Situation-Aware Linked heTerogeneous Enriched Data

D1.1: Report on semantic features extraction for heterogeneous data sources contextualization

Work package	WP 1
Task	Task 1.1, T2.1, T1.3
Due date	31/12/2022
Submission date	23/12/2022
Deliverable lead	IMT
Version	1.0
Authors	Roberto Minerva (IMT), Noel Crespi (IMT), Amir Reza Jafari Tehrani (IMT), Syed Mohsan Raza (IMT), Luis Sánchez (UC), Jorge Lanza (UC), Juan Ramón Santana (UC), Pablo Sotres (UC), Victor González (UC), Laura Martín (UC), Anja Summa (Kybeidos), Maren Dietzel (Kybeidos), Stephan Frenzel (Kybeidos), Benjamin Hebgen (NEC)
Reviewers	Ernö Kovacs (NEC), Luis Sánchez (UC), Stephan Frenzel (Kybeidos)

This project is co-financed by the Connecting Europe Facility of the European Union under the Action Number 2020-EU-IA-0274.

This project is co-financed by the Connecting Europe Facility of the European Union under the Action Number 2020-EU-IA-0274.

Abstract	This document, developed by the SALTED project, represents the D1.1 deliverable of the semantic features extraction for heterogeneous data sources contextualization. The focus of this document is to give an introduction to different data sources used in SALTED as well as smart data models for converting into the NGSI-LD architecture. Furthermore, D1.1 defines the injection chains of different sources and their contributions to the SALTED architecture with a few possible use cases.
Keywords	Heterogeneous Data Sources, Injection Chain, Smart Data Models

This project is co-financed by the Connecting Europe Facility of the European Union under the Action Number 2020-EU-IA-0274.

Table of Content

Table of Figures

1 INTRODUCTION

1.1 SCOPE OF DOCUMENT

Digitalization of contents and media, softwarization of many different activities and services, and increasing level of measurement of the environment are producing many distinctive data sets. They are, largely, characterized by specific formats as well as different ways and timing for collecting data. Many different data sets are collecting very similar information by means of slightly different formats. On the other side, the availability of usable data is governed by different processes that dictate the timing, and quality of the available data. These and other differences are a hurdle for the smooth usage and exchange of data.

In the SALTED project, the concept of injection chain has been defined [1] in order to fully describe the process of collecting, organizing and improving the data. The goal of an injection chain is to obtain well-formed usable data that is acceptable also from a formal perspective. In addition, this process adds additional functionalities and features to the data, in the form of linked meta-data, capable of contextualizing and improving the content of the data. In addition, the usage of established and formally defined Data Models helps in contextualizing the data as well as semantic extraction and organization of information.

The main outcome of the SALTED injection chains are datasets and data-streams conformant to the NGSI-LD information model and curated to assure its data quality so that homogeneous usage and exchange of them is enabled. Moreover, the linked-data and semantic characteristics of the NGSI-LD information model and the employed Data Models enable further enrichment of the data in subsequent steps of the SALTE Data Enrichment Toolchain (DET) on which further metadata can be linked to the individual pieces of data, as well as relationships can be established among them, thus creating reach knowledge graphs that can be queried through Linked-Data Context Brokers exporting the NGSI-LD API[1]

Data injection is a costly activity that requires investment and that should be framed into a well-suited process. Data collection, i.e. the acquisition of raw data, is only the starting point for a long process that should lead to the curation, certification and distribution of well-formed and reliable, trustable data. Data management, curation and completion are activities that should be automated as much as possible to avoid humans spending too much time on these processes and introducing additional errors or discrepancies in the datasets. Figure 1 represents the main phases and some of the sub-phases of the entire SALTED Injection Chain concept procedure.

[1] https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.06.01_60/gs_CIM009v010601p.pdf

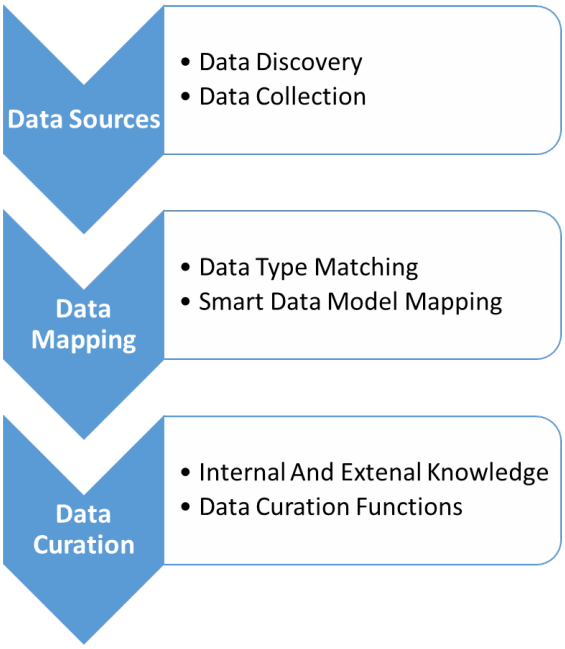


Figure 1 SALTED Injection Chain

One of the objectives of the SALTED project is the definition of chains of data preparation from the discovery of relevant data sources up to the consolidation of the available data and their storage in a distributed data management system. Different application domains need specific data and they need to be searched, acquired and organized according to the specific requirements of the Application Domain. There are many activities and steps in the extraction, curation and consolidation chain that are very specific to the application domain at hand. In addition, depending on the specific intended usage of the data, the application may need to access them in real-time, i.e., when the data is produced; or in batch format, i.e., after time series of a certain period have been created. Nevertheless, there are some high-level commonalities that have been described and consolidated in Deliverable D2.1 [1].

The so-called injection chain is a generic model presenting a process that takes care of the data from its discovery to the curation and injection of this data into a Linked-Data Context Broker. Each specific application domain, use case or service can rely on this general model and specialize it according to specific needs and treatments of the data in that context. This document refers to the general process and illustrates a set of use cases that may have specific requirements for the treatment of data. The Reader should be familiar with the injection chain of the SALTED project and, in general, the definition of the SALTED architecture as described in D2.1 [1].

Firstly, data has to be discovered and collected. Discovery of available data can be done directly by requesting data management platforms (e.g. IoT Platforms and/or Open Data Platforms) or crawling the web or social network channels. Furthermore, data collection also adds to the heterogeneity that the SALTED injection chain is aiming to harmonize. In this regard, we are mainly differentiating between data streams that are accessible in an asynchronous way in which data is notified to those consumers that have made the corresponding subscription, and datasets, that are available upon request (i.e. in a synchronous, request-response kind-of operation).

Of particular importance is the Data Mapping phase, in which the data collected from different sources are mapped to NGSI-LD based Data Models. This mapping increases the value of data because the data extracted are organized and represented accordingly to Data Models and supporting ontologies [2] [3] [4]. This process enables the possibility of reasoning and comparison of these data with other datasets organized in similar ways as well as its extension through other complementary data models and knowledge graphs.

The next step of the process, i.e., the Data Curation, increases the quality of data and it organizes the data in order to reflect the contexts in which they were generated in order to be better used in the specific application case. The knowledge of data and problem domain makes the data more aligned with the intended usage of the applications.

Under this perspective, the injection chain has a twofold outcome. On the one hand, it is capable of extracting the most relevant semantic features of the data and contextualizing them accordingly to the need of the specific applications. On the other hand, the linked-data-based information model employed enables the continuous extension of the available datasets through new metadata and the establishment of relationships.

1.2 Target Audience

The "Semantic features extraction for heterogeneous data sources contextualization" Document has a twofold goal: on one side, it is intended for internal use in order to drive the design and implementation of subsequent phases of the DET (i.e. data linking and enrichment) or application use cases for the project or internally to the partner organizations; on the other hand, it is publicly available in order to increase the awareness of the SALTED activities, its relation with the FIWARE ecosystem and it points to the important issue of the data treatment and consolidation for interoperable applications and domains.

Thus, the target audience is both the SALTED technical team including all partners involved in the delivery of work packages 2, and 3, but also it serves as a reference for developers of applications as defined by WP4 and in general situation-aware applications when the data handled through the described injection chains will be made available to a large audience of potential data consumers.

1.3 Structure of the Document

This document has identified a set of "interesting" use cases such as smart city, social media, government and statistical certified sources and meteorological applications. For each of the use cases identified, there is a section fully devoted to the description of the raw/original data and how to collect them, the chosen NGSI-LD data model employed, created or extended according to the needs of the application, the function identified for improving the quality of data and mitigate the errors, and how to "measure" the quality of the data at hand. When it is the case, a distinction between real-time flow and batch usage of data is indicated and described.

The goal of this editorial organization is to provide the Reader with the possibility to focus on the use case of interest and to extract all the relevant information, or to browse the entire

document in order to acquire knowledge about the issues and problems posed by different injection chains and how well the SALTED architecture can adapt to accommodate the changes.

The conclusion section is a short chapter aiming at providing some general considerations about the usage of the injection chain and its difficulties or advantages.

1.4 RESOURCES

There are two major resources that the Reader may consider knowing before hands: one is the Deliverable D2.1 with the specification of the SALTED Architecture and an in-depth description of the injection chain; the other is the set of Smart Data Models associated with the FIWARE Foundation[2]. In there, some Data models that have been used and modified in this document are present.

A knowledge of pub-sub mechanisms and of specific tools supporting the MQTT protocol could be useful for the Reader, but they are not essential for the comprehension of this document.

[2] https://www.fiware.org/smart-data-models/ [last checked 17/11/2022]

2 THE SMART CITY DOMAIN

2.1 CHARACTERIZATION OF DATA

2.1.1 Data Sources

There are many smart cities in the world and each of them is producing and dealing with a wealth of data. These data, however, are collected, stored and formatted in different fashions. There is a need to make these data sources comparable and, somehow, interoperable. NGSI-LD representations could be very helpful in order to harmonize and making the data comparable. However, the diversities in data are so relevant that a simple transformation from one format to another is not enough. For instance, the city of Dublin is organizing its traffic intensity data quite differently form Madrid. In order to make these data sets comparable, they have to go through a process of transformation of data formats. The city of Madrid is organizing its data in a similar way as Santander, but with different acquisition times. The integration of these differences requires the usage of common data models and particular attention to the timeline of the data. Hence, a longer process comprising collection, cleaning, curation, and augmentation of these data is needed. This chapter presents a set of different examples of how these data are generated and how they can be made more usable.

Data stemming from different cities are considered. In addition, different collection mechanisms (stream or batch) are used in order to demonstrate the applicability of the SALTED injection model.

Santander

The city of Santander publishes data from different categories in its large Open Data portal, available on its website[3]. Data are available in different modalities (quasi real-time and batch) in order to support different usages from applications. The data produced on a timely basis are formalized and they subsequently form the batch data sets. The traffic data is especially impressive, covering quasi real-time vehicle data, bicycle lane and docking station data, public buses location data and more.

In the Smart City domain, we will be focusing on the data we have been harvesting for the SALTED platform so far: Bike hires docking stations, bus fleet control, and historical datasets on road traffic status. All these datasets, except the last one, are provided on a periodic basis, ranging from 5 to 15 minutes. The coverage is very dense in the city center and sparser towards the outskirts, as is usually the case in most Smart Cities.

Moreover, the SmartSantander IoT infrastructure[4] has thousands of sensors providing several kinds of mobility and environmental data (cf. Figure 2). We are currently collecting air quality, battery sensor status, electromagnetic field, parking and temperature data. All the sensor observations are obtained through a REST subscription API, and therefore, we have labelled this data as stream data. Since data is received in an asynchronous way, it causes us to use a completely different collection procedure.

[3] https://datos.santander.es/data/
[4] https://api.smartsantander.eu/

Figure 2: SmartSantander IoT infrastructure

Finally, we are collecting parking data from the LoRaWAN sensors distributed across the city (cf. Figure 3). This data is provided through an MQTT endpoint, and therefore it qualifies as stream data as well.

Figure 3: Parking sensors deployment in the Santander's downtown

Madrid

The city of Madrid has an impressive system of sensors for monitoring different aspects of the urban environment. It has more than 5.000 traffic intensity sensors, and around 20 pollution/meteorological stations scattered in the territory. In addition, it gathers noise-related measurements. This information together with other data makes a large data catalogue[5]. This is a large data catalogue containing relevant information that represents the behaviour of a large city. In this version of the document, the relevant sources of information will be those related

[5] A description is available at https://datos.madrid.es/portal/site/egob/menuitem.9e1e2f6404558187cf35cf3584f1a5a0/?vgnextoid=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

to traffic[6] and pollution[7]. For this work, we are interested in the real-time data collected for the traffic and pollution data[8]. The traffic data are collected once every 15 minutes for each sensor, while the pollution data are collected hourly. The deployment of traffic sensors and air quality stations is very dense and it covers different scenarios. For instance, traffic sensors are deployed in every lane and not necessarily only at junctions. Air Quality stations are deployed in protected areas (e.g., close to parks) as well as in very dense populated areas. This gives the opportunity to study phenomena in different scenarios. Figure 4 shows the deployment of the sensing system in Madrid.

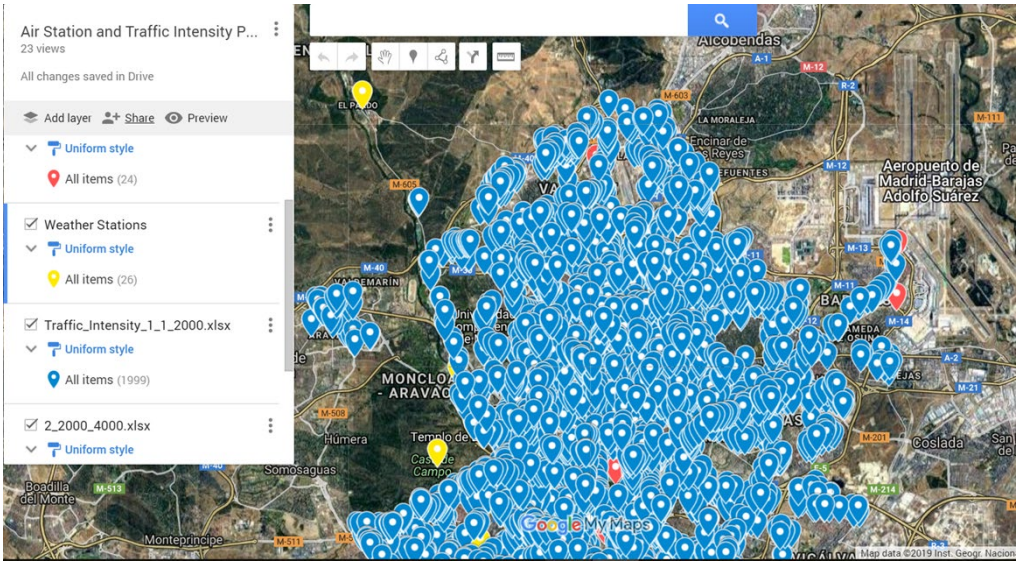


Figure 4: Sensing infrastructure in Madrid

Dublin

The City of Dublin makes available a large set of data related to traffic and a real-time API for getting the measurement related to Noise and Air quality. The sensing network (even if not as big as the Madrid one) is quite relevant and it covers a large part of the urban environment[9]. The traffic sensors are mainly deployed at junctions and their values are collected hourly. This makes some differences in the possible usage of information compared to Madrid and Santander in terms of granularity and time availability.

