Stage	Description	Contributor
0.1	04/01/2021	Draft	TOC	Eleni Ntzioni, Panagiota Syropoulou
0.2	08/01/2021	Draft	First version of document	Eleni Ntzioni, George Letsos
0.5	19/01/2021	Draft	Second version of document	Eleni Ntzioni, George Letsos
0.8	22/01/2021	Draft	Third version of document	Eleni Ntzioni, Panagiota Syropoulou
1.0	27/01/2021	Pre-Final	Document ready for internal review	Eleni Ntzioni, Panagiota Syropoulou
2.0	31/01/2021	Final	Document ready for submission	Eleni Ntzioni, Panagiota Syropoulou, Melanie Pellen, Igor Milosavljevic

Disclaimer

Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright message

Table of Contents

List of Tables

Table of Figures

List of acronyms

Acronym	Designation
API	Application Programming Interface
CE	Circular Economy
CPU	Central Processing Unit
DB	Database
GIS	Geographic Information System
GUI	Graphical User Interface
HDD	Hard Disk Drive
I/O	Input/Output
RAM	Random Access Memory
SME	Small and Medium Enterprise
SQL	Structured Query Language
US	User Story
WMS	Web Map Service

1 Executive summary

The DigiCirc project aims to boost the circular economy using digital tools, by supporting innovative SMEs in the development and marketing of solutions based on circular value chains through three acceleration programs on the following themes: "Circular City", "Blue economy" and "Bioeconomy". DigiCirc will offer to the SMEs four DigiCirc Digital Tools to support them in the development, demonstration and commercialization of their solutions. The focus of this report is one of the data-use facilitation tools of DigiCirc, namely the Circular economy data hub (CE data hub)[1], a multisource repository hosting geo-reference data and making them accessible to users.

This document outlines the design and development process for the Circular economy data hub, which considers the User needs and ensures that the tool will be tailored to the three domains of the project and will be of value to end users. The approach for identifying and collecting the data populating the CE data hub is also presented. The main features of the CE data hub are described and displayed in detail and the updated system architecture to support the services required for the tool is analyzed.

[1] The first name of the tool - Geolocation of materials - has been changed to Circular economy data hub to reflect more accurately its aim.

2 Introduction

The DigiCirc project aims to develop digital platforms to streamline data analysis and processing, and provide relevant datasets of real-world test beds for each accelerator. In order to achieve that, the Circular economy data hub digital tool provides an interactive platform where several types of heterogeneous open datasets received from a wide range of sources are available. The platform provides the available information in a combined environment, in various formats that may be useful to the users. Furthermore, the users have access to an easy-to-use Interface and they are able to visualize the information in multiple ways before they select the ones more suited to their needs.

The objective is to assist users with the development of products and services, which will exploit the datasets either by exporting them from the user interface or by using a RESTful API. Currently, the CE data hub provides information originating from many different sources, expanding under several different sectors. The acquired datasets are made publicly available for access and download. The retrieved information can be used for various purposes, such as:

i. the production of value-added solutions,
ii. training neural networks,
iii. calibration or visualization purposes,
iv. discovering insights based on big data analysis that can combine observations, time and geolocation information, etc.,
v. services of operational monitoring, warning and reporting to public institutions or businesses.

In section 3 of the document, the Circular economy data hub is described in detail. First of all, the aim of the tool and the target users are presented and following that and then the focus shifts to the data that are available in the CE data hub currently. After that, the main features and functionalities of the tool are presented in detail, accompanied by screenshots of the CE data hub. Section 4 contains the outline of the updated system architecture that supports the CE data hub with a description of its subsystems. The hardware and software requirements are presented in section 5 and the document concludes with section 6, where the hub expandability and the next steps are presented.

3 Circular economy data hub

The circular economy model is based on the reuse of resources (e.g., products, materials), the regeneration of natural systems and the reduction of waste and pollution. Open data can provide solutions to achieve the goals of circular economy by improving decision-making based on data insights[2]. Despite the availability of information in the form of open data, one of the major problems is that the information is decentralized. Practically every new project or initiative creates a small information hub to serve its needs. This clustering, leads to data fragmentation, and mostly inaccessible and unattainable information sources.

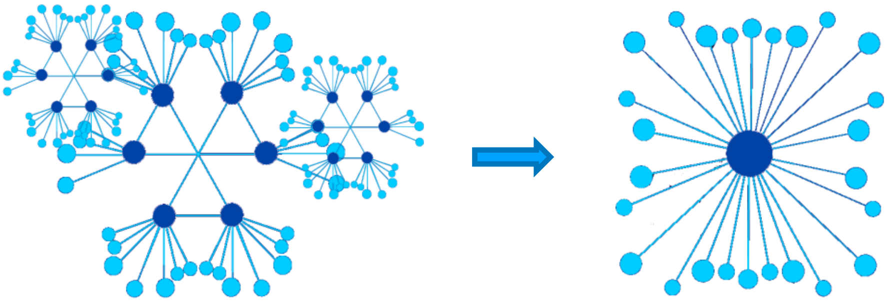


Figure 1: From clustered data hubs to centralized information structures

The challenge of the CE data hub is to collect and centralize the available data, and make this information readily available to the user through a convenient and friendly user environment. In order to achieve this, the first step is to unify the available data sources such that to create a single access data point, so the data will cease being decentralized into small data hubs and will fall under the big CE data hub umbrella. The main objective of the hub is to offer SMEs, a centralized, expedite access to valuable resources that is often key to realizing value-added solutions enabling circular economy.

In particular, the CE data hub is an automated data acquisition and pre-processing system, acting as a multi-source repository of information and measurements derived from instruments located directly at the point of interest. The hub is offering automated discovery, retrieval, harmonization and transformation services, collecting geospatial data from a wide range of sources: acquired from sensor networks, open data platforms, government repositories, citizen observatories and many others. The acquired information is made available for use in the data hub and where necessary, ingested into the hub's own database to be accessible to the users in a combined, uniform, analytical environment.

The CE data hub tends to the needs of non-technical managers but it is mainly oriented for programmers. More specifically, the hub is offering to the managers a user-friendly environment, with intuitive search and filtering options, providing small descriptions of each information source available and details such as latest dataset update and also the ability of dataset-preview and visual representations of datasets. At the same time, the hub exposes fully documented API services for data provision, allowing any programmer possessing with only basic programming skills, to access the data in the hub. Through the environment the programmer can acquire all the necessary documentation on the methods to access all the available datasets, including examples of dataset requests, filters and several other options to facilitate the programmer.

Currently, the hub integrates 59 datasets, from 10 different data sources and will be extended during the course of the project to include many more.

The Circular economy data hub can be accessed at: https://datahub.digicirc.eu/

[2] https://www.europeandataportal.eu/

3.1 Data included in the hub

In order to populate the CE data hub, the first step is to identify data sources that provide insights to support circular economy regarding, for example, food production, use of resources or reducing pollution. For the first version of the CE data hub, the focus of data sources identification shifted to the domain of **Circular cities**, that is the domain of the first open call of DigiCirc. The initial data sources research and identification was based on the sectors that were defined in the challenges of the first accelerator, namely Buildings and construction, Plastics, Food, Energy and Water. Furthermore, a template was created to identify and qualify the most appropriate datasets provided in the data sources, that are relevant to the domain of Circular cities. For the data sources identification and the datasets qualification through the template, guidance and assistance was provided by the Cluster partner with the expertise in the domain of Circular cities. This process will be described in detail in *D3.2 Circular cities testbed dataset*. The resulting datasets are included in the CE data hub and they fall under the following sectors: **Air quality, Buildings and construction, Energy, Mobility, Soft mobility, Waste management** and **Water.**

Currently, the hub integrates datasets from the following data sources[3]:

- Open Data of the City of Paris
- Open Data platform of the Île-de-France region
- Open Data Network Energies (ODRÉ)
- Paris Saclay Open Data portal
- Open platform for French public data
- Open data portal of the Toulouse metropolis
- Open data Hauts-de-seine
- Open data portal of the Republic of Bulgaria
- Ireland’s open data portal
- Poland's Open Data Portal

The hub delivers each dataset in one or multiple formats and it is accompanied by metadata, additional information that describe the dataset. In particular, the following information is collected and displayed for each dataset:

- Title of dataset
- Description
- Keywords
- Domain
- Sector
- Organization
- Language
- Access level
- Geographical area
- Date of creation
- Date of latest update
- License
- Link to original source
- Revisions of dataset

[3] With the support of Sofia Knowledge City and Baltic Eco-Energy Cluster, we were able to identity open data in Europe.

3.2 Main features of the Circular economy data hub

3.2.1 Structure

Dataset: The data publishing unit in the CE data hub is called “dataset”. A dataset is a parcel of data - for example, it could be the locations of car sharing stations in a specific city. A dataset consists of “metadata” and a number of “resources”, which hold the data itself. Data formats can include CSV or Excel spreadsheets, XML file, PDF documents, image files, linked data in RDF format etc. A dataset can contain any number of resources.

Metadata: “Metadata” are loosely defined as “data about data”. Metadata document and describe all aspects of a specific dataset (i.e. the who, why, what, when and where) that allow understanding of the physical format, content and context of the data. The metadata accompanying each dataset are described in Chapter 3.1.

Domain: The datasets are assorted in one of the three DigiCirc domains, i.e., Circular cities, Blue economy, Bioeconomy. For the first version of the data hub, the datasets are in the Circular cities domain.

Sectors: Each domain consists of sectors for further categorizing the datasets in the data hub. Each dataset may belong to one or more sectors. Currently seven sectors are available and they represent a simple way to help users search and access data thematically.

Administrator: The data hub has an administrator who is responsible for managing the content, managing the users and allocating editing authorization rights to the users.

User: The CE data hub can be publicly accessed but if a user wants to contribute by uploading data they have to register. The administrator will authorize the user to upload data related to a specific domain, by assigning the role Editor to them.

The following diagram depicts the structure of the Circular economy data hub schematically.