Figure 5 shows the Dublin traffic-sensing infrastructure:

[6] A description is available at https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=02f2c23866b93410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM1000000171f5a0aRCRD

[7] A description is available at https://datos.madrid.es/portal/site/egob/menuitem.754985278d15ab64b2c3b244a8a409a0/?vgnextoid=20d612b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextchannel=20d612b9ace9f310VgnVCM100000171f5a0aRCRD&text=contaminaci%C3%B3n&buscarEnTitulo=false&btn1=buscar.

[8] Available at https://datos.madrid.es/sites/v/index.jsp?vgnextoid=f3c0f7d512273410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD

[9] See https://data.smartdublin.ie/dataset/traffic-signals-and-scats-sites-locations-dcc

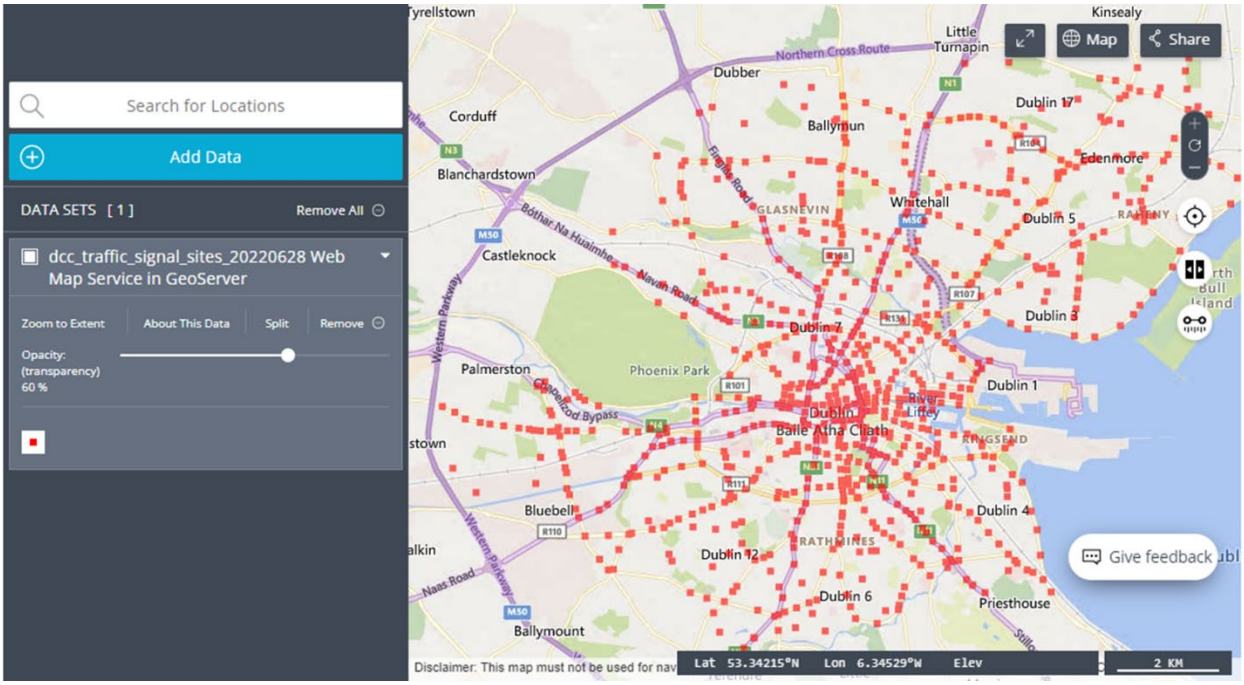


Figure 5: Traffic Sensors in Dublin

Rhine-Neckar Metropolitan Region

With the application of Smart City technologies, the Rhine-Neckar Metropolitan pursues the following goals:

- Provision and utilization of open data for holistic social, ecological and economic action in the region.
- Linking existing infrastructures and digital supplements.
- Building networks among service providers, guests and tourists through digital tools in order to work out potentials and use synergy effects.
- Active participation in processes (open government tools) to find long-term and sustainable solutions.
- Capture accurate needs analysis and identify mobility gaps through Smart Mobility.
- Achieve CO2 reduction through improved public transport services and set up a CO2 compensation scheme with the possibility to support (regional) compensation projects in order to increase environmental sustainability and demonstrate the impact of own actions.
- Increasing the resilience of cities - especially with regard to the consequences of climate change, e.g. heavy rainfall and heat islands.

Other Spanish cities

Most cities in Spain provide an Open Data portal that allows for the collection of batch data through periodic requests. We have made use of this possibility in the cities of Murcia[10],

[10] https://datosabiertos.regiondemurcia.es/

Barcelona[11], Valencia[12], Vitoria[13] and Bilbao[14]. We have collected air quality data from Murcia, which is updated every day, bike station data from Barcelona, and traffic status data from Barcelona, Valencia, Vitoria and Bilbao. These datasets are updated in a range of 5 to 15 minutes.

2.2 TYPE OF SMART CITY DATA

2.2.1 Traffic status and intensity data

Characteristics of batch data in Madrid and Dublin

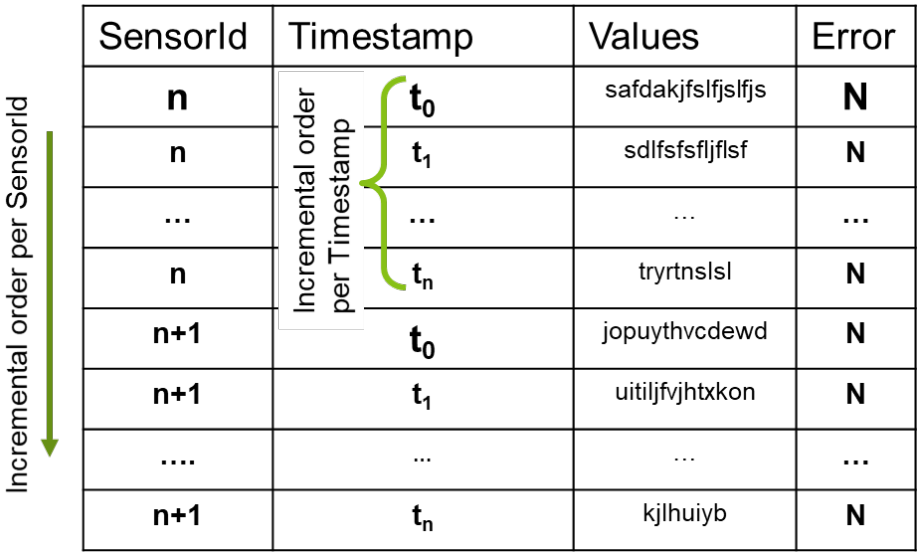
The collection and storage of data in Madrid and Dublin differ in several ways. Data in Madrid are collected in intervals of 15 minutes. The available files (in CSV format) are organized by sensor identifier and timestamp over periods of 15 minutes. Table 1 shows an example of these data.

Table 1: Traffic data organization in Madrid data sets

id	fecha	tipo_elem	intensidad	ocupacion	carga	vmed	error	periodo_integracion	
3731	4/1/2019 0:00	URB	68	3	17	0	N	13	
3731	4/1/2019 0:15	URB	45	1	8	0	N	15	
3731	4/1/2019 0:30	URB	39	3	11	0	N	15	
3731	4/1/2019 0:45	URB	30	0	6	0	N	15	
3731	4/1/2019 1:00	URB	13	0	6	0	N	15	
3731	4/1/2019 1:15	URB	32	2	7	0	N	15	
3731	4/1/2019 1:30	URB	19	1	6	0	N	15	
3731	4/1/2019 1:45	URB	17	1	6	0	N	15	
3731	4/1/2019 2:00	URB	13	1	5	0	N	15	
3731	4/1/2019 2:15	URB	24	0	4	0	N	15	
3731	4/1/2019 2:30	URB	7	1	4	0	N	15	
3731	4/1/2019 2:45	URB	21	0	1	0	N	15	
3731	4/1/2019 3:00	URB	16	1	6	0	N	15	
3731	4/1/2019 3:15	URB	13	1	4	0	N	15	
3731	4/1/2019 3:30	URB	5	0	1	0	N	15	
3731	4/1/2019 3:45	URB	29	1	6	0	N	15	
3731	4/1/2019 4:00	URB	39	0	8	0	N	15	
3731	4/1/2019 4:15	URB	43	1	8	0	N	15	
3731	4/1/2019 4:30	URB	37	2	10	0	N	15	
3731	4/1/2019 4:45	URB	16	5	7	0	N	15	
3731	4/1/2019 5:00	URB	22	0	9	0	N	15	
3731	4/1/2019 5:15	URB	32	1	6	0	N	15	

[11] https://opendata-ajuntament.barcelona.cat/data/es/dataset
[12] https://valencia.opendatasoft.com/pages/home/
[13] https://www.vitoria-gasteiz.org/j34-01w/catalogo/portada
[14] https://www.bilbao.eus/opendata/es/inicio

The information content also provides information about the intensity, the load and other information such as the average speed. In addition, the file indicates if for each represented measure, there is an error (field error in the head of the template). Aiming to the curation of data, this may represent a value related to the reliability of the data. These data are collected into files each month and then released to the public. Figure 6 represents the ordering policy of data in Madrid.



SensorId	Timestamp	Values	Error
n	t_0	safdakjfslfjslfjs	**N**
n	t_1	sdlfsfsfljflsf	**N**
…	…	…	**…**
n	t_n	tryrtnslsl	**N**
n+1	t_0	jopuythvcdewd	**N**
n+1	t_1	uitiljfvjhtxkon	**N**
….	…	…	**…**
n+1	t_n	kjlhuiyb	**N**

Figure 6: Ordering of data in Traffic Data of Madrid

An initial analysis of the data indicates the possible insurgence of some of the following problems (not indicated with the Yes value in the Error flag):

- Missing data for a timestamp
- Missing data for a period of timestamp
- Interruption of the time series for a long period of time
- Repetition of the same data for a set of timestamps (this may indicate a sensor is out of order)
- Null value
- Others

The organization of traffic data in Dublin is different. The ordering is for inverse time (i.e., LiFo, the most recent data goes first in the list), then Region, Site, and finally Detector (i.e., sensorid). Figure 7 shows the organization of a CSV file.

Incremental for Region, Site and Detector

End_Time	Region	Site	Detector	Sum_Volume	Avg_Volume	Weighted_Avg	Weighted_Var	Weighted_Std_Dev
20220731120000	SCITY	367	23	0	0			
20220731120000	SCITY	367	24	0	0			
20220731120000	SCITY	368	1	124	10			
20220731120000	SCITY	368	2	324	27			
20220731120000	SCITY	368	3	116	9			
20220731120000	SCITY	368	4	76	6			
20220731120000	SCITY	368	7	0	0			
20220731120000	SCITY	368	8	0	0			
20220731120000	SCITY	368	9	0	0			
20220731120000	SCITY	368	10	0	0			
20220731120000	SCITY	368	11	0	0			
20220731120000	SCITY	368	12	0	0			
20220731120000	SCITY	368	13	0	0			
20220731120000	SCITY	368	14	0	0			
20220731120000	SCITY	368	15	0	0			
20220731120000	SCITY	368	16	0	0			
20220731120000	SCITY	368	17	0	0			
20220731120000	SCITY	368	18	0	0			
20220731120000	SCITY	368	19	0	0			
20220731120000	SCITY	368	20	0	0			
20220731120000	SCITY	368	21	0	0			
20220731120000	SCITY	368	22	0	0			
20220731120000	SCITY	368	23	0	0			
20220731120000	SCITY	368	24	0	0			
20220731120000	SCITY	369	0	0	0			
20220731120000	SCITY	369	1	121	10			

Incremental order per TimeStamp

Sorted by:
- Timestamp (LiFo)
- Region and Site
- Detector (i.e., Sensor Id)

Figure 7: Organization and ordering of the Dublin Traffic data

These datasets refer to the dynamic values measured by the sensing infrastructure. Other files describe the Sensors in terms of their location and some of their characteristics. These are separated files in order to make the CSV files lighters in terms of bytes and readability.

Pollution

For the pollution data, this section refers to the description of data as collected from Madrid. The city counts with over twenty meteorological and pollution stations. Figure 8 shows the location of some of them.

Figure 8: meteo and pollution stations in Madrid

The stations are located in different parts of the city and they cover areas with heterogeneous characteristics (e.g., parks vs much-trafficked locations). This gives a great variety to the data collected. For pollution data, there are two types of data: hourly and daily data. The data made available in CSV files have the format represented in Figure 9.

Station Location | Type of Pollutant | Code for Location + Type of pollutant + type of measurement | Period | Hourly Value | Validation Check

PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO_MUESTREO	ANO	MES	DIA	H01	V01	H02	V02
28	79	4	1	28079004_1_38	2019	1	1	23	V	17	V

Daily Value | Validation Check

PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO_MUESTREO	ANO	MES	D01	V01	D02	V02
28	79	4	1	28079004_1_38	2019	1	18	V	20	V

Figure 9: Format of CSV pollution data

They indicate the pollution station with a code representing the region, the borough and the number of the station ("*provincia+municipio+estación*"). Then the type of pollutant ("*magnitud*"). "*Punto muestreo"* gathers the previous information (location + type of pollutant) plus the type of measurement (the technique for the measurement used to take the value). "*Punto Muestreo"* is a short representation for understanding, where and what has been measured, as well as a unique identifier for that specific pollutant sensor (i.e. the sensor within the station). Then there is the period of measurement. Since there are two types of records, those with hourly data and those with daily data, there are two slightly different formats: for hourly measurements, the period indicates the day ("*DIA"*) and for each of the 24 hours a value followed by a verification check. For a measurement of 24 hours, there will be H01 up to H24 values for that type of pollutant. For daily measurement for each day, there will be a value followed by a validation check.

The pollutants considered and measured are represented in the following Figure 10:

Magnitud		Abreviatura o fórmula	Unidad medida	Técnica de medida	
01	Dióxido de Azufre	SO_2	µg/m³	38	Fluorescencia ultravioleta
06	Monóxido de Carbono	CO	mg/m³	48	Absorción infrarroja
07	Monóxido de Nitrógeno	NO	µg/m³	08	Quimioluminiscencia
08	Dióxido de Nitrógeno	NO_2	µg/m³	08	Id.
09	Partículas < 2.5 µm	PM2.5	µg/m³	47	Microbalanza
10	Partículas < 10 µm	PM10	µg/m³	47	Id.
12	Óxidos de Nitrógeno	NOx	µg/m³	08	Quimioluminiscencia
14	Ozono	O_3	µg/m³	08	Absorción ultravioleta
20	Tolueno	TOL	µg/m³	59	Cromatografía de gases
30	Benceno	BEN	µg/m³	59	Id.
35	Etilbenceno	EBE	µg/m³	59	Id.
37	Metaxileno	MXY	µg/m³	59	Id.
38	Paraxileno	PXY	µg/m³	59	Id.
39	Ortoxileno	OXY	µg/m³	59	Id.
42	Hidrocarburos totales (hexano)	TCH	mg/m³	02	Ionización de llama
43	Metano	CH4	mg/m³	02	Id.
44	Hidrocarburos no metánicos (hexano)	NMHC	mg/m³	02	Id.

Figure 10: Types of pollutants measured in Madrid

Not all stations will measure the entire set of pollutants. Each station measures a subset of the listed pollutants.

2.2.2 Data collection mechanism

Madrid and Dublin

The Madrid and Dublin files are stored in the open data repository of the city organization. Some of this data can be also collected by API invocation when specific data are needed. The effort made with these large data sets is to collect them, provide a homogeneous format and curate the data in such a way that applications can access them with a well-defined and formed structure.

These raw data together with the associated information about the sensors will be downloaded from a site of the SALTED project and manipulated in order to produce NGSI-LD conformant description to be later injected into a FIWARE Context Broker. For the time being, due to the need to keep up with the changing definition of the data format and the variations of the original data, there is no possibility from a user perspective to request the downloading of a specific file (referring to a period of interest of the user) and to automatically start the pre-processing and curation of the dataset. Users will, instead, be able to request and access the available data by means of the Data Broker API.

Santander

As for Open Data portals, all the data is available through an API, which can be called with a RESTful request for the desired dataset. This way, the functionality of the injection chains that work with this data is very similar to those collecting data from Madrid and Dublin.

SmartSantander data is, as previously mentioned, available through a subscription API. A call-back endpoint must be specified so that all measurements generated by the SmartSantander sensors are forwarded in an asynchronous way. This requires a different methodology for the treatment of data, since measurements will arrive individually the moment they are generated, and there is no way to know when this will happen. This means that the injection chain, as a consumer of the raw data coming from the SmartSantander IoT Platform, will have to behave as the REST server that is always ready to receive new data and treat it with the corresponding techniques to be able to inject it as NGSI-LD data into the context broker.

The Santander TTN data is somewhat similar in nature, but the way of accessing the individual measurements is slightly different. In this case, the injection chain acts as an MQTT client, which is subscribed to the parking sensors and receives a notification every time there is an update. After this client receives the data, it is forwarded to the mapping and curation modules and finally injected into the context broker.

2.3 EMPLOYED DATA MODELS

2.3.1 TrafficFlowObserved

For representing the traffic information, the choice is to use the Traffic FlowObserved data model because it is similar to the model used for the available data and it offers extensibility and flexibility to represent the data available accordingly to their timestamp and other characterizing features.

An example of it taken from the Fiware page is depicted in Figure 11


```
{
    "id": "TrafficFlowObserved-Valladolid-osm-60821110",
    "type": "TrafficFlowObserved",
    "laneId": 1,
    "address": {
        "streetAddress": "Avenida de Salamanca",
        "addressLocality": "Valladolid",
        "addressCountry": "ES"
    },
    "location": {
        "type": "LineString",
        "coordinates": [
            [-4.73735395519672, 41.6538181849672],
            [-4.73414858659993, 41.6600594193478],
            [-4.73447575302641, 41.659585195093]
        ]
    },
    "dateObserved": "2016-12-07T11:10:00/2016-12-07T11:15:00",
    "dateObservedFrom": "2016-12-07T11:10:00Z",
    "dateObservedTo": "2016-12-07T11:15:00Z",
    "averageHeadwayTime": 0.5,
    "intensity": 197,
    "occupancy": 0.76,
    "averageVehicleSpeed": 52.6,
    "averageVehicleLength": 9.87,
    "reversedLane": false,
    "laneDirection": "forward"
}
```


Figure 11: TrafficFlowObserved Data Model.

The mapping between the available data and the representation should be straightforward and should allow the flexibility to represent the time series in a very effective way giving the possibility to collect the record in different ways (data per 15 min, hourly data and the like depending on the needs. In addition, the model can encompass and integrate the different representations of the considered cities (e.g., Madrid and Dublin).

For the initial version of the formatting, the SALTED Project will focus on the location, timestamp, the sensor-id and intensity value.

2.3.2 AirQualityObserved

Smart Data Models defined within the FIWARE activities cover many aspects related to air quality. The model that fits better the available data is the AirQualityObserved model.

This model allows us to represent different pollutants and their measure and to associate them to a location and to a timestamp. This organization is illustrated by this snapshot taken from the AirQualityObserved specification (Figure 12):

```
{
  "id": "urn:ngsi-ld:AirQualityObserved:Madrid-AmbientObserved-28079004-2016-03-15T11:00:00",
  "type": "Feature",
  "geometry": {
    "type": "GeoProperty",
    "value": {
      "type": "Point",
      "coordinates": [
        -3.712247222222222,
        40.423852777777775
      ]
    }
  },
  "properties": {
    "type": "AirQualityObserved",
    "dateObserved": {
      "type": "Property",
      "value": "2016-03-15T11:00:00/2016-03-15T12:00:00"
    },
    "areaServed": {
      "type": "Property",
      "value": "Brooklands"
    },
    "airQualityLevel": {
      "type": "Property",
      "value": "moderate"
    },
    "CO": {
      "type": "Property",
      "value": 500,
      "unitCode": "GP"
    },
    "temperature": {
      "type": "Property",
      "value": 12.2
    },
    "NO": {
      "type": "Property",
      "value": 45,
      "unitCode": "GQ"
```


Figure 12: a snapshot of the AirQualityObserved FIWARE's data model

This data model is flexible and extensible in such a way to allow the formatting of the air quality CSV data described in the previous chapter into a well-formatted data set. The considered data will be hourly data and they will be organized per location, hour and type of pollutants.

For the individual pollutants encountered in the different data sets, SALTED is going to define a description of it conformant to the data model and capable of representing the hourly value, the station and location in which it was measured and a timestamp.

2.3.3 BikeHireDockingStation

This Smart Data Model[15] describes a docking station for bike hiring, and gives information about its current status. The more significant properties, or the ones we have used for Santander data, are briefly discussed below:

- **freeSlotNumber**: Number of free slots available at a specific moment.
- **availableBikeNumber**: Number of bikes available for hiring at a specific moment.
- **totalSlotNumber**: Total number of bike slots available at this station. The sum of the two previous numbers should be equal to or less than this one, considering the possibility that there may be out-of-order slots.

[15] https://github.com/smart-data models/dataModel.Transportation/tree/master/BikeHireDockingStation

- **status**: Current status of the station. Can be “empty”, “full”, “working” or any other application specific.

2.3.4 FleetVehicleStatus

This Smart Data Model[16] describes a generic fleet vehicle and is applicable to many IoT applications. We have used it to store data about the position and status of Santander urban buses. The more significant properties, or the ones we have used for Santander data, are briefly discussed below:

- **speed**: The current speed of the vehicle. Default unit: km/h.
- **currentStatus**: Current status of the vehicle. Can be “servicing”, “finished”, or several others described in the JSON schema.
- **lastKnownPosition**: GeoJSON reference to the position of the vehicle. Equivalent to “**location**”.

2.3.5 BatteryStatus

This Smart Data Model[17] represents the current status of a physical battery. This is very commonly used for monitoring the battery status of IoT sensors deployed in a Smart City, as is the case with SmartSantander. The most significant properties, or the ones we have used for Santander data, are briefly discussed below:

- **refBattery**: Reference to the Battery Smart Data Model if used.
- **statusPercent**: Current battery level of the device.

2.3.6 ElectroMagneticObserved

This Smart Data Model[18] is used for measurements of electric and magnetic fields. These can be taken at several frequencies, usually, the bands used for mobile communications. The most significant properties, or the ones we have used for Santander data, are briefly discussed below:

- **eMF**: Level of electromagnetic field observed. Units can be specified using the UN/CEFACT Common Codes.
- **Reliability**: Percent for confidence factor regarding the measurement represented.

2.3.7 ParkingSpot

This Smart Data Model[19] represents a well-delimited area where one vehicle can be parked. Most Smart Cities will have hundreds, if not thousands, of parking sensors deployed and Santander is no exception. The most significant properties, or the ones we have used for Santander data, are briefly discussed below:

[16] https://github.com/smart-data-models/dataModel.Transportation/tree/master/FleetVehicleStatus
[17] https://github.com/smart-data-models/dataModel.Battery/tree/master/BatteryStatus
[18] https://github.com/smart-data-models/dataModel.Environment/tree/master/ElectroMagneticObserved
[19] https://github.com/smart-data-models/dataModel.Parking/tree/master/ParkingSpot

- **location**: While the location is a property widely used in every Smart Data Model, here it is especially important since the accuracy of the coordinates represented must be high enough to point to a specific parking spot.
- **status**: Occupancy of the spot. Can be “closed”, “free”, “occupied” or “unknown”.
- **refParkingSite**: Reference to the parking site this spot belongs to, if applicable.

2.3.8 SoundPressureLevel

This Smart Data Model[20] describes a measurement of sound pressure in dB. The most significant properties, or the ones we have used for Santander data, are briefly discussed below:

- **sounddB**: The sound pressure level in dB.

2.3.9 Extensions to available definitions

Currently, there are no further extensions to the available definitions. We will update if any possible extensions are added later in the project.

2.3.10 Mapping of raw data to the NGSI-LD Model

The FIWARE Data Models are larger in scope and definition. This means that some fields defined in them are not necessarily being defined in simpler and more specific data models. In addition, certain information is either non-existent or they are collected in different files (e.g., the location and address of sensors). There is also the case that one or more data in the original file need some treatment in order to derive the data in the expected format of the NGSI-LD data model. This means that the mapping from existing data models to the FIWARE models is to be detailed and pursued with specialized ad hoc converted. Figure 13 shows a possible mapping between data collected in different files and the construction of an NGSI-LD based container.

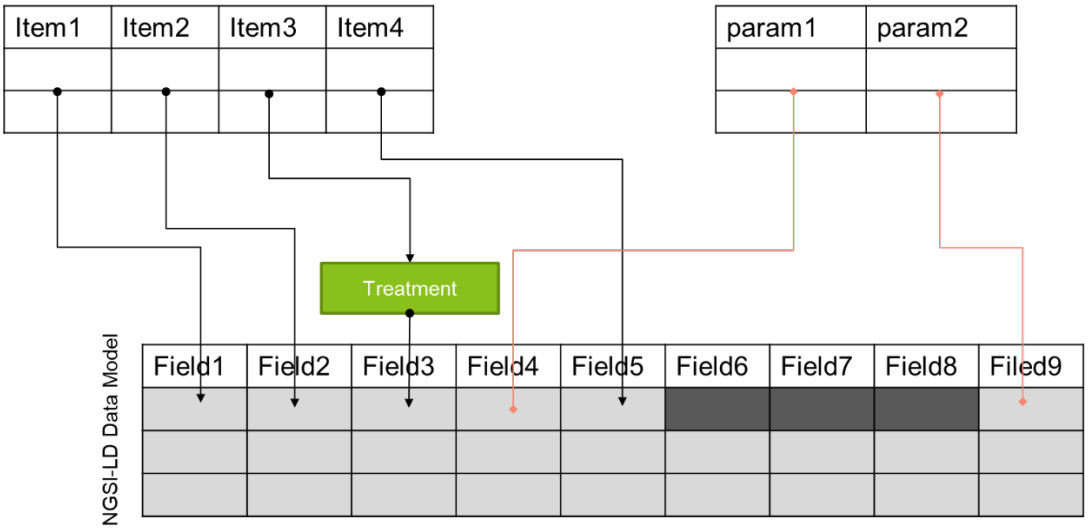


Figure 13: Mapping existing file structures into NGSI-LD one

Another relevant issue is to determine the range of values that are relevant for a possible linking of data pertaining to different data sets. For example, if the Madrid and Dublin traffic

[20] https://github.com/smart-data-models/dataModel.OCF/tree/master/SoundPressureLevel

data are to be harmonized, a set of issues will arise. Should the data be organized with the same timestamp? Moreover, in the positive case, what is the Timestamp period to utilize? Are the calculated measure comparable (i.e., the traffic intensity values)? A double choice is possible, to maintain the features of the original file and to tag the NGSI-LD description with notes related to these important characteristics or to harmonize the data by means of transformations. For this first organization of data sets, the former approach (do not touch has been considered). Figure 14 shows the mapping of two (or more) data sets.

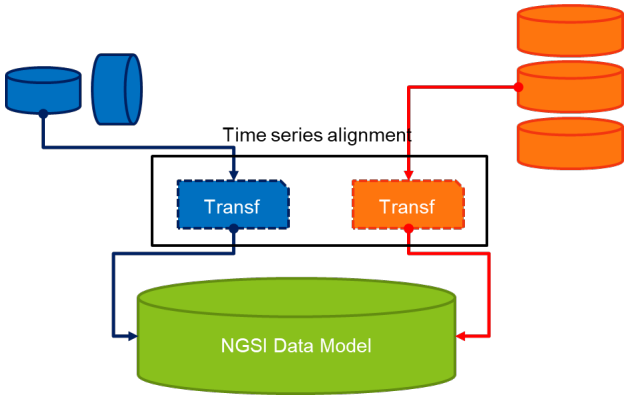


Figure 14: Mapping and harmonizing different data sets

There are two options with respect to pre-processing and curation in this scenario. One is to do the checking and adjustments of data directly on the raw original data, the second is to convert the data as soon as possible in an NGSI-LD compatible format and then do the adjustment and the curation. In the first case, an ad-hoc approach may have the advantage of specific transformations that will specifically tailor to the dataset under examination. The latter one, instead, loses a bit of specialization in favour of a generalized approach that can be applied to many more datasets and scenarios (if the data are in NGSI-LD format).

For the SALTED DET, we decided to use the second approach, and to map (also with some specific transformation) raw data to the well-formed NGSI-LD template to process, adjust and curate the data before the actual injection in the Context Broker. In this way, a data set organized accordingly to the data model is readily available and subsequent curations mechanisms are executed on it. The advantage is to define a set of operations on formalized data (supported by models) and not on specific data sets.

2.4 Data Curation

2.4.1 Features of the curated Smart City Data

The curation phase may take care of the following aspects of the dataset: the resolution of some problems related to the lack of data or their correctness in a short period; the broader issue of harmonizing the actual dataset to the rules and the expected quality of measurement in the large; and the evaluation of the goodness and usability of the formatted data in applications. The first aspect deals with the recovery of data in small periods of time and the interpolation of missing data in a short period of time. The second aspect will cope with the necessary sampling of data, the availability of data that can be semantically comparable to data of the same time of other data sets or previous data of the same source or the alignment to a system of measures. In other terms, the possibility of linking and comparing the transformed

data with other ones. The third aspect is related to an evaluation of the value and usability of the transformed data (it will be dealt with in the next section.

The punctual aspects of curation will cope with missing data or null data. The approach used is to identify the single missing data and by means of simple interpolation to calculate the possible value. At the same time, the datum will be tagged as “MI”, i.e., missing and interpolated or “NI”, i.e., Null and Interpolated. The alternative is to drop the missing record or to skip it. The sequence should be tagged as incomplete. This approach may be used with very sensitive data.