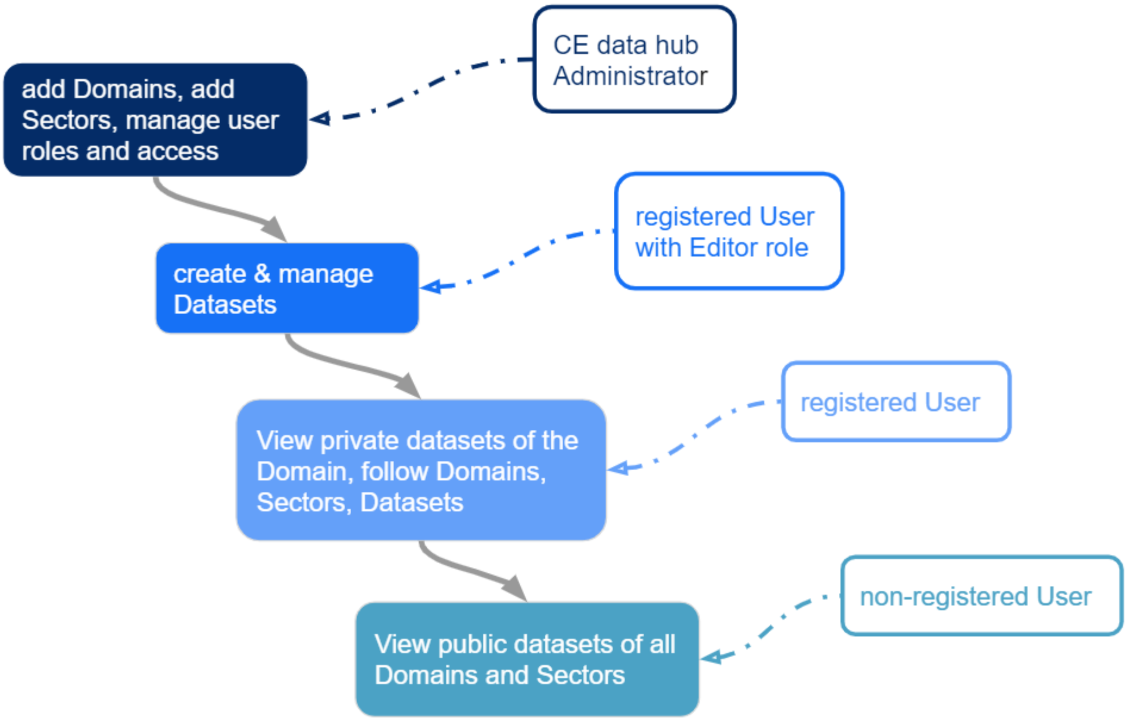
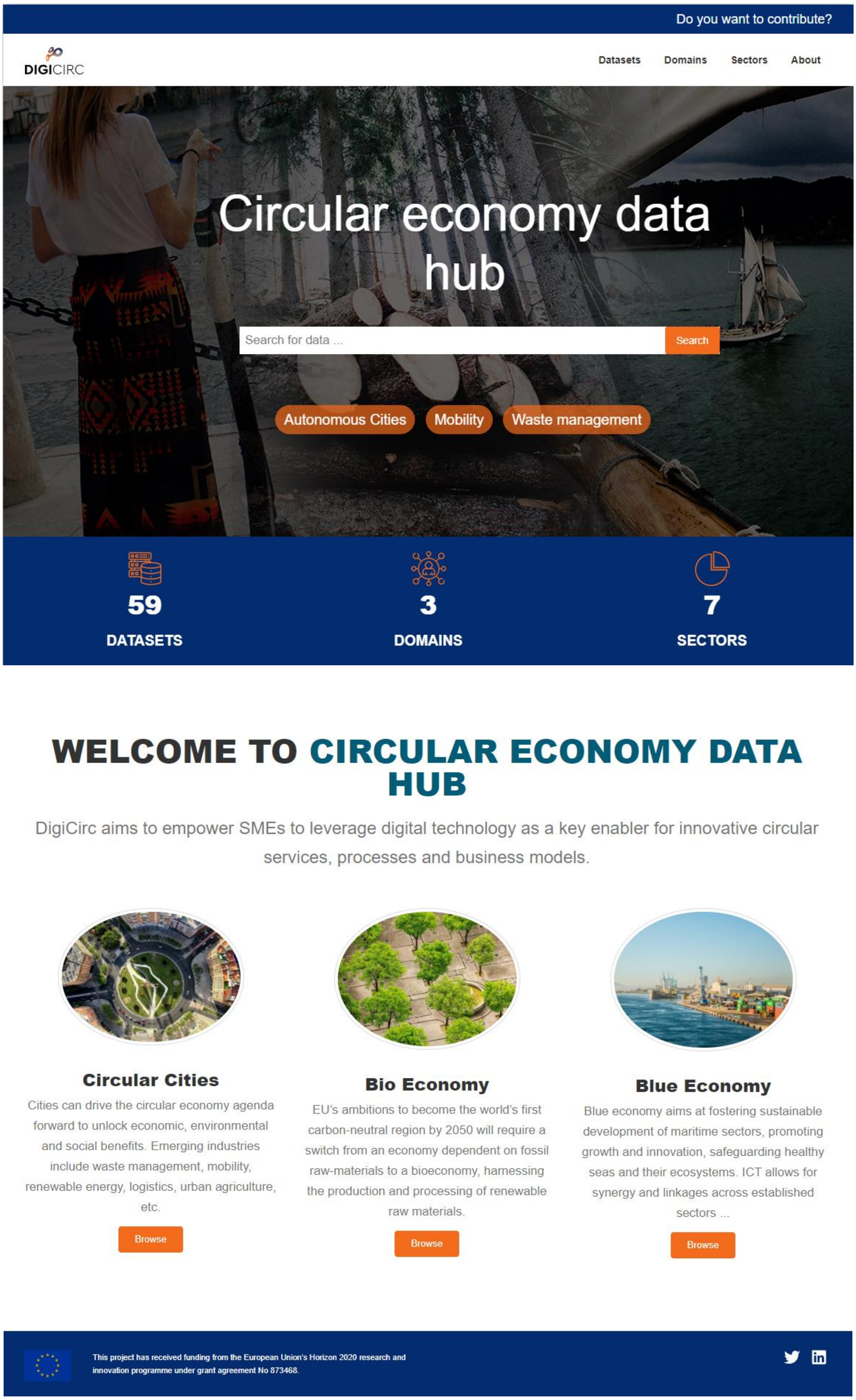

CE data hub Administrator
add Domains, add Sectors, manage user roles and access
registered User with Editor role
create & manage Datasets
registered User
View private datasets of the Domain, follow Domains, Sectors, Datasets
non-registered User
View public datasets of all Domains and Sectors

Figure 2: Structure of the Circular economy data hub

3.2.2 Welcome page



Do you want to contribute?

DIGICIRC

Datasets Domains Sectors About

Circular economy data hub

Search for data ...

Search

Autonomous Cities Mobility Waste management

59 DATASETS

3 DOMAINS

7 SECTORS

WELCOME TO CIRCULAR ECONOMY DATA HUB

DigiCirc aims to empower SMEs to leverage digital technology as a key enabler for innovative circular services, processes and business models.

Circular Cities

Cities can drive the circular economy agenda forward to unlock economic, environmental and social benefits. Emerging industries include waste management, mobility, renewable energy, logistics, urban agriculture, etc.

Browse

Bio Economy

EU's ambitions to become the world's first carbon-neutral region by 2050 will require a switch from an economy dependent on fossil raw-materials to a bioeconomy, harnessing the production and processing of renewable raw materials.

Browse

Blue Economy

Blue economy aims at fostering sustainable development of maritime sectors, promoting growth and innovation, safeguarding healthy seas and their ecosystems. ICT allows for synergy and linkages across established sectors ...

Browse

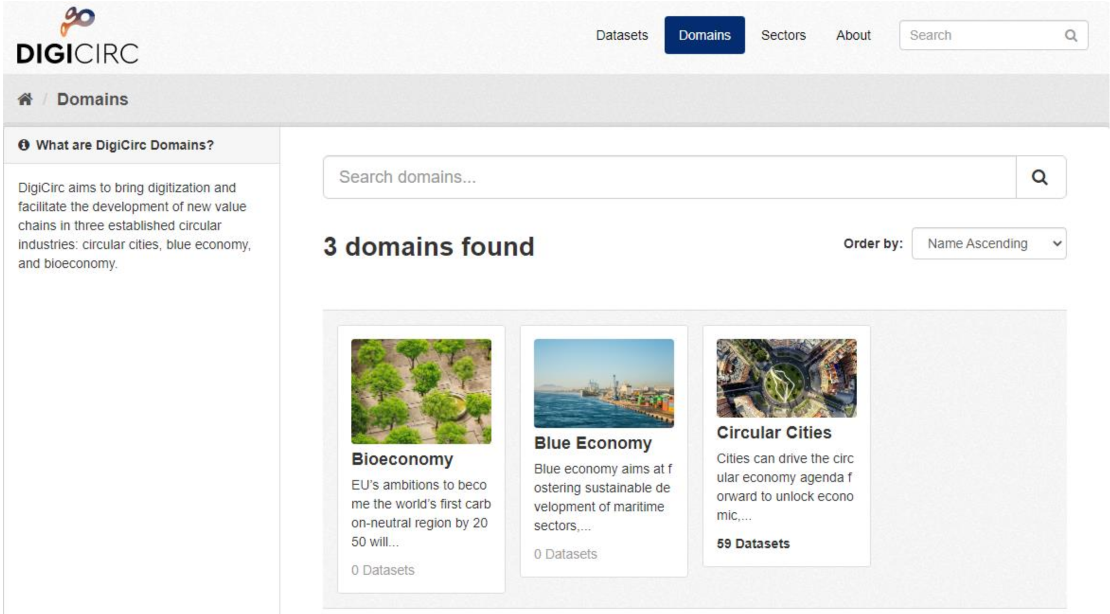
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 873468.

Figure 3: Welcome page of the Circular economy data hub

The welcome page of the CE data hub includes a menu for allowing the user to easily navigate to the Datasets page, the Domains page, the Sectors page and the About page. The main feature of the welcome page is the Search functionality, that directly produces a list of datasets based on the user's search input. The most common tags are available under the search bar, directing the user to the datasets related to those keywords. Statistics about the data hub are included in the next part of the page. Furthermore, descriptions about the three DigiCirc domains and links to respective domain pages are included.

3.2.3 Domain

The user can navigate to the Domains page of the data hub by selecting "Domains" from the menu (Figure 4). Each domain has a dedicated page, where users can find information about the domain (tab "About"), search within its datasets (tab "Datasets") and look at the latest activities relating to the datasets of that domain (tab "Activity Stream") (Figure 5).



What are DigiCirc Domains?

DigiCirc aims to bring digitization and facilitate the development of new value chains in three established circular industries: circular cities, blue economy, and bioeconomy.

3 domains found

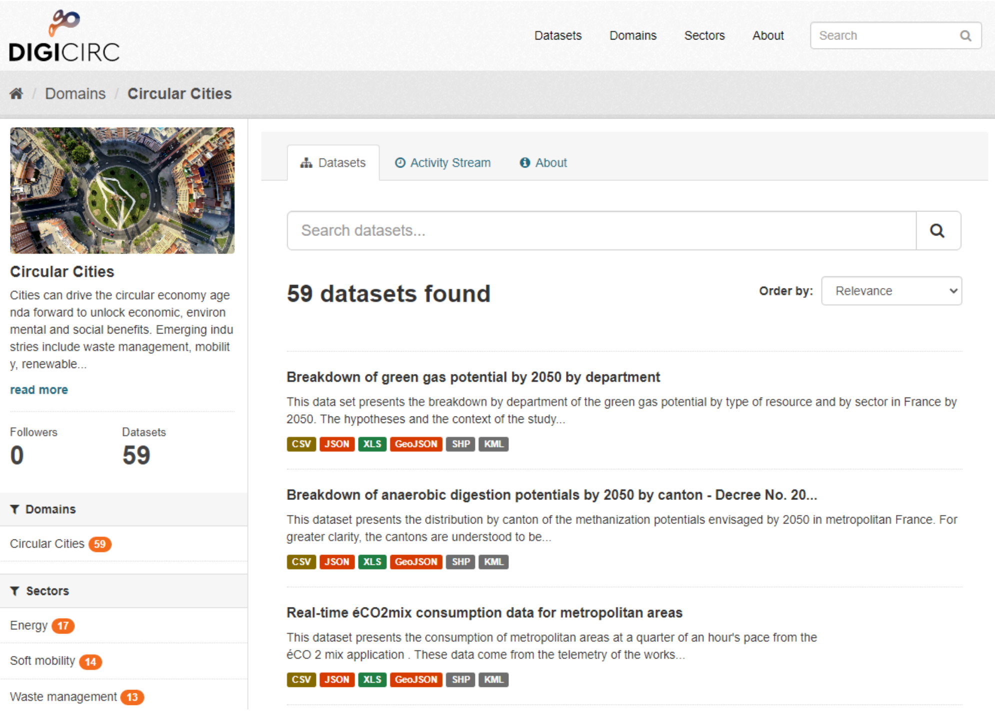
Order by: Name Ascending

Bioeconomy
EU's ambitions to become the world's first carbon-neutral region by 2050 will...
0 Datasets

Blue Economy
Blue economy aims at fostering sustainable development of maritime sectors,...
0 Datasets

Circular Cities
Cities can drive the circular economy agenda forward to unlock economic,...
59 Datasets

Figure 4: Domains page



DIGICIRC

Datasets Domains Sectors About Search

Domains / Circular Cities

Circular Cities

Cities can drive the circular economy agenda forward to unlock economic, environmental and social benefits. Emerging industries include waste management, mobility, renewable...

read more

Followers 0 | Datasets 59

Domains

Circular Cities 59

Sectors

Energy 17

Soft mobility 14

Waste management 13

Datasets | Activity Stream | About

Search datasets...