If missing or null data are more than one, an evaluation of the number of missing or null data will be done and in case of a small number of compromised data, a simple interpolation will be carried out, otherwise, some other interpolation strategies (e.g., KNN or others considering also past data of the same or sibling sensors) may take place. This method should be carefully applied in order to maintain a high quality of the dataset. In the case of Traffic Data, the strategy will be applied sensor per sensor in order to minimize the impact on the entire data set.

In case of a long interruption in the time series (e.g., more than 4 hours) then the sensor will be tagged as incomplete but an evaluation of the reliability of data will be expressed. A particular case is when the time series has many missing data in different periods (even a single missing or null value). In this case, an evaluation of the percentage of “holes” in the sequence will be calculated and the sensor for that period will be marked accordingly.

Repeated data and other errors. Looking for errors in this type of data is difficult and may involve knowledge of the particular situation in which the sensor is measuring. For instance, if the current value of traffic intensity is repeated for some time or if data are abnormal in a certain period of the day with respect to normal activities, there may be the case to warn the users by marking the data as Abnormal. The identification and resolution of these errors depend on how the measures are taken and the context in which the sensor is operating. Deriving a generalized approach (without using Machine Learning) could be difficult and out of scope.

Some of these mistakes could also require analysis and linking with other data. For instance, if the city system is providing an accident alarm, warning about road closure or working, then this information could be related to the abnormal one in order to correlate the issues. In addition, in this case, this additional treatment is very context-dependent and will not be considered in this phase.

2.4.2 Data quality dimensions metadata linking

The addition of specific metadata that is linked to the datasets in order to incorporate data quality assessment evaluation should be structured in a hierarchical way in order to evaluate and assess the most granular element (e.g., the sensor/detector) and to associate with it an evaluation related to a minimum period of time (e.g., one day). In this way, adjacent sensors can still provide good values and eventually be used by the applications. Groups of sensors may also be evaluated (for instance in the case of Dublin crossroad) in order to provide a good set of values even if some sensors are offering a low quality of data. A similar perspective could be adopted in other cities that have traffic sensors on different lanes (e.g., Madrid).

The ultimate goal of the metadata linking system (i.e. a kind of tagging of available pieces of data) should be one of allowing the final user to operate only on trusted data (e.g., the user may request only high-quality data) or by accepting the uncertainty of lower quality data. For large

elaborations, even low quality (and heavily interpolated data) may be used under the perspective that the quantity of available data will compensate for lower quality.

2.5 INGESTION PROCESS

2.5.1 Data streams collection

In order to inject stream data into the Scorpio Context Broker, we use the standard interface for the injection of individual entities. This is done with a POST request to the */entities* endpoint. The main drawback is that already existing entities, have to be updated instead of created, which requires a different request. However, we have written automation scripts for handling the creation/update process of individual entities.

2.5.2 Datasets collection

For the injection of datasets into the Scorpio Data Broker, the SALTED project will select a subset of available data and will locally transform and curate the data into an NGSI-LD complaint format. Differently from the real-time case, this process will be a one-time activity and it will result in a simplified (even if more time-consuming) set of requests to the Context Broker. The data will be stored in the most granular possible way, i.e., for each identifier (e.g., sensor_id) the data for each timestep will be created. In this way, the data can be afterwards aggregated accordingly to the needs of the requesting applications.

2.6 INTERACTION

The envisaged way for requesting the data is to make queries for collecting features and characteristics of sensors or stations. Then queries for specific values and attributes of the desired entity (sensor or station) in a location or in a time interval covered by the data availability will be possible.

2.7 ALIGNMENT TO THE SALTED ARCHITECTURE

These data will contribute to two major types of injection chain of the SALTED infrastructure. In fact, real-time flows, as well as large data sets, will be supported and two different processes will be implemented in order to format the available data in NGSI-LD, to curate them inject them into the Context Broker infrastructure. In addition, the raw data will be retained in order to allow the repetition of experiments and the improvement of the process. The specific contributions are related to:

- The identification of missing time series
- The interpolation of missing data accordingly to acceptable and practical functionalities
- The determination of the “quality” of the specific entity within a reasonable period of time (in the case of Madrid and Dublin, this could be on daily, weekly or monthly bases) and the corresponding tagging of the entity.
- The granularity of the data availability, in order to provide the possibility (in the phase of linking and enrichment) of creating different aggregations of the data.

These characteristics of the curation process are of great utility to developers and users of these kinds of data because they can focus their analysis on the right entities based on their attributes (like locations, and values) as well as on the quality of the data represented (e.g., low need for interpolation, verified data and the like).

3 THE SOCIAL MEDIA DOMAIN

3.1 CHARACTERIZATION OF DATA

3.1.1 Social Media data sources

Social media is a large network that is used to facilitate the sharing of ideas, thoughts, and information through the building of virtual networks and communities in the form of blogs, posts (tweets), likes, followers, clicks, shares, comments, or engagement rates. Communication between people and organizations has significantly changed because of social media, which includes blogs, discussion boards, and social networking websites [5].

Typically, using social media data is helpful to understand the audience and their need. Analyzing these data help content creators to provide better content for their audience in different forms on famous platforms. These analyses can also provide some indication of the level of brand awareness and customer satisfaction, which can help measure the effectiveness of marketing campaigns and social media strategies. Moreover, on one hand, it helps businesses to understand their competitors and on the other hand in academics, these data are considered very valuable since they can be used to answer a wide range of research questions from various disciplines.

Social media content mostly can be gathered from a varied range of social media platforms such as Facebook, Twitter, Instagram, LinkedIn, Snapchat, or TikTok in different forms including includes posts, likes, comments, shares, clicks, etc. As shown in Figure 15, the popularity of using different social media platforms is increasing significantly in recent years.

Here we take a dipper look at two main social media platforms' data used in research for data analytics, Twitter and Facebook.

Twitter

Launched in 2006, Twitter is a social network and microblogging platform, with currently 217 million active users in 2021[21]. Users may easily share and find real-time information on Twitter. It is the 4th most visited website worldwide[22].

Twitter data is information about what is happening in the world and what people are talking about right now. These data are highly valuable for every organization operating their marketing activities in the digital world nowadays.

Facebook

Facebook was first introduced in 2004 and has since grown to be one of the most widely used social networking sites. Among the world's most-used social platforms, Facebook comes on the top in 2022[23].

[21] https://www.omnicoreagency.com/twitter-statistics/
[22] https://www.semrush.com/website/top/
[23] https://datareportal.com/social-media-users

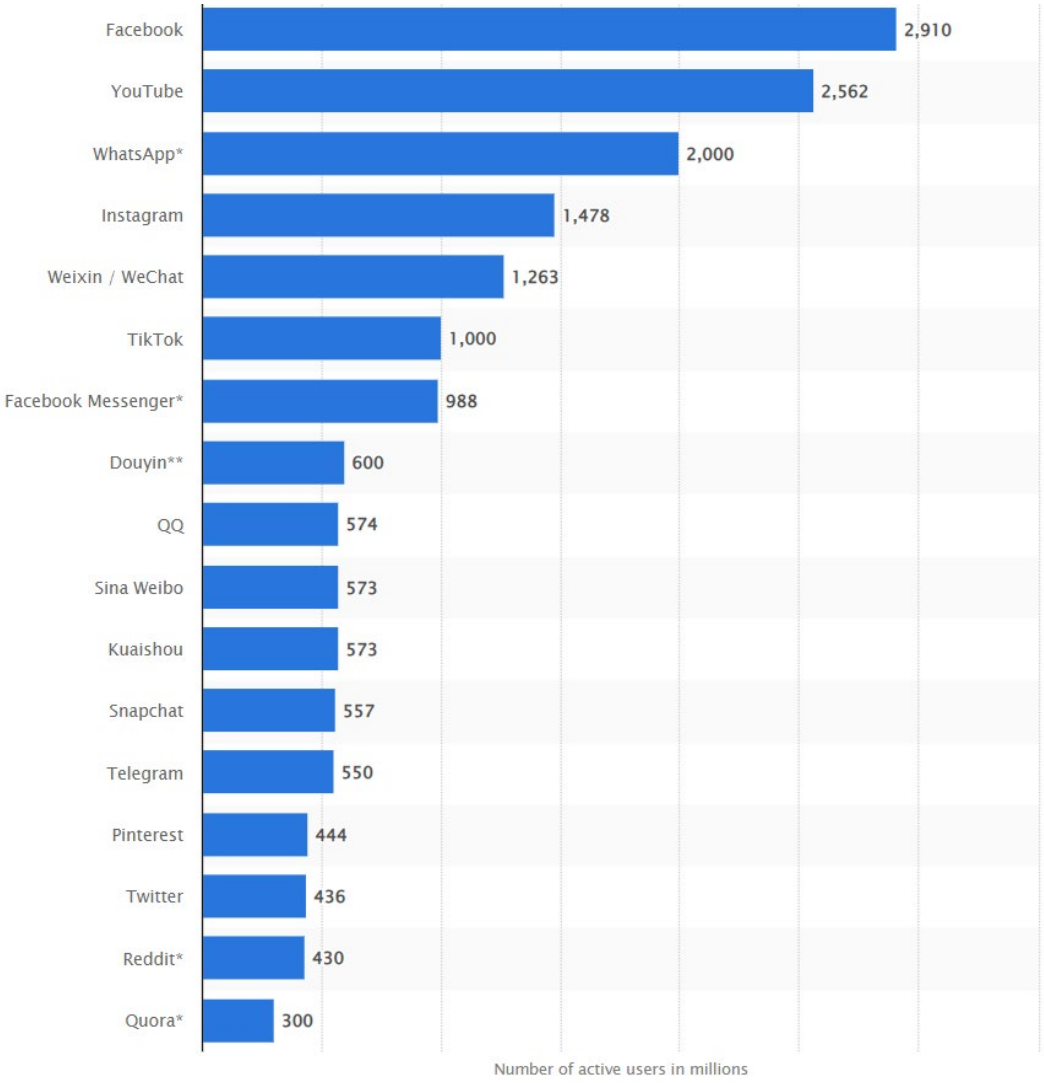


Figure 15: Global social media ranked by number of users[24]

3.1.2 Type of Social Media data

A variety of content in different forms can be found on social media. These data are generally in the format of text, images or videos. We can categorize these data as suggested[25]:

- User-generated content (UGC)
- Videos
- Polls and questions
- Contests and giveaways
- Statistics and data visualization
- Ephemeral content

[24] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[25] https://localiq.com/blog/types-of-social-media-content/

The data received from social media APIs are usually in JSON format containing different metadata such as profile, post, like and comment. Table 2 shows various data types of data retrieved from APIs.

Table 2: Social media data types based on received APIs output [6]

Type	Attribute	Type	Description
Profile Metadata	follower count	*numeric*	*audience size*
	followee count	*numeric*	*friend size*
	media count	*numeric*	*published posts*
	is verified	*boolean*	*verified by instagram*
	is private	*boolean*	*public or private*
	full name	*text*	*full name text*
	biography	*text*	*biography text*
	username	*text*	*account username*
	id	*numeric*	*unique id*
	profile pic url	*text*	*picture url*
	external url	*text*	*if exists*
	is business account	*boolean*	*if exists*
Post Metadata	caption	*text*	*post caption*
	date	*date*	*publish date*
	like count	*numeric*	*number of likes*
	comment count	*numeric*	*number of comments*
	shortcode	*numeric*	*unique id*
	hashtags	*text*	*list of hashtags*
	mentions	*text*	*list of caption mentions*
	is video	*boolean*	*video or image*
	video url	*text*	*if video == true*
	location	*numeric*	*location tag*
	tagged users	*text*	*tagged users in photo*
	thumbnail	*binary*	*content thumbnail*
	id	*numeric*	*unique post id*
Like Reaction	username	*text*	*username*
	id	*numeric*	*unique user id*
Comment Reaction	username	*text*	*username*
	id	*numeric*	*unique user id*
	date	*date*	*publish date*
	text	*text*	*comment text*

3.1.3 Limitations in data collection

Many social media platforms have set and updated their policies over time to maintain enough personal intimacy for their users, these access policies (Facebook since 2016, Twitter…) reduced the amount of information that may be collected and prevented access by the public to user timelines and personal profiles. In other words, researchers/developers are not permitted to gather information about specific users if they are not friends/related to them (Facebook…) or do not have their express consent. They can still gather information from public, likeable, and followable non-individual pages (like fan pages), which are public.

Even when having enough access to the data on social media platforms, consent to access personal information does not always imply that the information can be used at the collector's discretion. In fact, the ethical and legal ramifications of using the information collected must be taken into consideration, especially when we talk about platforms containing much personal and sensitive information. Therefore, the collectors should consider always the privacy laws and ethical guidelines set by the platforms' corporations[26].

The content might be limited due to the restriction set by the platform developer, For example, Twitter set a limited number of characters per post (since 2017, it went up to 280

[26] https://www.tandfonline.com/doi/abs/10.1080/01972243.2014.915276

characters[27]. This might be a bit challenging when aiming to explore the content itself (sentiment analysis…), or, when depending more on images to convey the concept of the information, Instagram, for example, might be a better choice.

3.1.4 Social Media data challenges

The nature of social media data can always be associated with many challenges for analysis. These challenges can have different dimensions:

Different language and unrelated text

For example, consider a scenario where we want to get data from Twitter about traffic in Spain based on specific hashtags. Firstly, since most users will their language to express their feeling and post their content, a related hashtag in the target language is needed. For instance, as shown in Figure 16 a post derived from twitter based on #Trafico indicates that although the comments and the hashtags are in Spanish, the main text of the post is in English, which is a part of a song lyrics. Only considering this text will not reveal any specific information about traffic, which might be far from users' intention to search for that hashtag. As a result, for a specific purpose, a combination of more hashtags is needed.

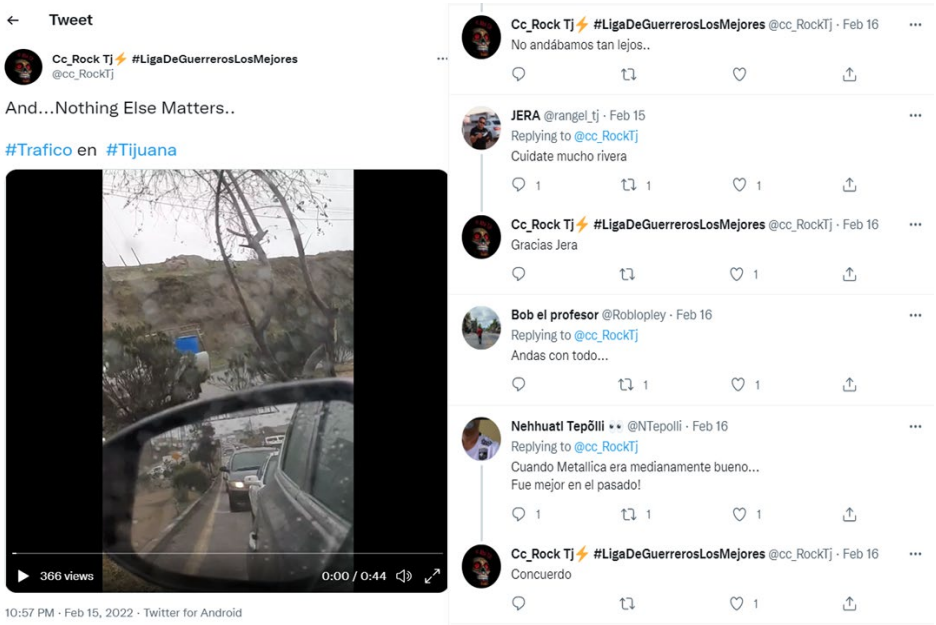


Figure 16: A post on twitter about traffic in Spanish language

Not expected data

Another challenge in social media data is getting unrelated data. As an example, assuming the intention of the user is to get the data of traffic in Bangkok, Figure 17 shows that even the combination of some hashtags to get the intended data might not be enough. Additionally, the comments are about the airplane in the photo, which again might be not well related to the

[27] https://developer.twitter.com/en/docs/counting-characters#:~:text=In%20most%20cases%2C%20the%20text,as%20more%20than%20one%20character

traffic. Moreover, considering different elements in the post together (such as video/picture, text, comments, etc.) would be more desirable to get the intended data.

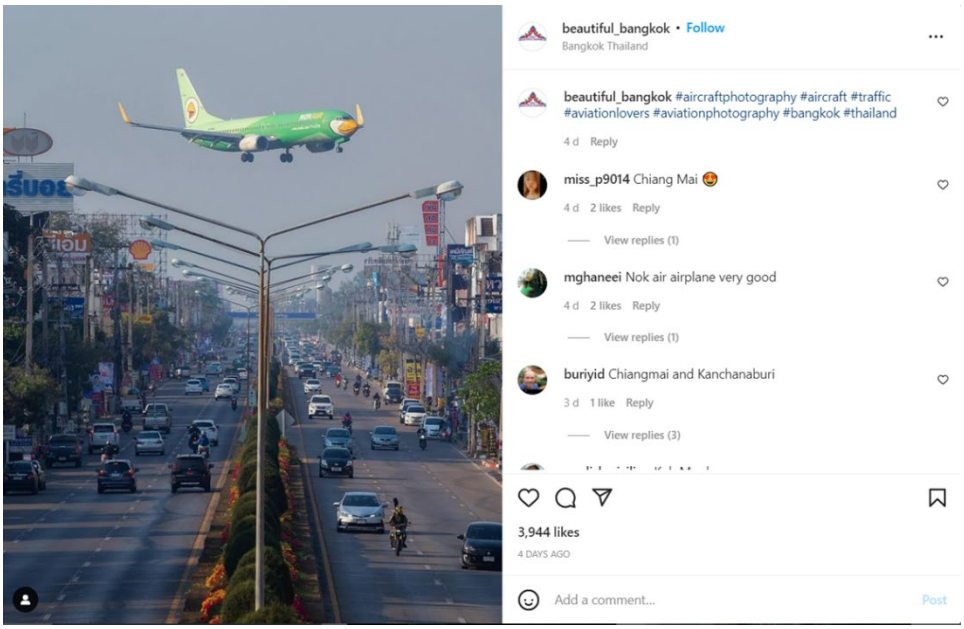


Figure 17: An Instagram post shows unrelated data

3.2 DATA COLLECTION MECHANISMS OF THE SOCIAL MEDIA

3.2.1 List of hashtags/ keywords/ users

Hashtags are a very helpful technique to interact with influential conversations as well as organize data. By definition, a hashtag is a keyword or phrase that is written within a post or remark and is followed by the hash symbol (#), highlighting it and making it easier to find it in searches. Social media networks utilize hashtags to index posts made on the same subject by many people. Although the hashtag was initially created for use on Twitter, numerous other social media platforms, such as Facebook and Instagram, have since adopted its use[28].

Firstly, this data includes the user's personal information. This specifies a user's username, account descriptions: the URL that leads to their profile page (homepage) and other crucial user data, such as names and the person's joining date for the network…

To establish personal connections (Follow and followers, adding friends…) and gain access to information, users can create personal accounts and post on their own behalf. Similar to this, business owners (or employees in public or investor relations) can set up accounts specifically for their businesses and post content on their behalf.

Based on this type of data, it is possible to determine user relationships within the social network of parties[29].

[28] https://gospeakeasy.com/2019/09/what-are-hashtags-and-how-to-use-them-on-social-media/

[29] https://www.researchgate.net/publication/321638167_Research_in_Social_Media_Data_Sources_and_Methodologies

One of the methods of crawling data from social media platforms is by the list of hashtags, keywords and users. In the Salted project, we use the official API of social media platforms the retrieve data. For example, using twitter's API, we get the data by the list of mentioned items to get the user timeline (i.e., the list of tweets posted by an account) and searching tweets (special hashtags or keywords).

For this project, APIs are developed to get the necessary parameters from users. These APIs will be integrated into the SALTED system for getting needed information from the application in order to extract related data from social media.

3.2.2 Continuous collection

The other method for crawling data from social media is continuous collection. Continuous data can include information about the text and graphic content of postings, comments on posts, likes and shares, links, photos and videos, how posts move across the social network, the timestamps of information exchanges… This enables us to examine issues including the latency in network post-migration, the social network architectures of knowledge and interest communities, and the detection of user mental and sentimental state[30].

For the continuous collection of Twitter's data, specific accounts, hashtags and keywords would be tracked to get the latest published content in the defined time span. About content uploaded, entities provide metadata and additional contextual information. Without needing to parse the text to extract that information, they give structured data from tweets, such as resolved URLs, media, 16 hashtags, and mentions. Entities may be found under the entities attribute in all Tweet Objects via the REST API and Streaming API endpoints[31].

3.3 Employed Data Models

3.3.1 The Social Media Model

For representing the social media information, the SocialMedia[32] data model is chosen since it offers extensible and flexible to represent the data available based on the characterization of social media data.

Table 3 and Table 4 show the Entity types and the properties of this proposed data model. The entity types available are:

- **SMAnalysis**: This entity contains a harmonized description of a generic SMAnalysis made for the Social Media domain. This entity is primarily associated with the process of analysis of Social Media applications' posts.

[30] https://www.tandfonline.com/doi/abs/10.1080/01972243.2014.915276

[31] https://www.researchgate.net/publication/321638167_Research_in_Social_Media_Data_Sources_and_Methodologies

[32] https://github.com/smart-data-models/dataModel.SocialMedia

- **SMCollection**: This entity contains a harmonized description of a generic SMCollection made for the Social Media domain. This entity is primarily associated with the process of collection of Social Media posts (primarily Twitter).
- **SMPost**: This entity contains a harmonized description of a generic SMPost made for the Social Media domain.
- **SMRefLocation**: This entity contains a harmonized description of a generic SM Reference Location (SMRefLocation) made for the Social Media domain.
- **SMUser**: This entity contains a harmonized description of a generic SMUser made for the Social Media domain. This entity is primarily associated with the description of a user of Social Media applications.

Table 3 Entity types of social media smart data model

Entity Types	User/Profile	associated with the description of a user of Social Media applications like Instagram/Twitter/Facebook etc.
	Post	associated with the description of a post (or a tweet in Twitter) created by a social media user
	Analysis	associated with the process of analysis of Social Media applications' posts
	Collection	associated with the process of collection of Social Media posts based on different subject (like a specific subject, hashtag, etc)
	RefLocation	associated with the description of a generic SM Reference Location

Table 4 Properties of social media smart data model

	Entity	Required	Other
Properties	User/Profile	id, platform, type, username	Name , dateCreated, bio, dateModified, etc.
	Post	postId, platform, type, postCreatedAt	URL, description, hasAnalysis, hasHashtags, hasImages, hasInteractionCount, hasLanguage, hasMentions, hasText, hasVideos, location, etc.
	Collection	collectionId, hasPost, description, type	dateCreated, description, hasAnalysis, hasPosts, etc.
	Analysis	AnalysisID, description, analyzedAt, type	dateCreated, description, hasAnalysisValue, hasAnalysisType, isAnalysisOf, etc.
	Location	locationID, location, type	dateCreated, description, locationReferencedBy, location

There are several relationships between defined entity types in this data model. A detailed description of these relationships is shown in Table 5.

Table 5 Relationships of social media smart data model

	Entity	In relation with	Entity	Description
Relationships	User/Profile	CreatedPosts	Post	Each user can create several posts
		isMentionedBy	Post	Each user might be mentioned in a post
	Post	CreatedBy	User/Profile	Each post is created by a user
		hasMentions	User/Profile	Each post can mention several users
		hasRefrencedLocation	Location	Each post might have a reference location
		hasAnalysis	Analysis	Each post might be analyzed by different services
		belongToCollection	Collection	Each post might belong to different collections
	Collection	hasPosts	Post	Each collection might have several posts
		hasAnalysis	Analysis	Each collection might be analyzed by different services
	Analysis	isAnalysedof	Collection	Each analysis service might be used by several collection
	Location	locationRefrencedBy	Post	Each location can be referenced by several posts

3.3.2 Extensions to available definitions

Currently, there are no further extensions to the available definitions. We will update if any possible extensions are added later on in the project but Figure 16 reveals a better perspective of all entities and relationships of the proposed data model.

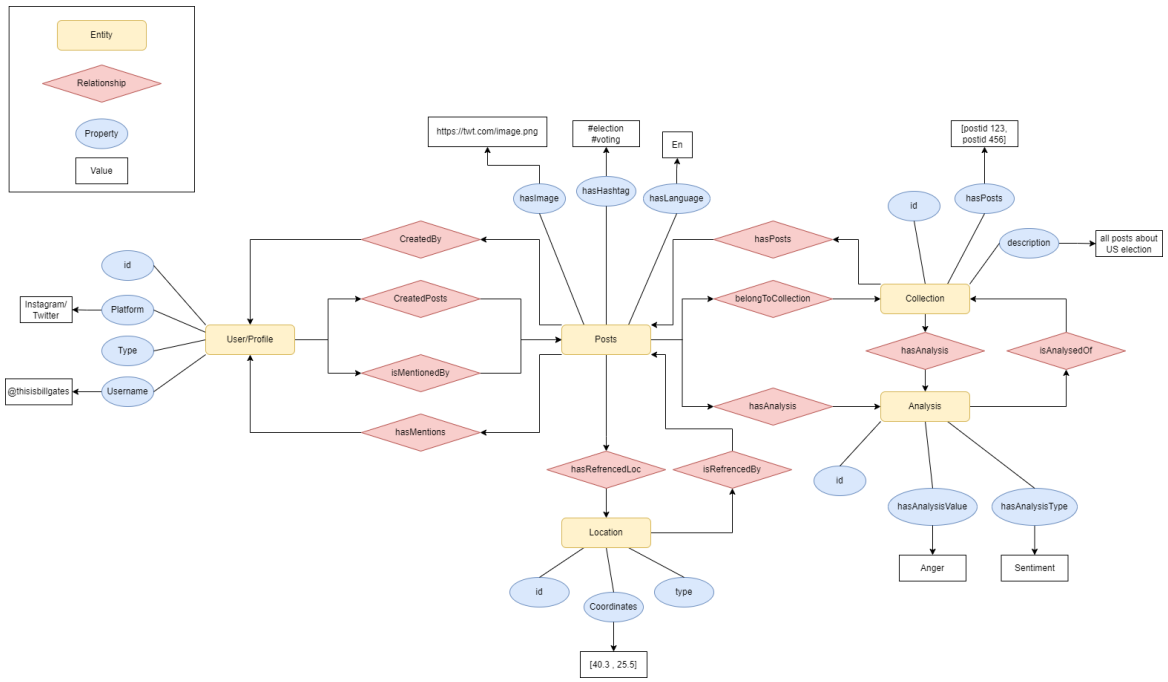


Figure 16: Social media smart data model overview

3.3.3 Mapping of raw data to the NGSI-LD Model

The data received from official social media APIs usually contain a huge amount of information. Not all of the information is needed to be modelled based on the proposed smart

data model. So extracting the necessary information from the output JSON file is an important step for mapping the raw data to the NGSI-LD model. As seen in Figure , a part of the JSON output of Twitter API contains a variety of information which will not be used in the NGSI-LD model. In fact, the only information that can be assigned to the current data model will be used.

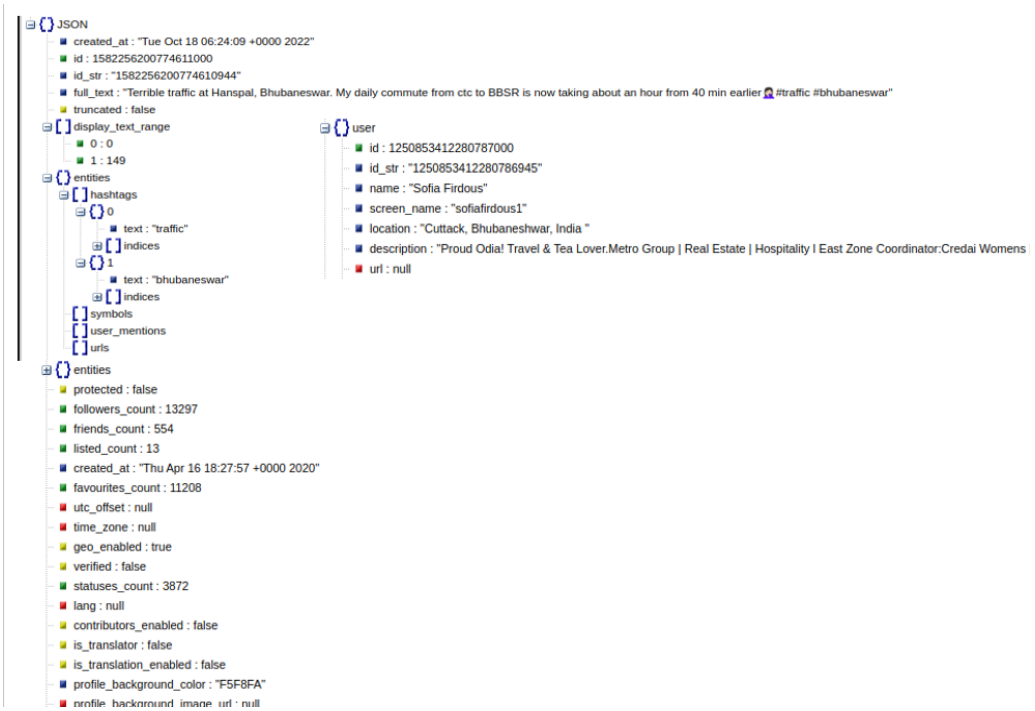


Figure 17: A part of twitter API output example in JSON format (raw data)

3.4 DATA CURATION

The data curation of social media data can be divided into separate steps inside the injection chain. The first step is pre-processing of the extracted data which is described in detail in Deliverable 2.1. This pre-processing step will be applied to the data after extracting the raw data from social media in order to clean the noisy data before converting it to NGSI-LD format. This step contains the following tasks:

- Transforming emojis and emoticons
- Removing URLs, emails, mentions, duplicate whitespaces, and punctuation from text
- Substitution of contractions
- Spell correction

The next data curation step is about checking and removing sensitive data before pushing it to the broker. Since social media data can contain some sensitive information which can be later used for malicious objectives, this step is considered a very important step before publishing the data. To prevent this problem, after converting data to the NGSI-LD format, we remove the username of the data Publisher (for example the username of the person who published that certain tweet). Moreover, more restrictions might apply to the data before publishing.