59 datasets found

Order by: Relevance

Breakdown of green gas potential by 2050 by department

This data set presents the breakdown by department of the green gas potential by type of resource and by sector in France by 2050. The hypotheses and the context of the study...

CSV JSON XLS GeoJSON SHP KML

Breakdown of anaerobic digestion potentials by 2050 by canton - Decree No. 20...

This dataset presents the distribution by canton of the methanization potentials envisaged by 2050 in metropolitan France. For greater clarity, the cantons are understood to be...

CSV JSON XLS GeoJSON SHP KML

Real-time éCO2mix consumption data for metropolitan areas

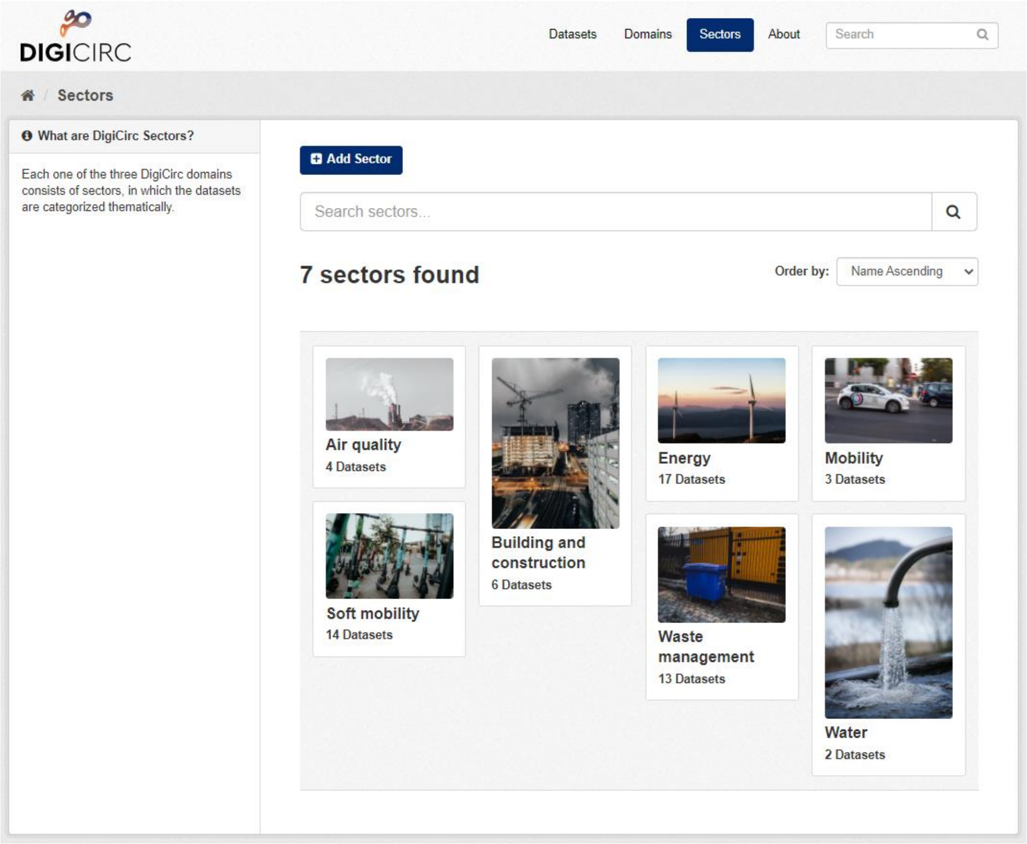
This dataset presents the consumption of metropolitan areas at a quarter of an hour's pace from the éCO 2 mix application . These data come from the telemetry of the works...

CSV JSON XLS GeoJSON SHP KML

Figure 5: Page of Circular cities domain

3.2.4 Sector

In the CE data hub, the user will find different sectors related to the certain thematic aspects of the three DigiCirc domains, by selecting “Sectors” in the menu (Figure 6). Each sector has a dedicated page, where users can search within its datasets (tab “Datasets”) and look at the latest activities relating to the datasets of the sector (tab “Activity Stream”) (Figure 7).



DIGICIRC

Datasets Domains Sectors About Search

Sectors

What are DigiCirc Sectors?

Each one of the three DigiCirc domains consists of sectors, in which the datasets are categorized thematically.

Add Sector

Search sectors...

7 sectors found

Order by: Name Ascending

Air quality – 4 Datasets

Building and construction – 6 Datasets

Energy – 17 Datasets

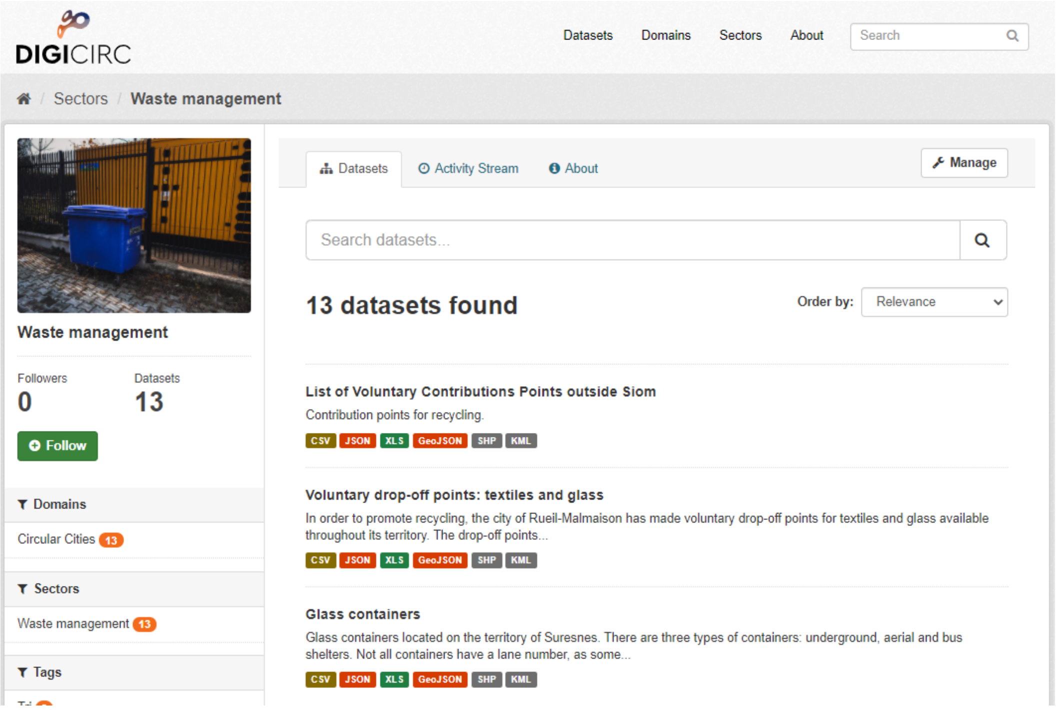
Mobility – 3 Datasets

Soft mobility – 14 Datasets

Waste management – 13 Datasets

Water – 2 Datasets

Figure 6: Sectors page



13 datasets found

List of Voluntary Contributions Points outside Siom

Contribution points for recycling.

CSV JSON XLS GeoJSON SHP KML

Voluntary drop-off points: textiles and glass

In order to promote recycling, the city of Rueil-Malmaison has made voluntary drop-off points for textiles and glass available throughout its territory. The drop-off points...

CSV JSON XLS GeoJSON SHP KML

Glass containers

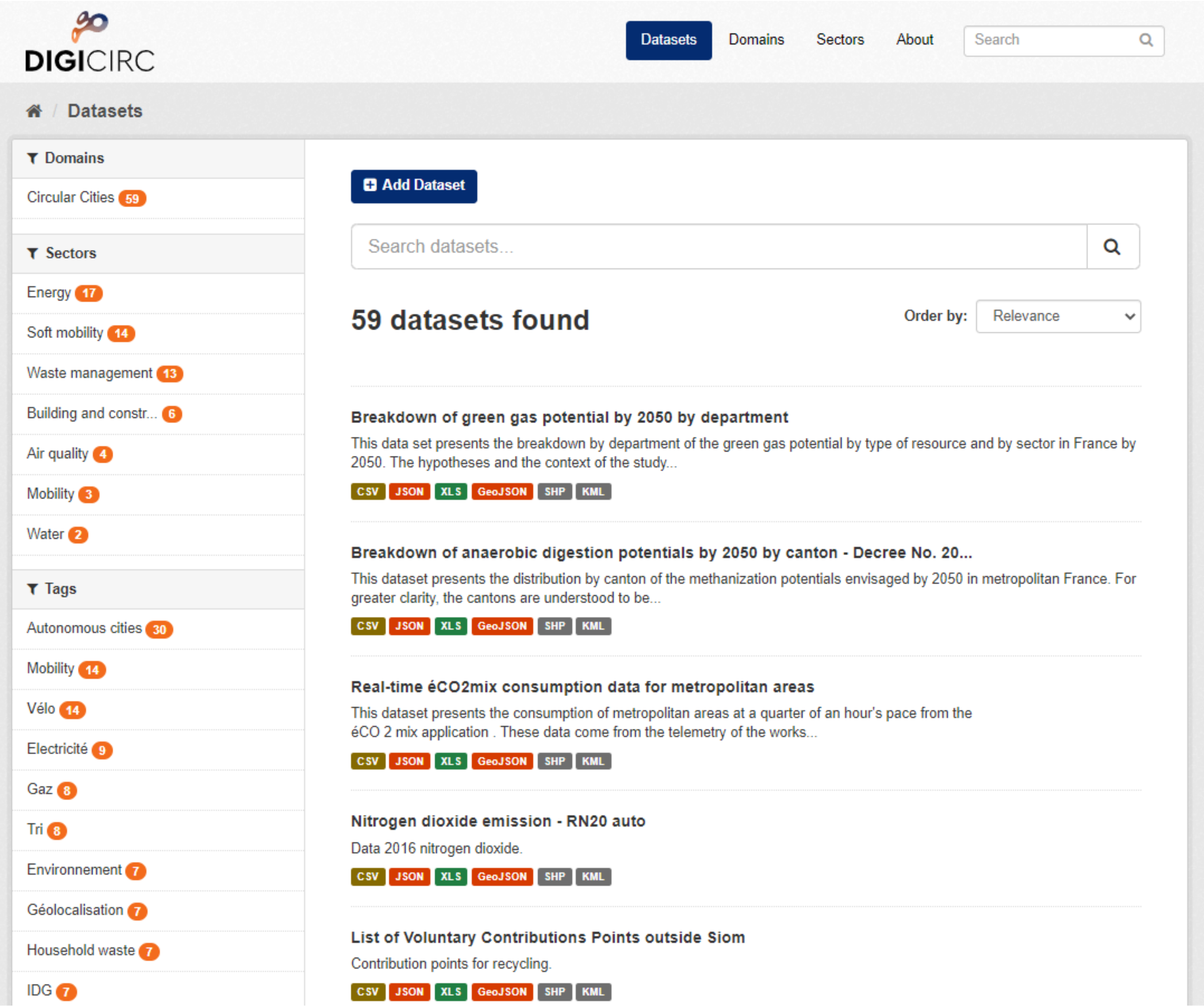
Glass containers located on the territory of Suresnes. There are three types of containers: underground, aerial and bus shelters. Not all containers have a lane number, as some...

CSV JSON XLS GeoJSON SHP KML

Figure 7: Page of Waste management sector

3.2.5 Search for datasets

To find datasets in the CE data hub, the user can type any combination of search words (e.g., “water”, “recycle”, etc.) in the search box on any page. The CE data hub will then return all corresponding search results as a list (Figure 8).



DIGICIRC

Datasets | Domains | Sectors | About | Search

Datasets

Domains

Circular Cities 59

Sectors

Energy 17

Soft mobility 14

Waste management 13

Building and constr... 6

Air quality 4

Mobility 3

Water 2

Tags

Autonomous cities 30

Mobility 14

Vélo 14

Electricité 9

Gaz 8

Tri 8

Environnement 7

Géolocalisation 7

Household waste 7

IDG 7

Add Dataset

Search datasets...