3.5 INGESTION PROCESS

For the ingestion of social media data into the satellite Scorpio broker, after converting the data to the standard NGSI-LD format, we use the standard interface for the injection of converted data to the local satellite broker.

The service developed for the ingestion process; always checks the same entities and the attribute's value to avoid duplication of same entities.

3.6 INTERACTION

The social media injection chain will be connected to the additional interface acting as an MQTT client in order to get the requested data from users to start crawling from different social media platforms based on the needs of external apps.

As explained in Section 3.2, to get the social media data, the developed API will get the necessary parameters for collecting the related data from different platforms. Therefore, the interaction with the interface acting as an MQTT client is essential in this loop. At this stage, the structure needed for the injection chain is designed and implemented and the connection of the injection chain with the control broker is under deployment.

3.7 ALIGNMENT TO THE SALTED ARCHITECTURE

Social media data as one of the sources of data with a different structure compared to IoT data is a valuable source for SALTED. The social media injection chain includes crawling data from Twitter as a good platform for extracting real-time information from what is happening in the world in a different format (such as text, image, etc.) and planning to integrate more platforms in future. After that, the raw data will be stored in the local databases. Moreover, the injection chain includes pre-processing of the data, converting to NGSI-LD architecture and pushing the entities into the Scorpio Context Broker. Additionally, the enrichment services presented in Deliverable 2.1 would be integrated as a service to this injection chain. The contribution of this data to the SALTED project can be seen in Figure 18.

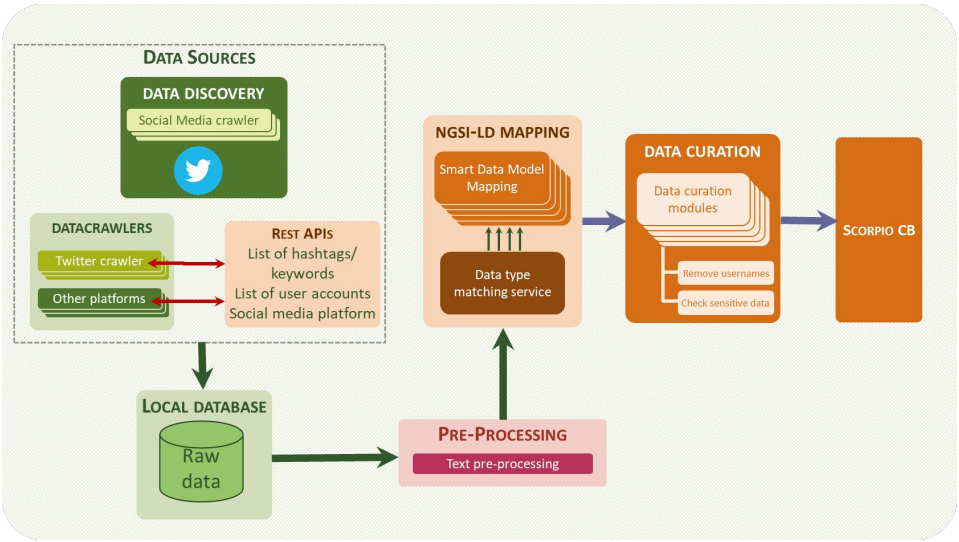


Figure 18: Social media injection chain overview

4 HARVESTING AVAILABLE WEB-STORED DATA (SEMISTRUCTURED AND GEO-REFERENCED)

4.1 CHARACTERIZATION OF DATA

4.1.1 Web as Data Source

The public web is one of the main sources of data, along with social media and IoT. Its key characteristics are that it is accessible to the majority and equal for all[33]. Over the years, beginning with the first ever-published website in 1991, the increase of websites was exponential, resulting in more and more data used for any conceivable purpose. The following graphic underlines this movement with numbers:

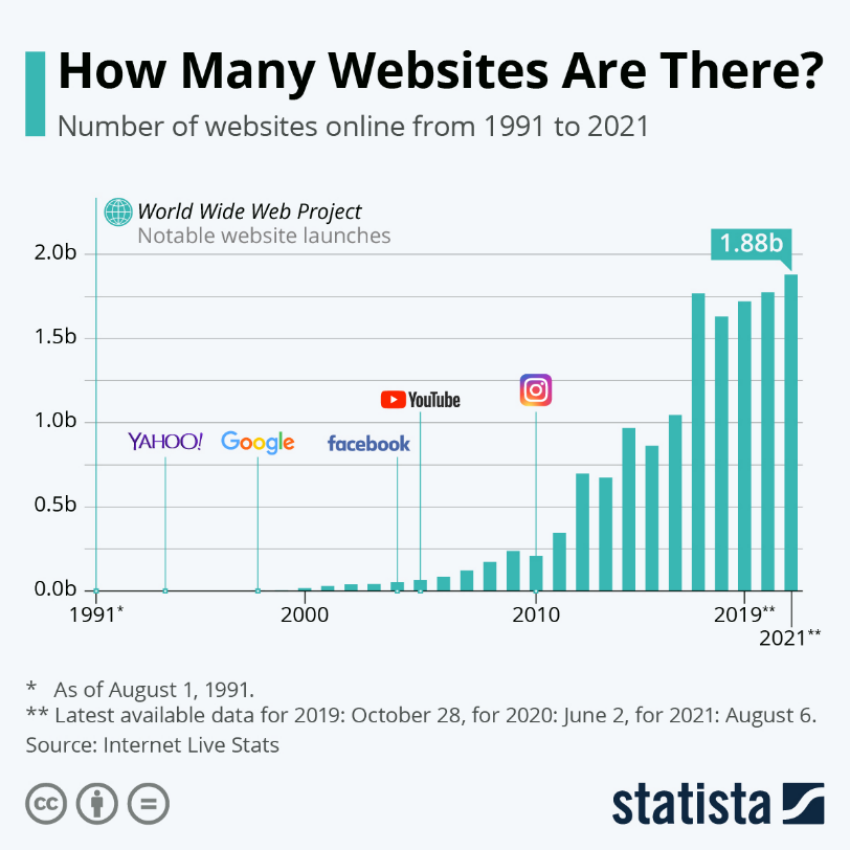


Figure 19: Number of Websites from 1991 to 2021[34]

[33] Detailed information on worldwide internet use and access can be found in a compacted form and visualized here: https://ourworldindata.org/internet#:~:text=Globally%2C%20the%20number%20of%20Internet,online%20for%20the%20first%20time.

[34] https://www.statista.com/chart/19058/number-of-websites-online/

4.1.2 Mechanisms for data discovery on the web

The web offers a multitude of different data sources in itself. The initial challenge is to identify useful data sources on the web in the first place. Useful means in this sense that it is possible to generate information or knowledge from the data[35] and that this information or knowledge also fits the use case. This is in contrast to the use of sensor data, where the data source is provided in the beginning and you can start processing it directly. The following gives a short overview of available data sources, that we already used within the project[36]. Each section also tries to underline what exemplary information can be extracted from those data sources, to give the incentive followed in the project. In the next section of this paper, the process of actually extracting information is explained in more detail. This list does not claim to be complete and will probably be extended in the course of the project:

There are commonly known **websites** e.g., homepages of companies or businesses or public administration sites. This is the most known data source on the web and the immense scope has already been shown in the above graphic, which is representative for the entire web. Those hold mostly small-scale owner/publicist-specific or sometimes topic-specific information.

The web also serves freely accessible interfaces to so-called **open data portals**. One known example, which this project also contributes to, is the European Open Data Portal, holding any type of information held by the Commission and other EU institutions and bodies[37]. However, there are also local interest groups that maintain data in a bundled representation and make it available to the public via the web. One example would be the data portal of the Rhine Neckar Metropolitan Region. It is part of the regional infrastructure and provides, among others, information about demographic, traffic and construction activities[38].

Free web encyclopedias, like Wikipedia, offer information about almost any topic. Generated, completed, deleted or edited in any way it is an open project of the general public, which offers a great amount of data, but with unsure quality standards[39].

The web also offers **geo information** in multiple ways and across multiple sources. OpenStreetMap is one example of free accessible geo information. Like Wikipedia, it is crowdsourced, which leads to highly detailed data, but again with unsure quality standards.

In the present project, usage of all the above-mentioned data would be achievable to state the different possibilities of web data and serve proof of concept studies of usage for each scenario. Additionally, this project also tries to integrate many data sources into the implementation of every single end-user application to showcase the possibilities of web data if used in a connected way, which can be visualized through the following graphic:

---

[35] See data pyramid approach: https://en.wikipedia.org/wiki/DIKW_pyramid

[36] But not only this web content data referenced in the following can be of interest. The web also offers the metadata alongside this content: web usage data and web structure data. Our main objective is to use content data and possibly, for better understanding, web structure data. Web usage data is more a topic of behavioural analysis and will not be focussed at this point.

[37] https://digital-strategy.ec.europa.eu/en/policies/open-data-portals

[38] https://digitale-mrn.de/projekte/kooperative-dateninfrastrukturen/datenportal

[39] https://www.wikipedia.org/

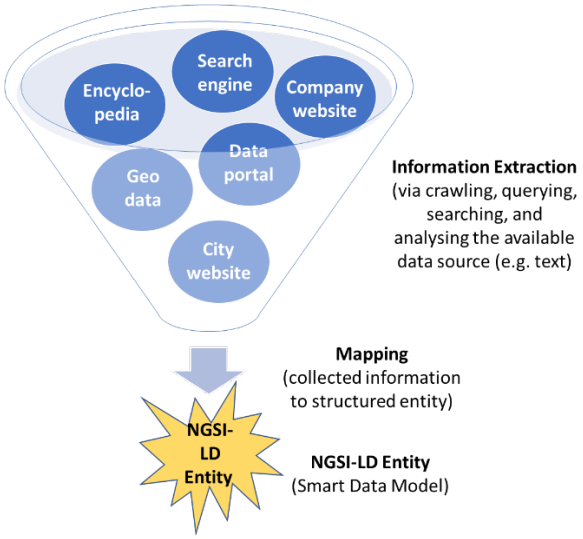


Figure 20: Integration of different data source for the creation of one NGSI-LD entity

The above section listed different data sources accessible through the world wide web. As already mentioned briefly before they differ mainly in the ways they are structured, their source (private/public administration) and their quality control. Since these points infer specific requirements for the handling of these data sources the following chapters explain the resulting handling mechanisms differentiated (as detailed as possible at the current state of the project).

4.1.3 Limitations in collecting the data

Web data is characterized by its enormous amount that is available. Users of this data must make an informed choice / a sound selection of the data that can actually benefit the desired use case. Many criteria play a decisive role here: accessibility, legal possibilities of use, quality and the necessary processing requirements to get/reach the data.

For the ultimate feasibility, however, the necessary time and the amount of data to be processed are also crucial, as this is determined by the available and usable infrastructure. Necessary processing requirements, same as storage possibilities need to be scaled at present capabilities at each site. Use cases on the other side need to find the most efficient and effective trade-off in each dimension.

4.2 WEB DATA CRAWLING MECHANISM

4.2.1 Top-Down approach

As laid out by the limitations of collecting web data, an informed choice with a use case in mind is necessary to master the amount of available data. To follow this guideline a top-down approach was chosen: First, a use case was specified, that would use web data. Afterwards the steps of developing processors that search, collect, and finally transform the chosen / relevant data into linked-data in NGSI-LD compatible format and publish through the Scorpio Linked Data could take place.

The use case decided upon was Agenda Analytics since it was the first validated Salted use case with potential users. The following paragraph will give a short introduction to the use case,

trying to make it understandable why certain decisions were made in the following processing steps.

Agenda Analytics has the matching of given agendas to the activities of companies and public administration as its content. The obtained matching scores can be used for detailed reporting and visualization, as can be seen in the next illustrations:

Figure 21: Agenda Analytics Report and Visualization

Agendas of interest could not only be the sustainable development goals[40] (SDGs) or environmental social governance risks in banks[41] (ESGs) but also the e. g. the public governance codex[42], addressing the emerging interest on governance topics in public companies. But for initial implementation, the SGDs were chosen.

The following sections give an introduction to the search and collection as necessary steps of the Salted injection chain. (The mapping to NGSI-LD and curation will follow in a dedicated section.) Within each of those steps, the differences in the handled web data regarding type and structure force a distinction in the procedure, which leads to the following separation according to these criteria. The explanations are based on the Agenda Analytics use case as an illustrative example. The next use cases in the project will use different data, but the following requirements depending on data types remain generally valid and will be applied and completed by them.

4.2.2 Search

Websites

For crawling (= searching and if applicable downloading) open websites, a crawler is used:

> "Web crawler is defined as a program or software which traverses the Web and downloads web documents in a methodical, automated manner."[43]

One visual example of what "traversing the web" means can be seen in the following graphic. Here a crawler was used with the starting point of the project's website https://salted-project.eu/, finding all referenced web documents/websites:

[40] https://sdgs.un.org/goals
[41] https://home.kpmg/xx/en/home/insights/2021/05/esg-risks-in-banks.html
[42] https://publicgovernance.de/html/de/Public-Corporate-Governance-Kodizes-und-Beteiligungsrichtlinien.htm
[43] Source: AbuKausar, Md.; S. Dhaka, V.; Kumar Singh, Sanjeev (2013): Web Crawler: A Review. In: IJCA. (https://research.ijcaonline.org/volume63/number2/pxc3885125.pdf)

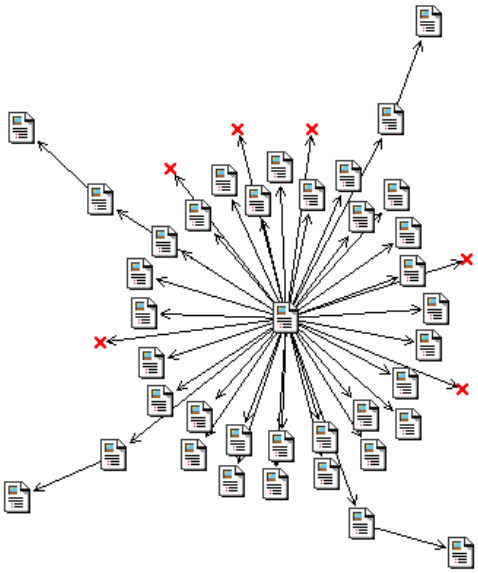


Figure 22: Crawlers traverse the www as directed acrylic graph[44]

Web crawlers can be divided into 3 categories, depending on the type of knowledge they crawl: general-purpose crawling, focused crawling and distributed crawling. General-purpose crawling collects as many websites as possible from a defined set of seed URLs and their referenced links. Even though this allows many websites to be collected from a wide variety of sources, this extensive/unrestricted approach results in reduced speed and slower network bandwidth. Focused crawling crawls only websites on a specific topic, using predefined specifications. This leads to less network bandwidth and hardware resources needed. Distributed crawling uses multiple processes are used to crawl and collect websites [7].

As mentioned above a more conservative approach is needed when facing unlimited data on the web and restrictions in network bandwidth, storage and processing. To handle this, meta-search engines are developed within the project, which uses an automated query process on already implemented search engines e. g. Google or Bing. They can deliver websites of interest within a definable scope and give that as input for the desired crawler. They allow using a more focused crawling approach within the project. An exemplary procedure in the Agenda Analytics use case is the search of main company websites via meta-search engines. Then, starting from the main page, a crawler can crawl all accessible websites of the same domain up to a defined depth, which meets criteria that fulfil the affiliation to sustainability communication.

In addition to the general crawling approach, at least one crawler implementation must also be selected for the project. There are a lot of frameworks in different programming languages. Commonly known examples are Scrapy[45] and Mechanical Soup[46] for python and Heritrix[47] and Stormcrawler[48] in Java. Within Scrapy and Stormcrawler projects will be evaluated due to their divergent strengths: Scrapy is easy to learn, fast and easy to implement. Stormcrawler is complex to understand/implement and maintain but offers the building of distributed, stable web

[44] Own simulation with https://www.cs.cmu.edu/~rcm/websphinx/
[45] https://scrapy.org/
[46] https://mechanicalsoup.readthedocs.io/en/stable/
[47] https://github.com/internetarchive/heritrix3
[48] http://stormcrawler.net/

crawlers. Their pros and cons are evaluated for the project's use cases before a decision takes place. A helpful framework that could be used for this evaluation can be found at https://nlp.stanford.edu/IR-book/html/htmledition/web-crawling-and-indexes-1.html . They provide a list of features a crawler must provide (politeness, robustness) and a list of features a crawler should provide, depending on the use case / the priorities in the project (distributed, scalable, performance and efficiency, quality, freshness, extensible).

After websites are found with the crawler, the search for the useful data within them takes place. Websites consist of unstructured text data. In addition to identifying useful websites, the challenge of extracting the needed parts is essential. Within Agenda Analytics, this is necessary when corporate public communications on sustainability need to be captured, ultimately to be matched against the SDGs reference corpora in order to make an assessment of efforts[49]. An example website of the Deutsche Bahn AG can be seen below.

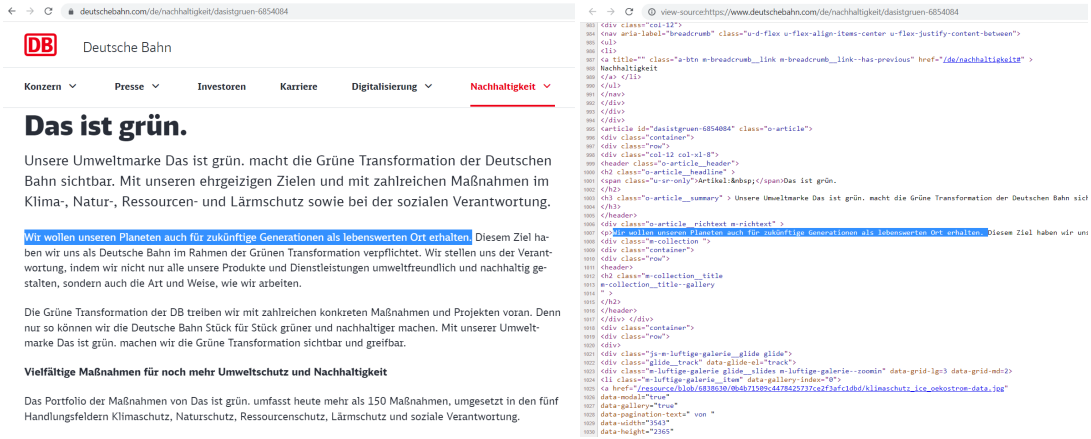


Figure 23: Company website in compiled and raw format

The raw content of the website contains HTML tags, CSS attributes and finally content that is not to be used, such as navigation bars or advertisements, in addition to the text data that is to be used later. Furthermore, it would be desirable not only to extract the text for agenda analytics but also to perform further NLP in it, e.g. in the context of named entity recognition or relationship identification between already identified entities.

Open Data Portals

Open Data Portals are designed to make it easier to find reusable information and datasets made available to the public. In some cases, the administrators of these portals offer access via application programming interfaces (APIS), which enables direct and automated access to software applications. Within the project, open data portals of interest are identified.

In the project, the given data basis is evaluated for each use case. If there are open data portals that contain relevant information, these will be analyzed. For the Rhine-Neckar metropolitan region, the data from the MRN association[50] and Geonet e.V. are particularly interesting, as they publish region-specific data in open data portals. Therefore, cooperation is

[49] An alternative approach is to use only PDF files identified on websites that can be identified as sustainability reports and convert them to text.
[50] https://contextbroker.digitale-mrn.de/entities

developed to ensure the best use and integration of this data. Up until now, they are not integrated into any use case.

Free web encyclopedias

Free encyclopedias like Wikipedia impress with their huge amounts of information on certain topics. In the project, they are used only sporadically due to the uncontrollable truth content, e.g., to provide sample data for prototype developments. An example would be the data demand for companies under state control in Germany, for which the evaluation by Agenda Analytics is particularly interesting. In order not to stop the development of the ingestion chain due to the complications in finding open data sets containing this data, existing data from Wikipedia[51] is used at the beginning, before coverage of the data to qualitatively higher sources can be covered. In this example an analysis of the raw website is needed again to extract data from the corresponding Wikipedia table:

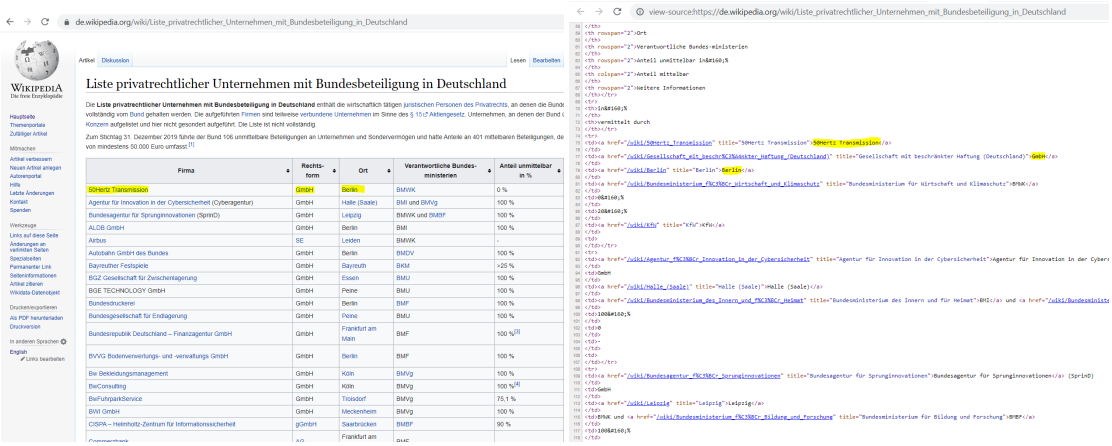


Figure 24: Company wikipedia table in compiled and raw format

Geoinformation

Geodata can be found in various places. Nowadays, federal states have often set up geo-data portals. An overview for Germany can be found at: https://de.digital-geography.com/open-data-deutschland-freie-geodaten-von-bund-und-landern/ .

Like the Open Data Portals, there is also project cooperation with e.g., units from the MRN region (Geonet e.V. https://ckan.geonet-mrn.de/), that supply structured geo data over data portals.

Geo data is also available through collaborative projects like OpenStreetMap[52]. Due to the widespread use of this data source, a large number of publications[53] deal with the quality assessment of the data and the evaluation of quality assurance measures of OpenStreetMap.

[51] Referenced Wikipedia content: https://de.wikipedia.org/wiki/Liste_privatrechtlicher_Unternehmen_mit_Bundesbeteiligung_in_Deutschland

[52] https://wiki.openstreetmap.org/wiki/About_OpenStreetMap

[53] Examples: https://www.researchgate.net/publication/220144096_Quality_Analysis_of_OpenStreetMap_Data_Based_on_Application_Needs, https://www.tandfonline.com/doi/pdf/10.1080/10095020.2016.1151213

An own representation of quality assurance can be found at https://wiki.openstreetmap.org/wiki/Quality_assurance. OpenStreetMap provides an Overpass API[54] for read-only access to the data. Within this project OpenStreetMap is used e.g. for enriching already defined entities with address information and further data provided, like web URLs, using the organization name and headquarters location. An example Overpass API Query is represented in the following python code used and exemplary result in the Overpass Frontend:

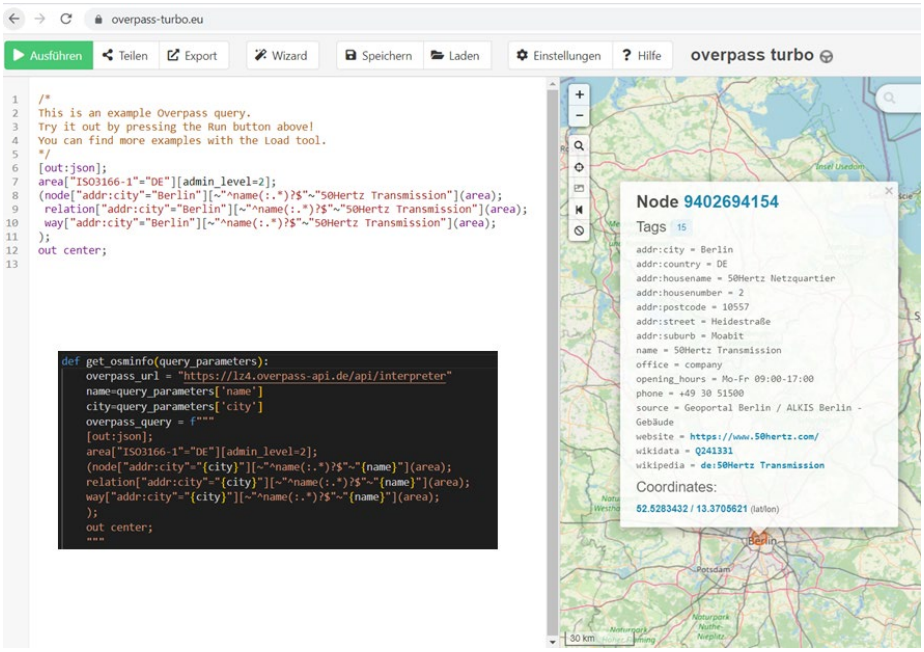


Figure 25: OpenStreetMap Overpass API Query

4.2.3 Collection

Regarding the collection from previously identified data sources, two perspectives play a decisive role:

In some cases, only extracted information from the web data is further processed or stored. Examples are address information from OpenStreetMap or company information from Wikipedia. This ensures an efficient approach without too much overhead.

In some cases, however, such as the crawling of websites, it can make sense to store the entire content of a website, among other things in order to maintain transparency with regard to the origin of the information. It may be necessary to demonstrate which data basis has led to which computations (e.g. Compliance Scores in Agenda Analytics). Furthermore, it can be useful to store information about the last crawling time in order to avoid unnecessarily short crawling intervals.

These different views are selected and implemented depending on the use case and data source. The exemplary collection implementations already implemented for the Agenda Analytics use case are bundled in a service called “DiscoverAndStore”. Chosen API framework is FastAPI. As a storage layer, PostgreSQL is used for companies found. The service endpoints handle the start of different services regarding the search for new organizations (using a

[54] https://wiki.openstreetmap.org/wiki/Overpass_API

Wikipedia table or OpenStreetMap information) or the enriching of the ones already in the database (getting website URLs through google or information about specific organizations within OpenStreetMap). Additionally, CRUD operations on the database records are possible. The services output is always a list of JSON objects, where each organization is represented as a JSON object. The following schema tries to visualize this sequential exemplary approach of search and collection of company data in the Agenda Analytics use case:

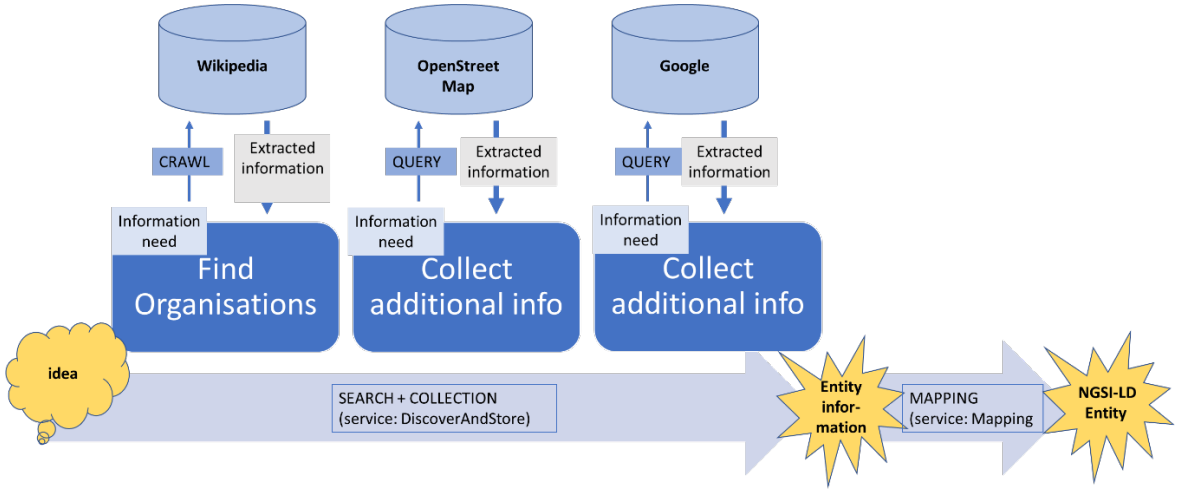


Figure 26: Exemplary flow of search, collection and processing of relevant organization data for Agenda Analytics

The implementation of the search and collection of company website data regarding sustainability will follow.

4.3 Employed Data Models

"Semantic annotation is regarded as the key step in the whole processing architecture, which means adding semantic notifications to pre-processed data. Generally, semantic annotation is composed of two steps, semantic modelling and instance annotation. Semantic modelling serves an important role, and users may define new or reuse existing semantic models depending on the situation. The pre-processed data would be instantiated based on predefined semantic models to finish the process of semantic annotation." [8]

In Salted a cross-domain ontology is provided by the NGSI-LD, which uses a semantic expression format based on DF/RDFS/OWL and partially on JSON-LD. Further Salted aims at using smart data models as common data models for specific entity types.

Within the short explanations on data collection for web data, the 2 views have already been pointed out. In the already realized implementations, information was extracted from web data and flowed into use case specific entities of the type Organization. For this purpose, a new smart data model was defined and contributed. Further, possibly necessary data models for the first use case Agenda Analytics, like PointOfInterest are already defined within the smart data models.