59 datasets found

Order by: Relevance

Breakdown of green gas potential by 2050 by department

This data set presents the breakdown by department of the green gas potential by type of resource and by sector in France by 2050. The hypotheses and the context of the study...

CSV JSON XLS GeoJSON SHP KML

Breakdown of anaerobic digestion potentials by 2050 by canton - Decree No. 20...

This dataset presents the distribution by canton of the methanization potentials envisaged by 2050 in metropolitan France. For greater clarity, the cantons are understood to be...

CSV JSON XLS GeoJSON SHP KML

Real-time éCO2mix consumption data for metropolitan areas

This dataset presents the consumption of metropolitan areas at a quarter of an hour's pace from the éCO 2 mix application . These data come from the telemetry of the works...

CSV JSON XLS GeoJSON SHP KML

Nitrogen dioxide emission - RN20 auto

Data 2016 nitrogen dioxide.

CSV JSON XLS GeoJSON SHP KML

List of Voluntary Contributions Points outside Siom

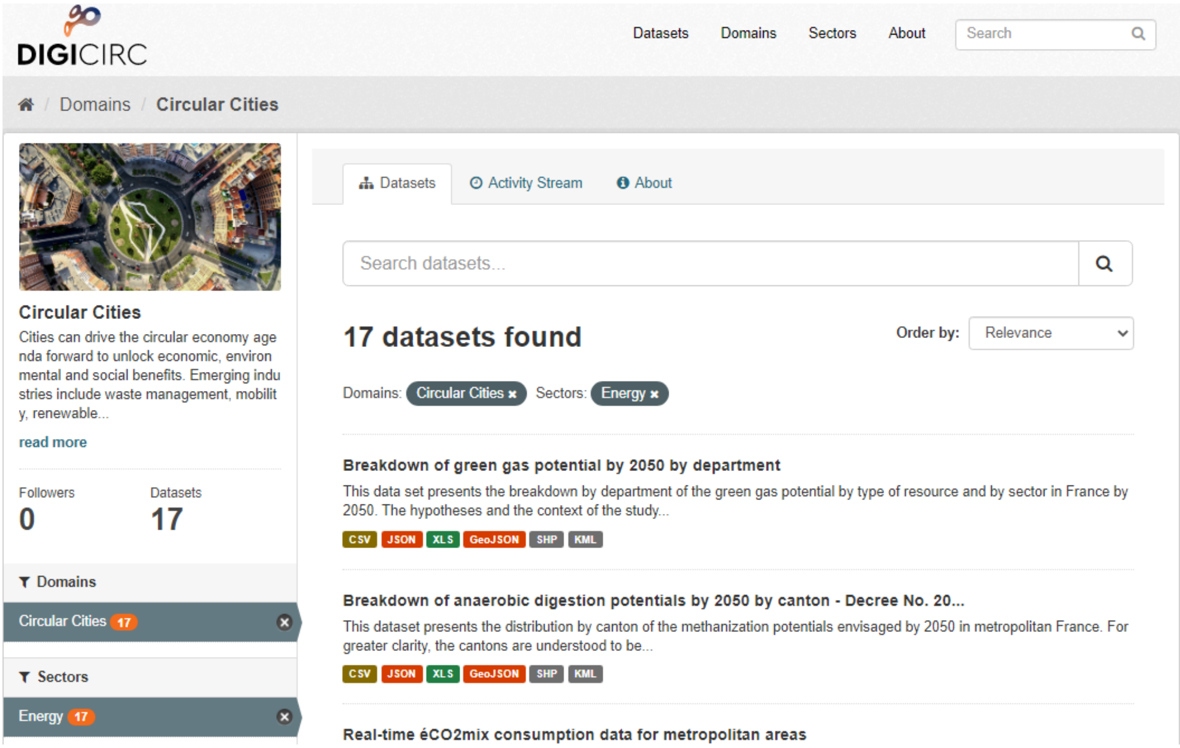
Contribution points for recycling.

CSV JSON XLS GeoJSON SHP KML

Figure 8: Datasets page with the search field and the search results

On the search result page, the user can sort the results according to relevance, name, modification date or popularity by selecting “Order by”. They can also limit the results using the filters on the left column (Domains, Sectors, Tags, Formats and Licenses). The user can combine filters, selectively adding and removing them, and modify and repeat the search with existing filters still in place.

Additionally, the user can select “Domains” from the menu to view the three DigiCirc domains and then select the domain they are interested in and they will be directed to the domain’s page. By typing a search query in the main search box on the page the data hub will return search results as described above but restricted to datasets from the specific domain. Apart from typing in the search box, the user can explore the datasets in that domain. Respectively, they can select “Sectors” from the menu and follow the same process to explore the datasets thematically.



Circular Cities

Cities can drive the circular economy agenda forward to unlock economic, environmental and social benefits. Emerging industries include waste management, mobility, renewable...

read more

Followers 0

Datasets 17

Circular Cities 17

Energy 17

17 datasets found

Domains: Circular Cities Sectors: Energy

Breakdown of green gas potential by 2050 by department

This data set presents the breakdown by department of the green gas potential by type of resource and by sector in France by 2050. The hypotheses and the context of the study...

Breakdown of anaerobic digestion potentials by 2050 by canton - Decree No. 20...

This dataset presents the distribution by canton of the methanization potentials envisaged by 2050 in metropolitan France. For greater clarity, the cantons are understood to be...

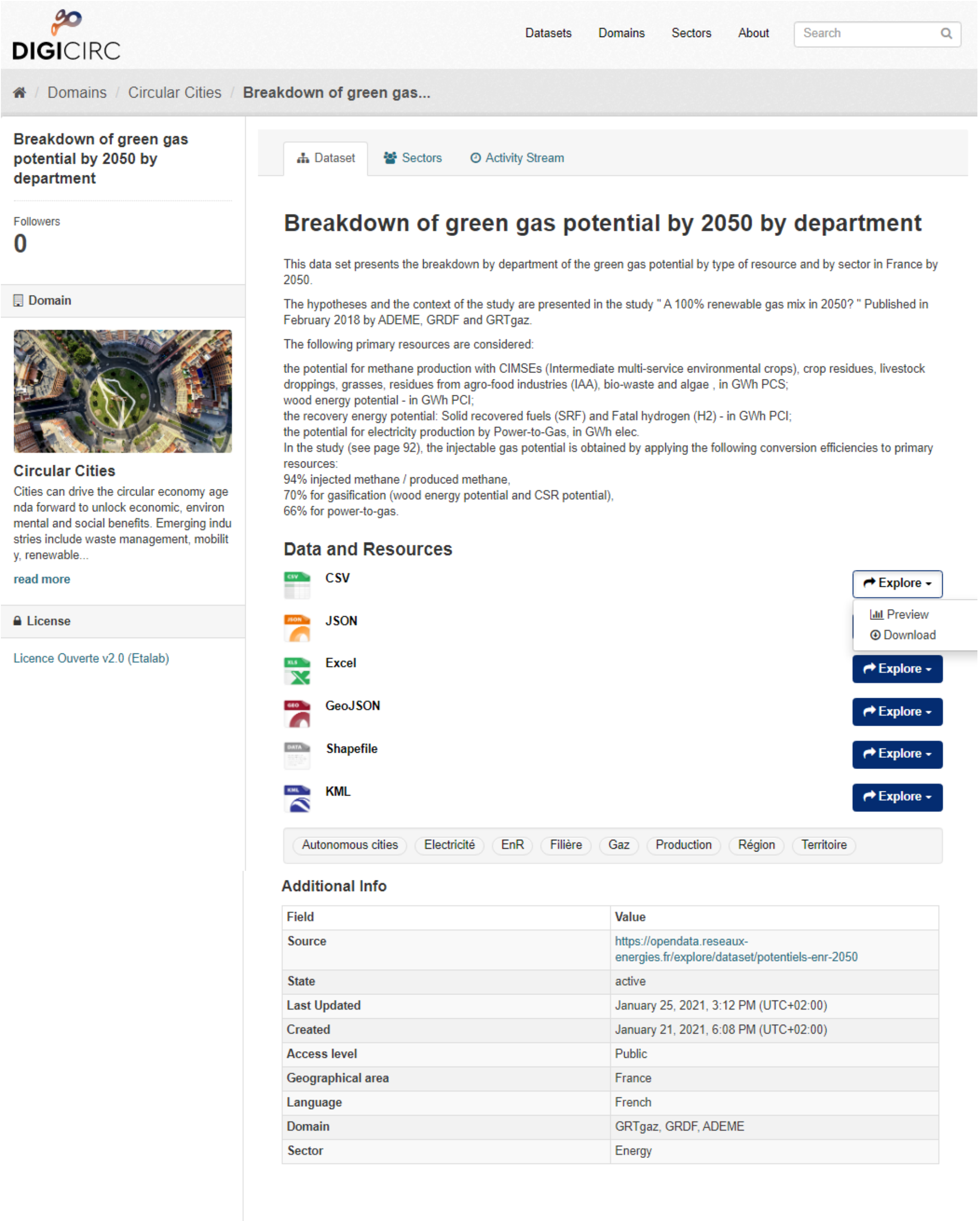
Real-time éCO2mix consumption data for metropolitan areas

Figure 9: Datasets in the Domain's page with applied filters

3.2.6 Dataset

Once the user finds a dataset they are interested in and selected it, the CE data hub will display the dataset page (Figure 10). On the overview page of a dataset, the user will find three tabs: "Dataset", which shows the data and resources belonging to this dataset as well as additional info (metadata), "Sectors", which shows the sectors this dataset belongs to and "Activity stream", which shows the history of recent changes to the dataset. On the left part is a static column that displays the title of the dataset, the domain that it relates to and the license of the dataset.

On the "Dataset" tab the user can see all the information of the dataset including the title, the description, the list of data and resources, the keywords associated to the dataset and the additional info. The "Explore" button on the right of each resource offers the following options to the user: "Preview", which shows the page of the resource including additional information, a preview and the API and "Download", which downloads the file directly. The list of keywords is collected both from the template prepared for the dataset and also from its original data source. The additional information presents the metadata of the dataset, collected both from the template prepared for the dataset and also from its original data source.



Domains / Circular Cities / Breakdown of green gas...

Breakdown of green gas potential by 2050 by department

Followers
0

Domain

Circular Cities

Cities can drive the circular economy agenda forward to unlock economic, environmental and social benefits. Emerging industries include waste management, mobility, renewable...

read more

License

Licence Ouverte v2.0 (Etalab)

Breakdown of green gas potential by 2050 by department

This data set presents the breakdown by department of the green gas potential by type of resource and by sector in France by 2050.

The hypotheses and the context of the study are presented in the study " A 100% renewable gas mix in 2050? " Published in February 2018 by ADEME, GRDF and GRTgaz.

The following primary resources are considered:

the potential for methane production with CIMSEs (Intermediate multi-service environmental crops), crop residues, livestock droppings, grasses, residues from agro-food industries (IAA), bio-waste and algae , in GWh PCS;
wood energy potential - in GWh PCI;
the recovery energy potential: Solid recovered fuels (SRF) and Fatal hydrogen (H2) - in GWh PCI;
the potential for electricity production by Power-to-Gas, in GWh elec.
In the study (see page 92), the injectable gas potential is obtained by applying the following conversion efficiencies to primary resources:
94% injected methane / produced methane,
70% for gasification (wood energy potential and CSR potential),
66% for power-to-gas.

Data and Resources

CSV
JSON
Excel
GeoJSON
Shapefile
KML

Additional Info

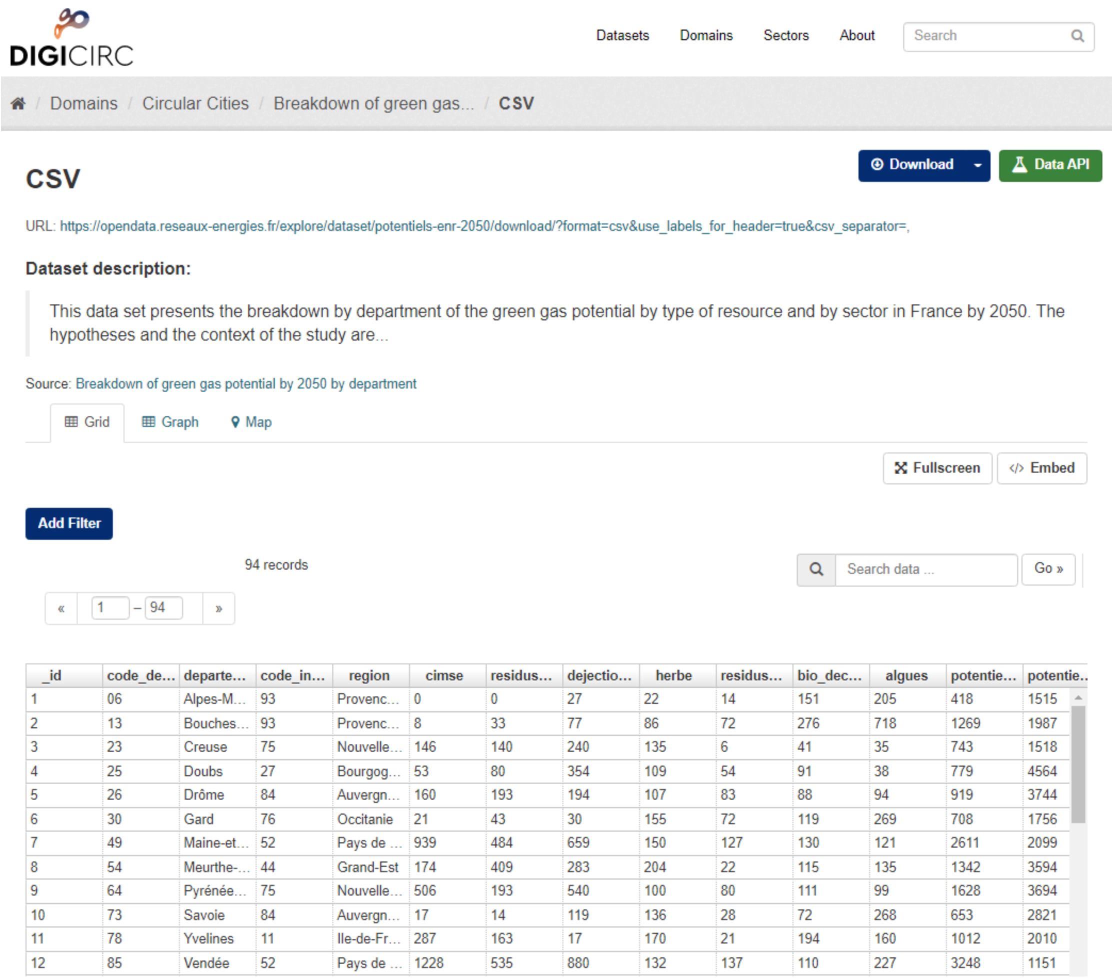
Field	Value
Source	https://opendata.reseaux-energies.fr/explore/dataset/potentiels-enr-2050
State	active
Last Updated	January 25, 2021, 3:12 PM (UTC+02:00)
Created	January 21, 2021, 6:08 PM (UTC+02:00)
Access level	Public
Geographical area	France
Language	French
Domain	GRTgaz, GRDF, ADEME
Sector	Energy

Figure 10: Dataset overview page

3.2.7 Data preview and visualization

In the resource page, information for the specific file is presented and the user has several preview options. The type of the file is presented on top, along with a link to its original source and the description of the dataset. Files in the format of CSV and XLS spreadsheets can be previewed in a grid view, with map (after definition of the coordinate data fields in the file) and graph views also available if the data is suitable. The resource page will also preview resources if they are common image types, PDF, or HTML.

By selecting the “Grid” option, the user can see in a table view the contents of the file. Sorting functionalities are available. Below the table, the user has access to the data dictionary, that includes information for the different fields the file consists of (Figure 11).



CSV

URL: https://opendata.reseaux-energies.fr/explore/dataset/potentiels-enr-2050/download/?format=csv&use_labels_for_header=true&csv_separator=,

Dataset description:

> This data set presents the breakdown by department of the green gas potential by type of resource and by sector in France by 2050. The hypotheses and the context of the study are...

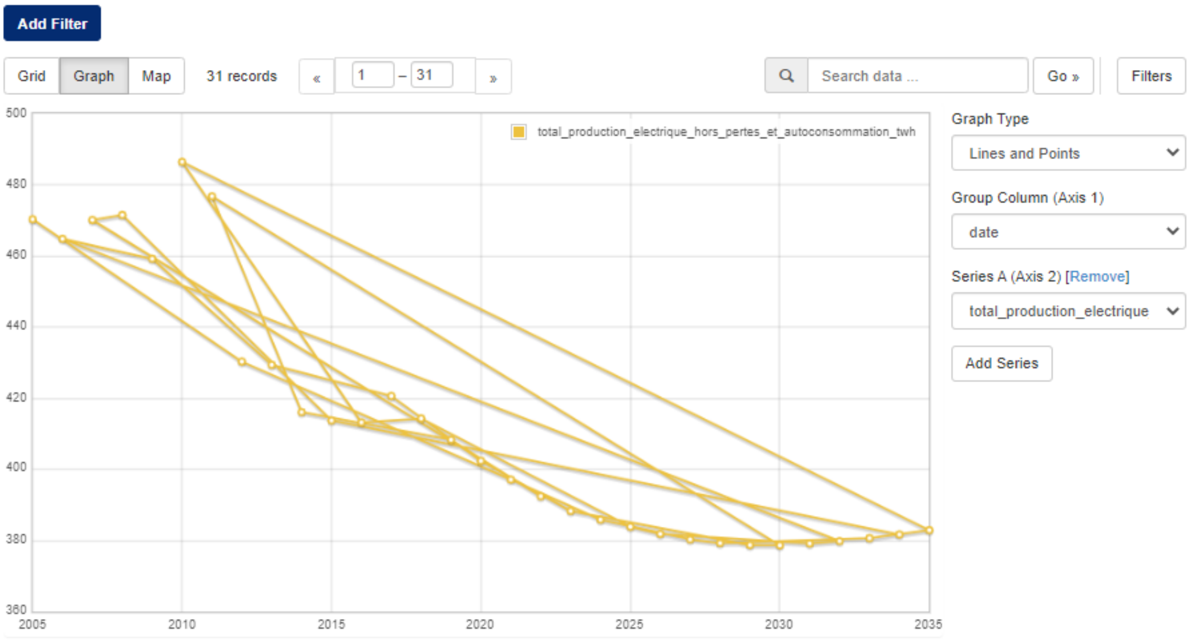
Source: Breakdown of green gas potential by 2050 by department

94 records

_id	code_de...	departe...	code_in...	region	cimse	residus...	dejectio...	herbe	residus...	bio_dec...	algues	potentie...	potentie...
1	06	Alpes-M...	93	Provenc...	0	0	27	22	14	151	205	418	1515
2	13	Bouches...	93	Provenc...	8	33	77	86	72	276	718	1269	1987
3	23	Creuse	75	Nouvelle...	146	140	240	135	6	41	35	743	1518
4	25	Doubs	27	Bourgog...	53	80	354	109	54	91	38	779	4564
5	26	Drôme	84	Auvergn...	160	193	194	107	83	88	94	919	3744
6	30	Gard	76	Occitanie	21	43	30	155	72	119	269	708	1756
7	49	Maine-et...	52	Pays de ...	939	484	659	150	127	130	121	2611	2099
8	54	Meurthe-...	44	Grand-Est	174	409	283	204	22	115	135	1342	3594
9	64	Pyrénée...	75	Nouvelle...	506	193	540	100	80	111	99	1628	3694
10	73	Savoie	84	Auvergn...	17	14	119	136	28	72	268	653	2821
11	78	Yvelines	11	Ile-de-Fr...	287	163	17	170	21	194	160	1012	2010
12	85	Vendée	52	Pays de ...	1228	535	880	132	137	110	227	3248	1151

Figure 11: Dataset explore page, preview of data in grid

The "Graph" option allows the user to create diagrams of five different types and select which fields of the file they want to display on the two axes of the graph. As a result, they are able to create dynamically any combination of data and get valuable insights for the dataset (Figure 12).



31 records

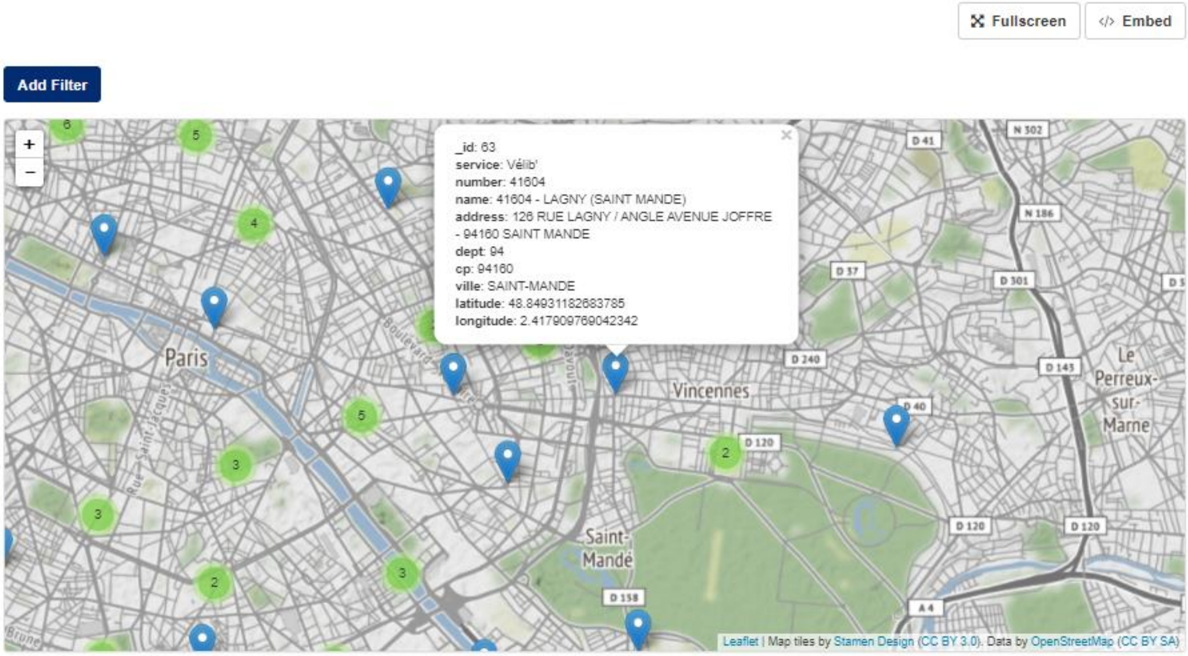
Graph Type: Lines and Points

Group Column (Axis 1): date

Series A (Axis 2) [Remove]: total_production_electrique

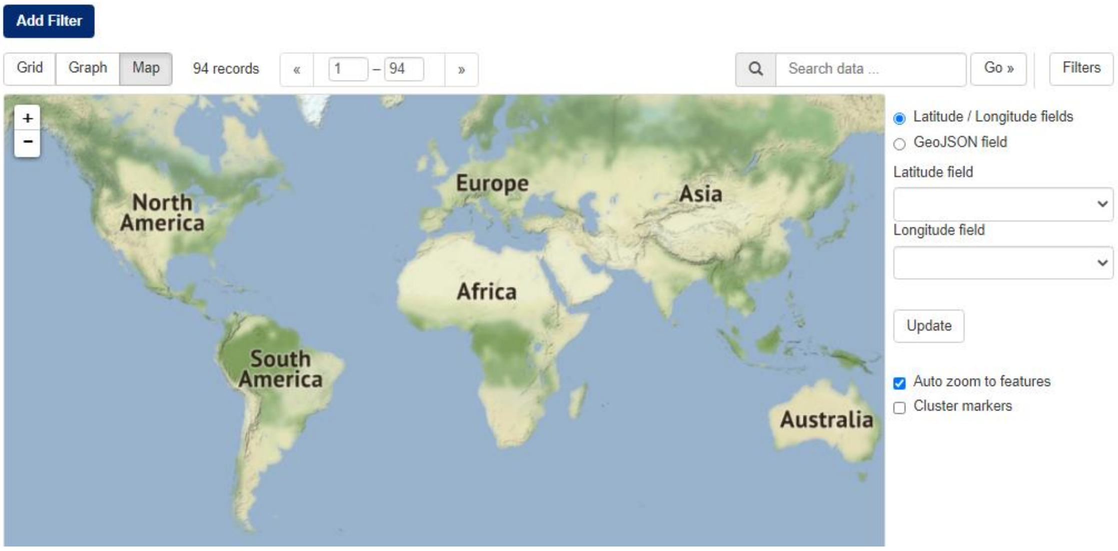
Figure 12: Dataset explore page, preview of data in a line graph

If the resource contains geospatial information, then a third option is available, to display the contents on a map by clicking on “Map” (Figure 13). The map is created automatically, but the user has the option to select the fields they want to display on the map (Figure 14). Clicking on points or polygons on the map displays the information of each asset.


Fullscreen
Embed
Add Filter
+
−
Paris
Vincennes
Saint-Mandé
Le Perreux-sur-Marne
Leaflet | Map tiles by Stamen Design (CC BY 3.0). Data by OpenStreetMap (CC BY SA)

_id: 63
service: Vélib'
number: 41604
name: 41604 - LAGNY (SAINT MANDE)
address: 126 RUE LAGNY / ANGLE AVENUE JOFFRE - 94160 SAINT MANDE
dept: 94
cp: 94160
ville: SAINT-MANDE
latitude: 48.84931182683785
longitude: 2.417909769042342

Figure 13: Dataset explore page, preview of data on a map

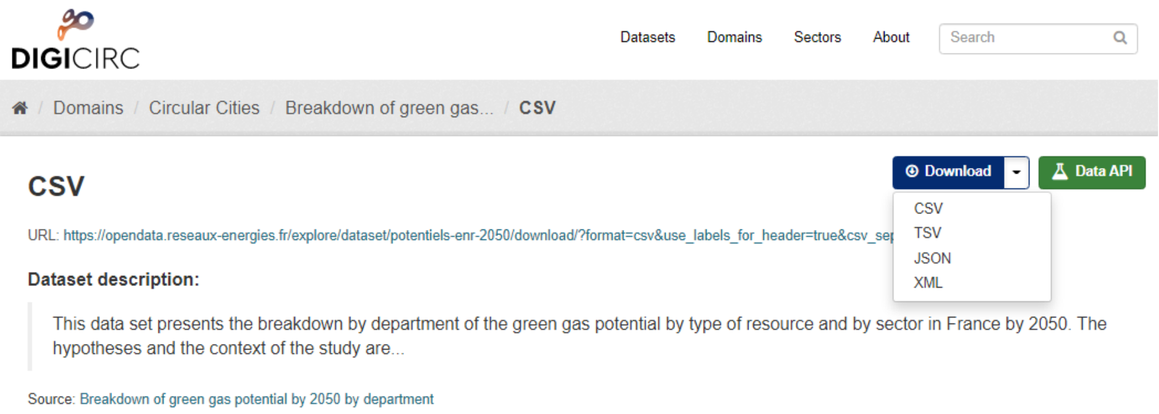


● Latitude / Longitude fields
○ GeoJSON field
Latitude field
Longitude field
☑ Auto zoom to features
□ Cluster markers

Figure 14: Dataset explore page, preview of data on a map with user selection

3.2.8 Download and API

Apart from downloading a resource directly from the dataset page, the user can navigate to the resource page where they are presented with more options. The download button is still available at the top of the page, where they can select the file format they want to download (Figure 15). Furthermore, the API option is available, which allows the user to access the data directly for use in their solutions. By clicking on the “Data API” option, the user is presented with a pop-up, where all the information for the API is available. More specifically, the available Endpoints are presented, along with several examples to guide the user. Links to documentation are also included (Figure 16).



CSV

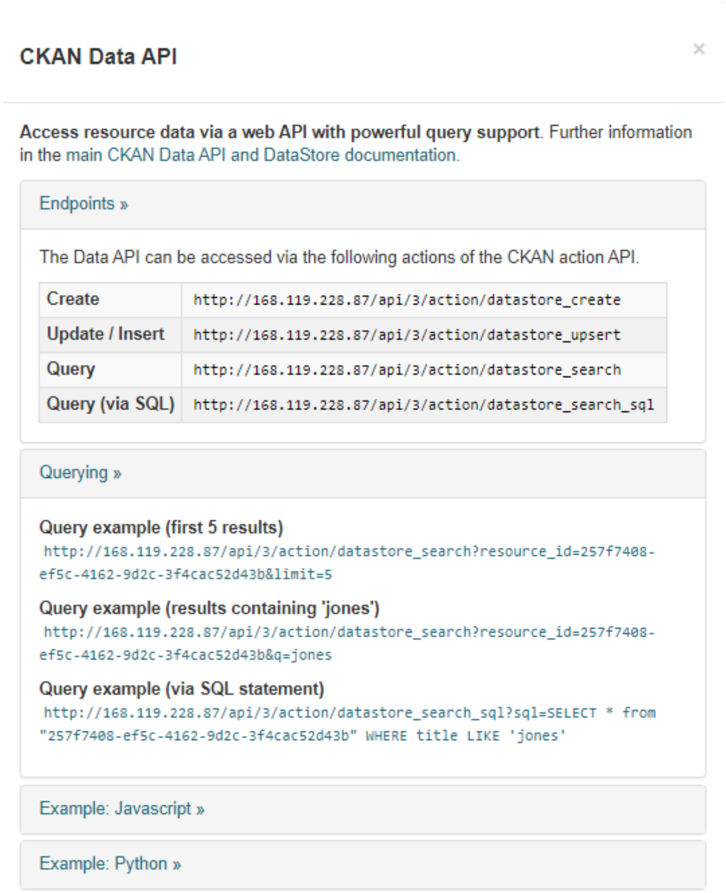
URL: https://opendata.reseaux-energies.fr/explore/dataset/potentiels-enr-2050/download/?format=csv&use_labels_for_header=true&csv_sep

Dataset description:

> This data set presents the breakdown by department of the green gas potential by type of resource and by sector in France by 2050. The hypotheses and the context of the study are...

Source: Breakdown of green gas potential by 2050 by department

Figure 15: Download files in resource page



CKAN Data API

Access resource data via a web API with powerful query support. Further information in the main CKAN Data API and DataStore documentation.

The Data API can be accessed via the following actions of the CKAN action API.

Create	`http://168.119.228.87/api/3/action/datastore_create`
Update / Insert	`http://168.119.228.87/api/3/action/datastore_upsert`
Query	`http://168.119.228.87/api/3/action/datastore_search`
Query (via SQL)	`http://168.119.228.87/api/3/action/datastore_search_sql`

Query example (first 5 results)

```
http://168.119.228.87/api/3/action/datastore_search?resource_id=257f7408-ef5c-4162-9d2c-3f4cac52d43b&limit=5
```

Query example (results containing 'jones')

```
http://168.119.228.87/api/3/action/datastore_search?resource_id=257f7408-ef5c-4162-9d2c-3f4cac52d43b&q=jones
```

Query example (via SQL statement)

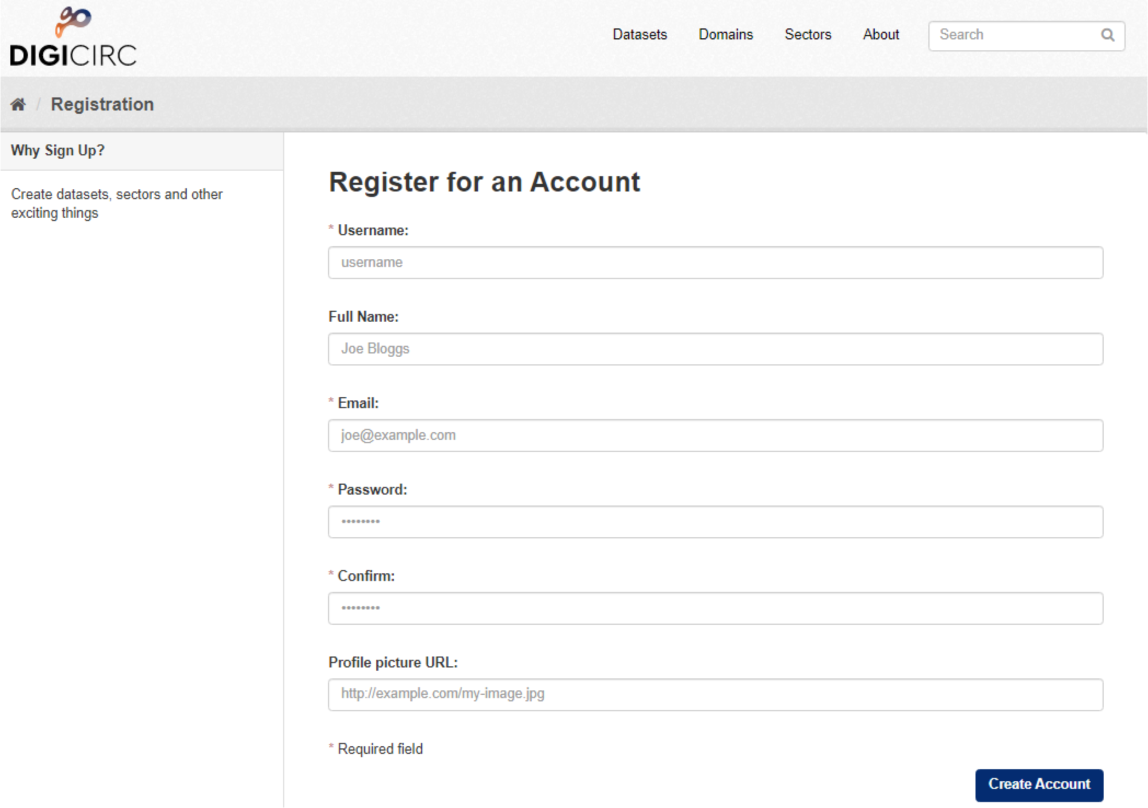
```
http://168.119.228.87/api/3/action/datastore_search_sql?sql=SELECT * from "257f7408-ef5c-4162-9d2c-3f4cac52d43b" WHERE title LIKE 'jones'
```

Figure 16: API available in resource page

3.2.9 Register and login

Users of the CE data hub can search for and find datasets without any restrictions. However, in order to upload a dataset, a login with appropriate permissions is needed. Anyone – not only DigiCirc project partners – can apply for registration. Registration is available through the welcome page when clicking “Do you want to contribute?” at the top right of any page. The required information for the registration is username, full name, e-mail address and password (Figure 17).

After the registration, the Administrator will assign editing authorization rights to the user, for them to be able to upload a dataset. The Administrator will add the new user to the three domains and will assign to them the “Editor” role, so they can edit and publish datasets to each domain (Figure 18).



Registration

Why Sign Up?

Create datasets, sectors and other exciting things

Register for an Account

* Username: username

Full Name: Joe Bloggs

* Email: joe@example.com

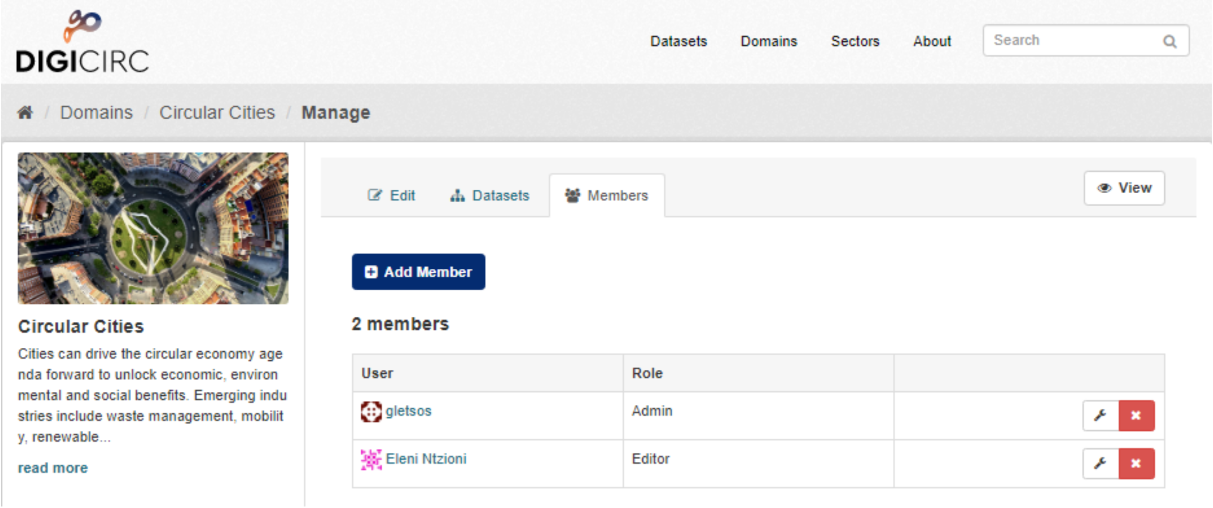
* Password: ••••••••

* Confirm: ••••••••

Profile picture URL: http://example.com/my-image.jpg

* Required field

Figure 17: Registration page



Circular Cities

Cities can drive the circular economy agenda forward to unlock economic, environmental and social benefits. Emerging industries include waste management, mobility, renewable...

read more

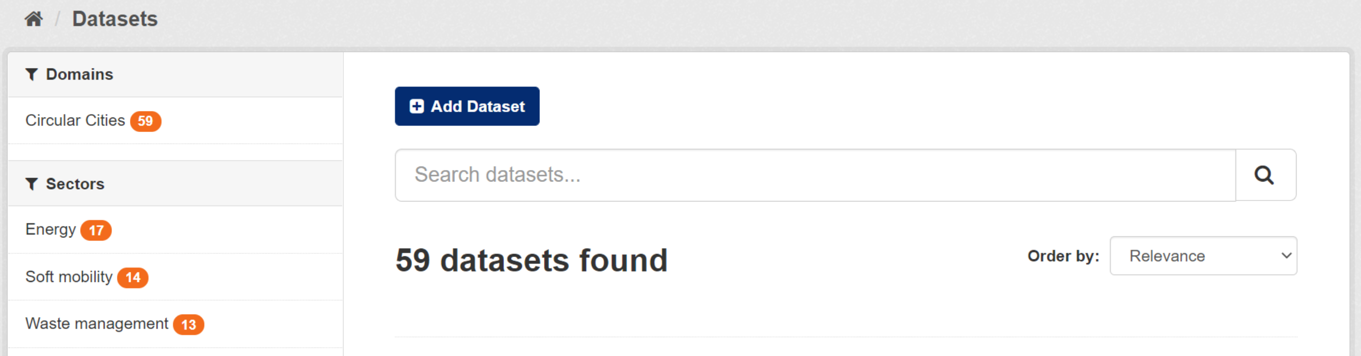
2 members

User	Role
gletsos	Admin
Eleni Ntzioni	Editor

Figure 18: Domain's page in the editing mode, tab "Members

3.2.10 Upload

In order to upload a file, the logged-in user can either select the "Add Dataset" button in the Datasets page (Figure 19) or navigate to the domain that the dataset is in relation to and then select the "Add Dataset" button above the search box (Figure 20). The "Add Dataset" button will be available after they have been assigned editing authorization rights by the Administrator.



Datasets

Domains

Circular Cities 59

Sectors

Energy 17

Soft mobility 14

Waste management 13

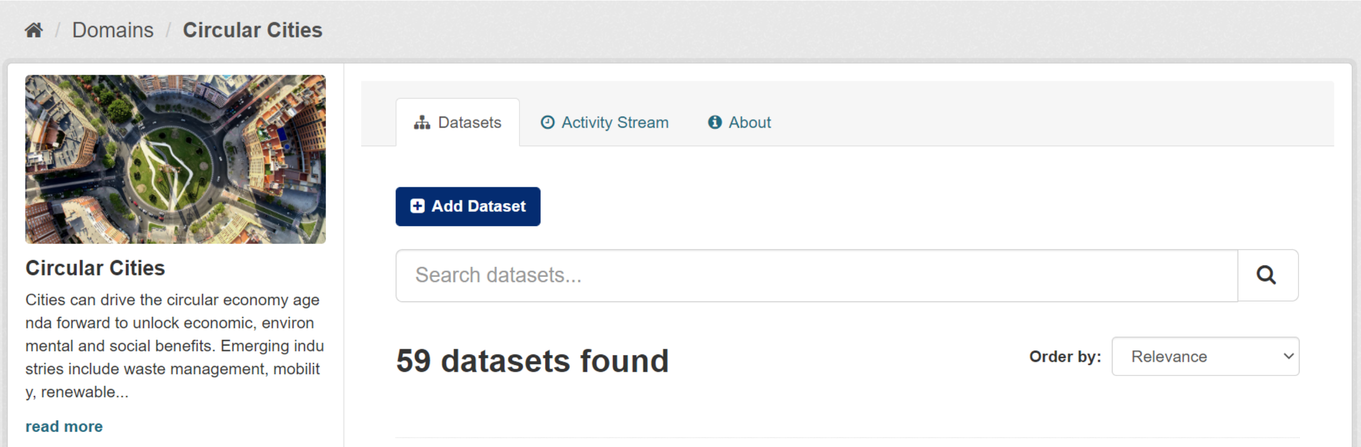
Add Dataset

Search datasets...

59 datasets found

Order by: Relevance

Figure 19: "Add Dataset" in Datasets page



Domains / Circular Cities

Circular Cities

Cities can drive the circular economy agenda forward to unlock economic, environmental and social benefits. Emerging industries include waste management, mobility, renewable...

read more

Datasets | Activity Stream | About

Add Dataset

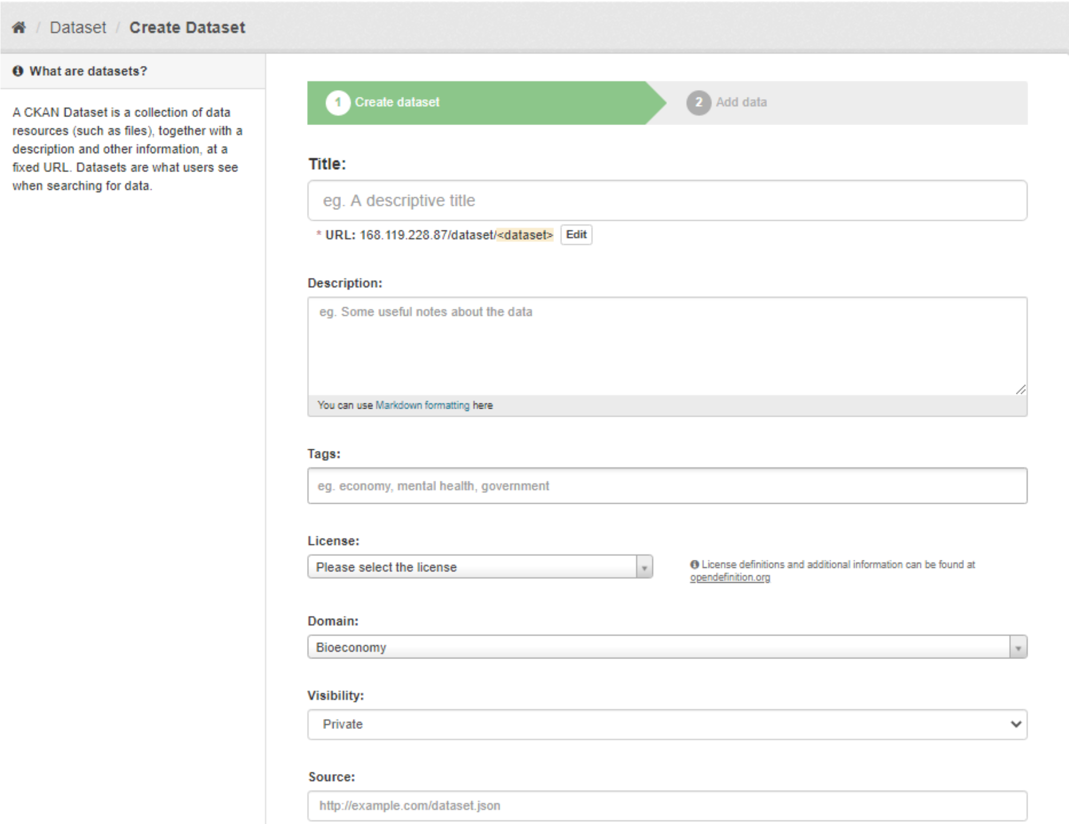
Search datasets...

59 datasets found

Order by: Relevance

Figure 20: "Add Dataset" in a Domain's page

For each new dataset, the user has to enter specific information. During the metadata entry the user can select if a dataset should be private or public. Before uploading any resources containing the actual data, the user needs to describe the dataset by providing metadata on the “Create dataset” page (Figure 21,). On the “Add data” the user can add one or more “resources” which contain the data for this dataset. To add a resource, the user can either choose a file from their local directory by pressing the “Upload” button or link to their data resource by clicking the “Link” button. Additionally, they should add a name for the resource - different resources in the dataset should have different names – and define the format of the resource, e.g., CSV, XLS, JSON, PDF, etc. This field can also be left blank as it will be guessed automatically (Figure 22).



Dataset / Create Dataset

What are datasets?

A CKAN Dataset is a collection of data resources (such as files), together with a description and other information, at a fixed URL. Datasets are what users see when searching for data.

1 Create dataset 2 Add data

Title: eg. A descriptive title

* URL: 168.119.228.87/dataset/<dataset> Edit

Description: eg. Some useful notes about the data

You can use Markdown formatting here

Tags: eg. economy, mental health, government

License: Please select the license

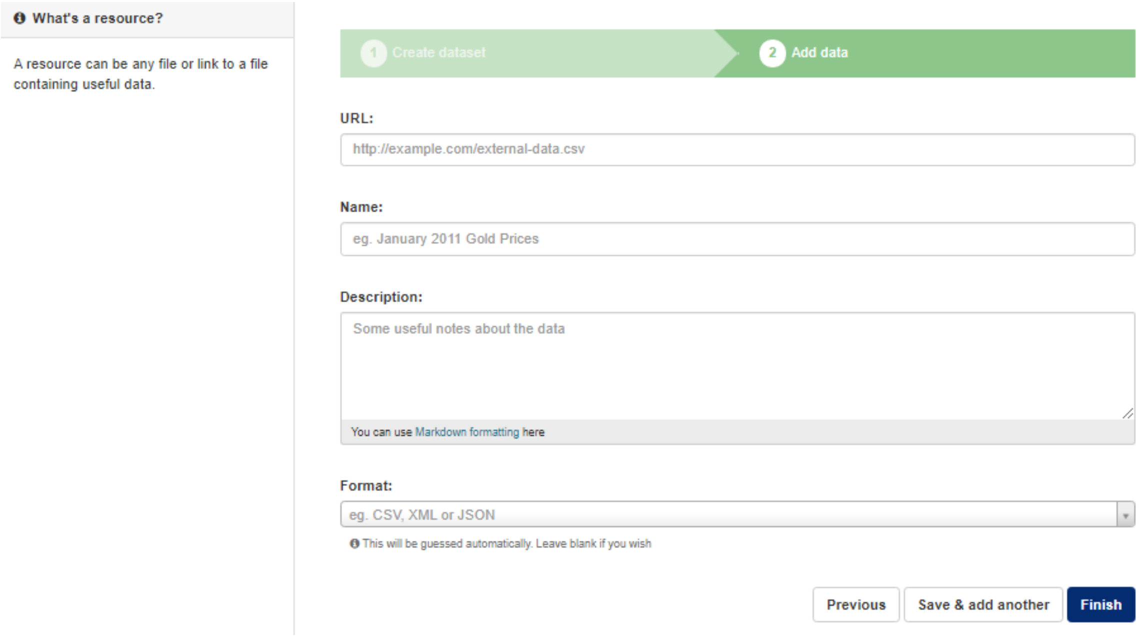
License definitions and additional information can be found at opendefinition.org

Domain: Bioeconomy

Visibility: Private

Source: http://example.com/dataset.json

Figure 21: Create dataset page - metadata fields



What's a resource?

A resource can be any file or link to a file containing useful data.

1 Create dataset 2 Add data

URL: http://example.com/external-data.csv

Name: eg. January 2011 Gold Prices

Description: Some useful notes about the data

You can use Markdown formatting here

Format: eg. CSV, XML or JSON

This will be guessed automatically. Leave blank if you wish

Figure 22: Create dataset page, upload of resources section

4 System Architecture

The design and development of the Circular economy data hub was based on the User needs that were produced with the help of the Development reference groups and are described in *D3.1 User Needs and Development Assessment*. The user needs were thoroughly analyzed and research was conducted for open-source data management systems that integrate tools to streamline publishing, sharing, finding and using data as well as previewing and visualizing data to discover insights before using them. CKAN open-source software[4] was selected and the system architecture was tailored to include this addition.

The first version of the architecture consisted of suitable subsystems that were selected to support the services required to meet the user needs. That version served as a good starting point and during the development phase it was slightly modified to fit the needs of the requirements and the functionalities of the data hub and the introduction of CKAN software. The following description is an updated version of the system architecture description in *D3.1 User Needs and Development Assessment.*

The architecture consists of four main subsystems whose purpose is the implementation and support of the overall process of data gathering, data import, data searching and browsing by the users and data accessing. It includes subsystems responsible for the collection, processing and storage of data relevant to the three thematic areas of DigiCirc as well as the search and presentation of results to the end user through a friendly online environment or accessing through APIs.

The system architecture will meet basic technical requirements such as:

- Availability: Continuous provision of services to the end user
- Extensibility: Ability to extend the architecture to support new services
- Security: Protection against risks, viruses, access breaches, publication of incorrect data
- Scaling: Ability to upgrade requirements for maximum performance
- Reliability: Accuracy and consistency of services provided
- Ease of management: Monitoring procedures to ensure quality service delivery

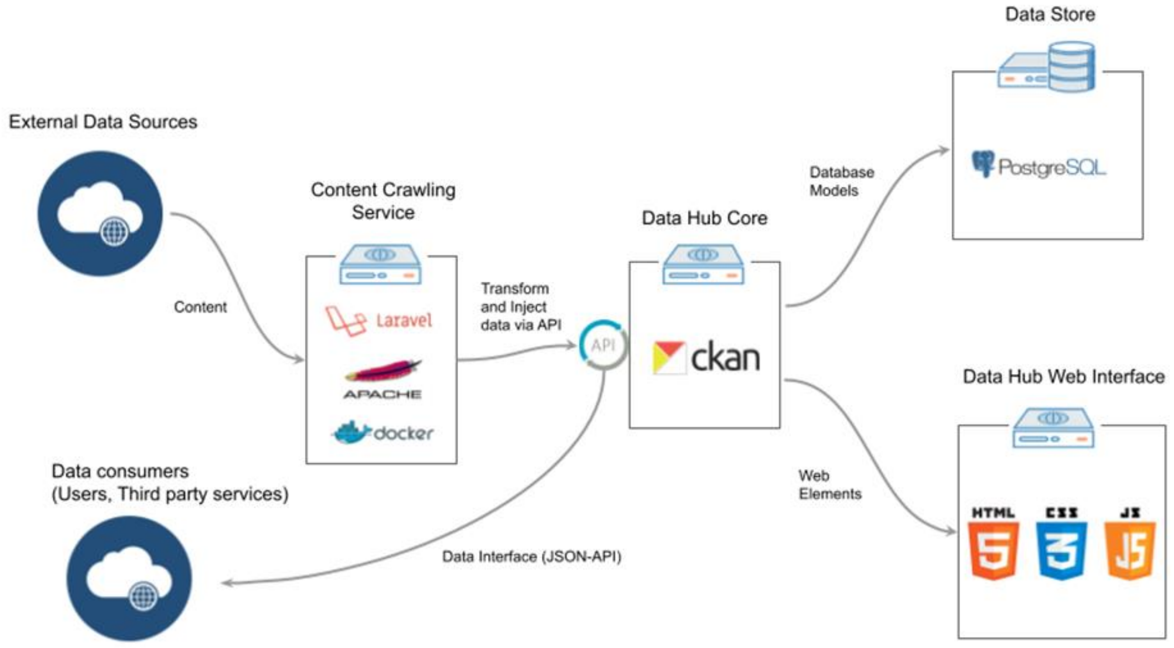


Figure 23: Circular economy data hub architecture

Figure 23 shows the architectural diagram of the four main subsystems. The flow of data starts with the "Content Crawling Service", which is responsible to gather the data from all available data sources, to transform them in

[4] https://ckan.org/

appropriate types compatible with the rest and finally injects the data to the core subsystem (Data Hub Core) via exposed API. Storing of the processed data will be handled by the "Data store" and they will be provided to the users either through APIs directly to other applications or through the "Data hub Web interface" which will be the means of interaction between the system and the users.

The Data Sources contain all the information that will be defined in Task 3.1 during the course of the project with the assistance of the cluster partners, which will also be described in future deliverables *D3.2 Circular cities testbed dataset*, *D3.3 Bioeconomy testbed dataset* and *D3.4 Blue economy testbed dataset*. The sources gathered so far are available in many forms and formats either as a part of a web page, as a structured web service (e.g. a RestFul API), as a file (e.g. an excel or CSV file) or even as an open database. They mainly involve official data sources and contributions that will be uploaded by the users and are the starting point of our information flow.

4.1 Data Hub Core

The Data hub core subsystem, based on Python programming language and CKAN Portal, acts as an Orchestrator of the system while applying the business processes to the data it draws from the peripheral subsystems.

The main functionalities of the Data hub core subsystem are:

- The receipt of data coming from the crawling service through the User Interface Subsystem.
- The storage of files and geospatial data in the corresponding components of the database subsystem.
- Apply the business rules of the platform.
- Exchange data between the components of the Data handling subsystem.
- Convert the results to a user-friendly format and display them through the Interface system

A mechanism responsible for the security of the system and its contents is also implemented. Access control lists include lists of users, groups and roles that have access to the content (e.g., access to sensitive data) or interact with components of the platform. This unit also shields the application from third-party attacks such as SQL injection, cross-site scripting (XSS), cross-site request forgery (CSRF), and cookie hijacking.

4.2 Content Crawling service

In order to get the data from the sources, appropriate software in the form of several crawlers was created to perform Data Ingestion. The data sources are static or they get updated at various time intervals and sometimes they contain measurements at different units for the same parameters or even provide values with different time granularity (e.g., hourly, daily, weekly etc), therefore some initial transformation and harmonization techniques needed to be applied to the crawled data. Finally, data enrichment was required to add further attributes to the data to allow for easier data searching.

A typical process of a crawler is the following:

1. Start the crawler manually or automatically
2. Gather the data from the source
3. Quality control
4. Harmonization (pattern handling)
5. Enrichment
   5.1. Enrich the dataset with information needed in the user interface
   5.2. Metadata enrichment to include other needed fields
6. Transformation of the data in order to be in the correct format for the preview functionality
7. Store the data in the Data Store component

The source code of the subsystem is based on the PHP programming language implemented by the PHP Framework, Laravel[5]. Laravel is a complete PHP Framework that not only facilitates development through built-in programming tools, but facilitates installation as well, compared to other frameworks. The design of the service is based on the model *Model – View – Controller*, an architectural pattern that separates an application into three main logical components. This model will ensure the better distinction between the business logic of the application and its presentation, thus allowing the separation during the development process and the easier maintenance of the subsystems in the future.

4.3 Data Store Component

The Data Store Component is used to store all the data of the system. The data types deriving from the Content Crawling service are stored in a modeled DB format. The types of data that refer to the business data, defined as the data of users and generally data related to the business operation of the system, are better handled by an object-relational database, such as PostgreSQL. It offers a variety of powerful index types to best match a given query workload. Furthermore, it offers performance optimizations including parallelization of read queries, table partitioning, and just-in-time (JIT) compilation of expressions. Apart from these characteristics, what makes PostgreSQL a proper choice, is its spatial database extender, PostGIS[6], that supports geographic objects and allows location queries in SQL. GIS data will be stored in the database creating attribute tables, where every entry represents geographical geometries. Since a great percentage of the collected data will have geospatial information, directly querying a PostGIS database will allow for more powerful and precise insights.

4.4 Web Interface subsystem

The User Interface subsystem is the human-computer point of interaction, where the graphical interface (GUI) of the web application is developed. The Web Interface is based on HTML and JavaScript, in combination with Jinja[7], a modern and designer-friendly templating language for Python, embedded by CKAN. Jinja is fast, expressive and extensible. Special placeholders in the template allow writing code similar to Python syntax and then the template is passed data to render the final document.

[5] https://laravel.com/

[6] https://postgis.net/

[7] https://jinja.palletsprojects.com/en/2.11.x/

5 Hardware and Software requirements

The infrastructure of the entire system is based on a single server running under Ubuntu Linux operating system and consisting of 9 Docker Containers. A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application. The Dockers allow the application to run anywhere, regardless of the operating system, the existence or not of a cloud infrastructure, or whether it is a physical or virtual environment.

The table below depicts the hardware needs of the system and the software running.

Table 1: Characteristics and Specifications of the infrastructure

Characteristics	Specifications
Operating System	Linux Ubuntu 18.04 (or higher)
Required Installed Software	Docker
CPU	4 Cores
RAM	8GB
HDD	300GB

The instance is based on cloud infrastructure offering vertical scale capabilities (scale up) via an easy-to-use control panel. The cloud-based instance offers flexibility for adding more CPU, memory, or I/O resources to the existing instance, or replacing with a more powerful instance without manual intervention to the physical server. Daily processes are responsible to produce database and uploaded data offsite backups.

6 Hub expandability

The CE data hub architecture is developed in such a way that allows future integration and expandability of data provision services. The software infrastructure of the hub is able to accommodate any datasets which will facilitate programmers to develop their solutions. During the next period, before the beginning of the accelerator for the Circular cities domain, additional datasets will be incorporated and offered through the CE data hub. These datasets fall under the existing sectors and they will be retrieved from the existing data sources.

Some of the additional datasets will be accessed through requests to certain APIs and they refer to constantly updated data that need to be collected at certain time intervals. For that purpose, automated tasks will be scheduled periodically with cron-jobs (software used for scheduling tasks to run on the server), that will trigger the crawling mechanism enabling any new data to be crawled in the CE data hub. Specifically, in the case that new data exists, it will be grabbed and ingested in the system and then forwarded to the content parser which will extract the information and store it in the database.

Furthermore, the same process will apply to the Blue economy and Bioeconomy domains, later in the course of the project. As with the current case, a careful evaluation of the possible sources of information will be conducted for the other two domains with the assistance of the cluster partners, a qualification of the datasets using template will be performed and the datasets will be incorporated and offered through the hub. Throughout this process, datasets of stakeholders, open sensor measurements available on the internet, and information from web page tables will be retrieved, parsed and ingested into the CE data hub in a structured way.

DIGICIRC

End of Document

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 873468.