The developments for crawling, which could make it partially necessary to store whole websites, are still in the work. Here it could be useful to design data models for whole websites, which contain a link to the persistent storage under an attribute name like "content". The same

approach might be useful for the agendas used for comparison. This would ensure a representation of used data within the Scorpio.

4.3.1 Extensions to available definitions

An already finished contribution to the smart data models is the model for the entity type Organization, which can be found here: https://github.com/smart-data-models/dataModel.Organization [55]


```
{
  "id": "urn:ngsi-ld:Organization:8b599219-b9ae-49ba-a0de-5623aca71919",
  "type": "Organization",
  "dateCreated": {
    "type": "Property",
    "value": "2022-11-22T13:40:30.513763+00:00"
  },
  "dateModified": {
    "type": "Property",
    "value": "2022-11-22T13:40:30.514148+00:00"
  },
  "name": {
    "type": "Property",
    "value": "Stadtwerke Reutlingen"
  },
  "location": {
    "type": "GeoProperty",
    "value": {
      "type": "Point",
      "coordinates": [
        48.4938379,
        9.1878359
      ]
    }
  },
  "address": {
    "type": "Property",
    "value": {
      "addressLocality": "Reutlingen",
      "postalCode": "72762",
      "streetAddress": "Hauffstraße 89"
    }
  },
  "areaServed": {
    "type": "Property",
    "value": null
  },
  "url": {
    "type": "Property",
    "value": "www.stadtwerke-reutlingen.de"
  },
  "legalName": {
    "type": "Property",
    "value": "Stadtwerke Reutlingen "
  },
  "taxID": {
    "type": "Property",
    "value": null
  },
  "@context": [
    "https://raw.githubusercontent.com/smart-data-models/dataModel.Organization/master/context.jsonld",
    "https://smartdatamodels.org/context.jsonld"
  ]
}
```


Figure 27: Organization data model (example)

[55] A reference to the contributor project can be found here: https://github.com/smart-data-models/dataModel.Organization/blob/master/CONTRIBUTORS.yaml

Within this contribution process generated context file[56] for this data model can be used within the project to upload information about organizations to a broker, ensuring everyone else can understand and use the data as well.

As soon as the need arises for further data models for the web-generated data, the contribution process thus tested is run through again.

4.3.2 Mapping of raw data to the NGSI-LD Model

Regarding the technical approach of the mapping, firstly the usage of templates was targeted.

Since the services leveraged for each collection step provide clear results (e.g. through JSON formatted data), a mapping through templates is sufficient and no bias needs to be introduced through AI-based engines.

For mapping purposes, the source of "to be mapped" data is crucial.

So far, the project has implemented 2 mappings, regarding web data. Those are bundled within one service called "Mapping". Chosen API framework is FastAPI. The first service endpoint accepts data in form of the output of the "DiscoverAndStore" service, a list of JSON objects. The services output is a list of NGSI-LD Organization entities. The second service endpoint is a proof of concept of using data from the MRN Context Broker. Input is the data from the Context Broker API about BikeHireDockingStation and output is a list of NGSI-LD mapped BikeHireDockingStation entities. This enables the use of MRN data, which can be extended when needed, since the source representation is "almost" NGSI-LD with only the context missing, which can be added through the known used data models.

Each mapping uses unified resource identifiers generated by the python package UUID[57], but this can be adapted within the project.

4.4 Data Curation

The unstructured web data with unlimited scope imposes again additional challenges on the curation of data. Due to the large number of data sources that flow into e.g. a representation of an entity, the curation must already take place within all services of the injection chain, i.e. at the service level. Different approaches are currently discussed and tested.

One proposed exemplary workflow for the collection of the initial company data (legal name, address information and web URL) intended to store raw data in a PostgreSQL database. Additionally stored associated metadata (when was this company found and by what service? When was the information updated?) should support the information transparency and quality.

This approach is supposed to ensure the logging of how the data was discovered, for traceability and proof of origin.

[56] Accessible at the public address: https://raw.githubusercontent.com/smart-data-models/dataModel.Organization/master/context.jsonld

[57] https://docs.python.org/3/library/uuid.html

4.5 INGESTION PROCESS

For the ingestion of the mapped data into the Scorpio broker a service called “Publish” is developed. Chosen API framework is FastAPI. This transfers NGSI-LD formatted entities into the local Scorpio. The input to this service is a list of NGSI-LD entities. The output contains all entities / their information that is successfully represented in the broker afterwards.

The challenge here is the handling of possible exceptions and the automated adaptation of the procedure. If, for example, an entity ID already exists, then the existing entity and the entity to be added must be compared regarding their attributes and attribute values to be able to make overrides, deletions, or additions if necessary.

It is equally important to recognize existing entities that do not have the same ID but may have the same attribute values. These must be recognized as the same entity.

4.6 INTERACTION

The services developed use FastAPI as an API framework and are made deployable independently using docker-compose. The implementation via FastAPI allows a flexible extension of the endpoints of all services in order to successively include e.g. additional parameters in the service address options.

At the beginning of the project, the focus was on developing the content of the services. The above-mentioned flexible approach ensures that the content can be expanded at a later stage.

Subsequently, the implementation of access to the services was and still is the present focus. After the successful definition/concretization of the orchestration of all SALTED services, the interfaces to the Control Plane will be implemented for all FastAPI services as well as for all other SALTED services. Those interfaces make them reachable via MQTT (this allows them to communicate, read & write on topics). Alignment to the SALTED Architecture

On a theoretical level, the consideration/inclusion of web data brings another view of the often unstructured, undefined, and quantitatively indistinguishable/assessable data into the project. It challenges the architecture of the project in different ways than IoT data does, for example. Additionally, the services contribute by proposing a microservice development approach with FastAPI, that touches on the topics of deployment (Docker), logging, testing (Pytest) and observability (Grafana, Prometheus, Loki, Tempo). Created GitHub repos and documentation that can help in the further development of services.

At the practical implementation level, the following illustrations show the services already implemented and those planned for the first Agenda Analytics use case, visualizing the exemplary channelling[58] of services, that were mentioned in the sections before.

Managing these services on a VM with a dedicated partner Scorpio that federates with the master Scorpio allows the testing of flexible implementations of the injection chain within the framework/architecture concepts defined in the project.

[58] Channeling, because automatic orchestration takes not place yet. Fill follow.

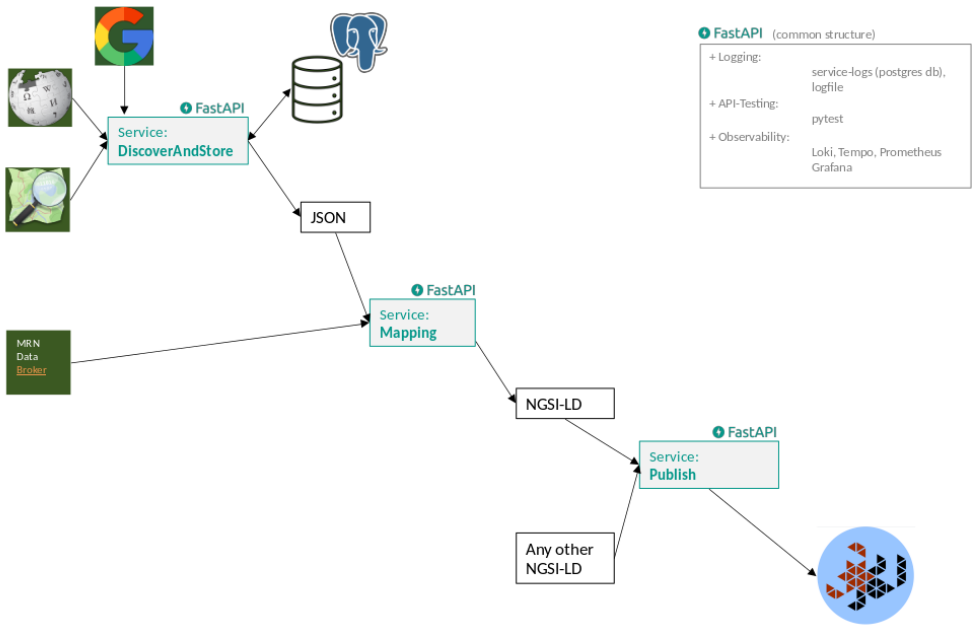


Figure 28: Implemented first injection chain (Agenda Analytics)

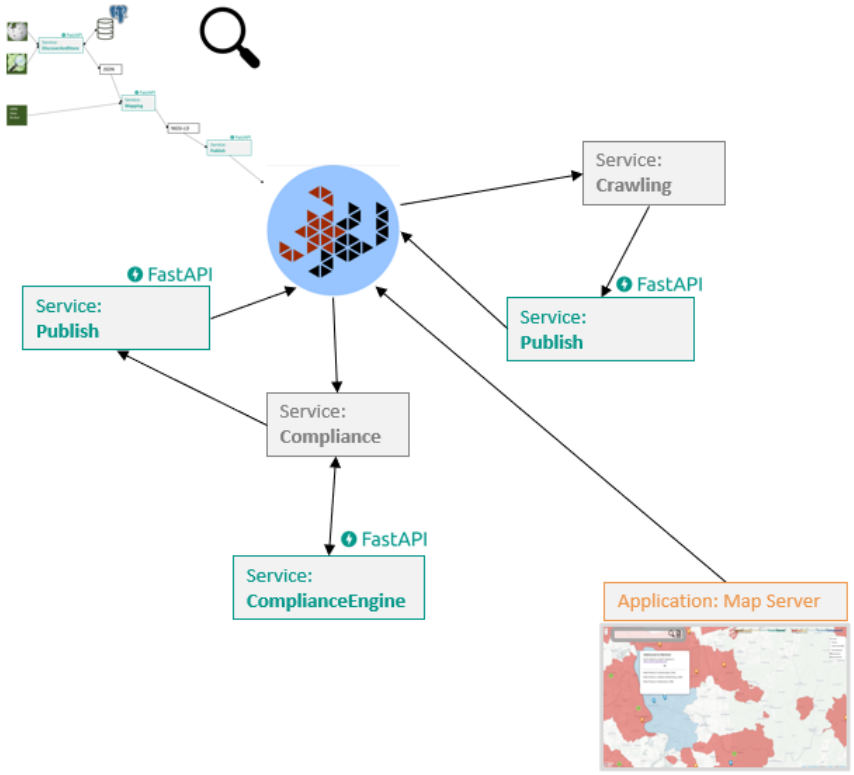


Figure 29: Planned enriching loops and applications (Agenda Analytics)

5 SOCIOECONOMIC STATISTICAL DATASETS

In the context of Smart Cities, Open Data is usually talked about in view of data related to the control and use of infrastructures (mobility, infrastructure, ...) and to the evaluation of environmental conditions (weather, air quality, ...).

What is often not considered is that there is another vital domain of open data - the area of open government.

"Open government is the governing doctrine which sustains that citizens have the right to access the documents and proceedings of the government to allow for effective public oversight."[59]

The goals to be promoted by Open Government include more transparency, more participation, more intensive cooperation between civil society and administration, innovation and a strengthening of community interests.

The publication of data that can be attributed to open government is not a fundamentally new idea. Historically, of course, administrative units at all levels have always collected statistical information in order to steer the development of society and the economy. Since the beginning of the modern era, the social sciences, especially economics, have supported this concern by providing relevant research data. Therefore, it makes sense to always consider research data in addition to open administrative data.

An important role in the provision of open government data is played by the statistical offices that exist in developed countries at all administrative levels - from municipalities and states to global institutions such as the UN.

In the field of Open Government, too, the aim is to make data available in a more timely manner, for shorter periods of time, in relation to smaller local units and in relation to current events, through more automated processes, in order to be able to use them as a basis for new control processes.

5.1 ENVISIONED APPLICATION: A SALTED BOT FOR ACCESSING SOCIO-ECONOMIC STATISTICS

This development coincides with SALTED's concern for "situation-awareness" - where "situation" includes the aspects of place, time and context.

It is therefore to be expected that SALTED, as a platform for the provision of open data, can also generate added value in the area of open government.

This is to be demonstrated within the framework of the SALTED project through the development of a "bot" that provides current socio-economic data based on the location of the user.

[59] https://en.wikipedia.org/wiki/Open_government

This demonstrator will be based on data published by the German Federal Statistical Office.

The class of data from the Open Government context was also chosen in order to meet the claim formulated in the project application:

“From the beginning, the project starts with a large portfolio of comprehensive datasets in Smart City and Smart Agriculture. The datasets are available as public data (e.g. from the EPD Portal) or from previous projects, e.g. sustained Smart City projects continuing after the lifetime of the respective R&D project. This includes […] socio-economic data about SALTED example cities like Santander, Paris or Heidelberg."

It should be mentioned that the technology of a "bot" chosen for the demonstrator is of course also applicable in relation to the mentioned data classes from the context of Smart Cities.

5.1.1 Draft of frontend and layout of statistics

The following illustration shows a communication flow that could ideally take place in a messaging environment:

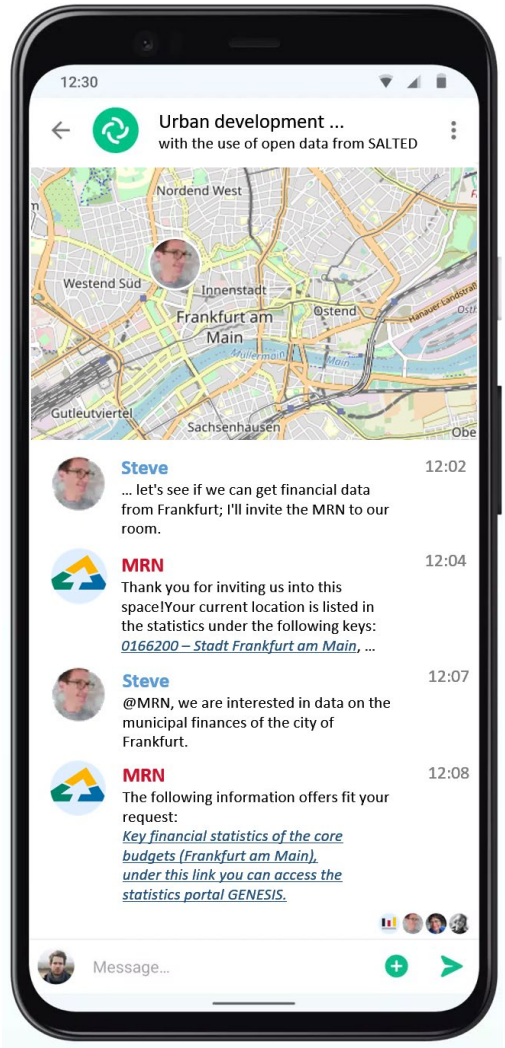


Figure 30: Exemplary GUI of Element Messenger App

This illustration shows the messaging app of the Matrix/Element platform; Kybeidos has already gained experience with the implementation of bots on this platform in previous projects.

The following illustration shows the layout of the statistical evaluation that will be displayed when the relevant link is activated in the messaging app.

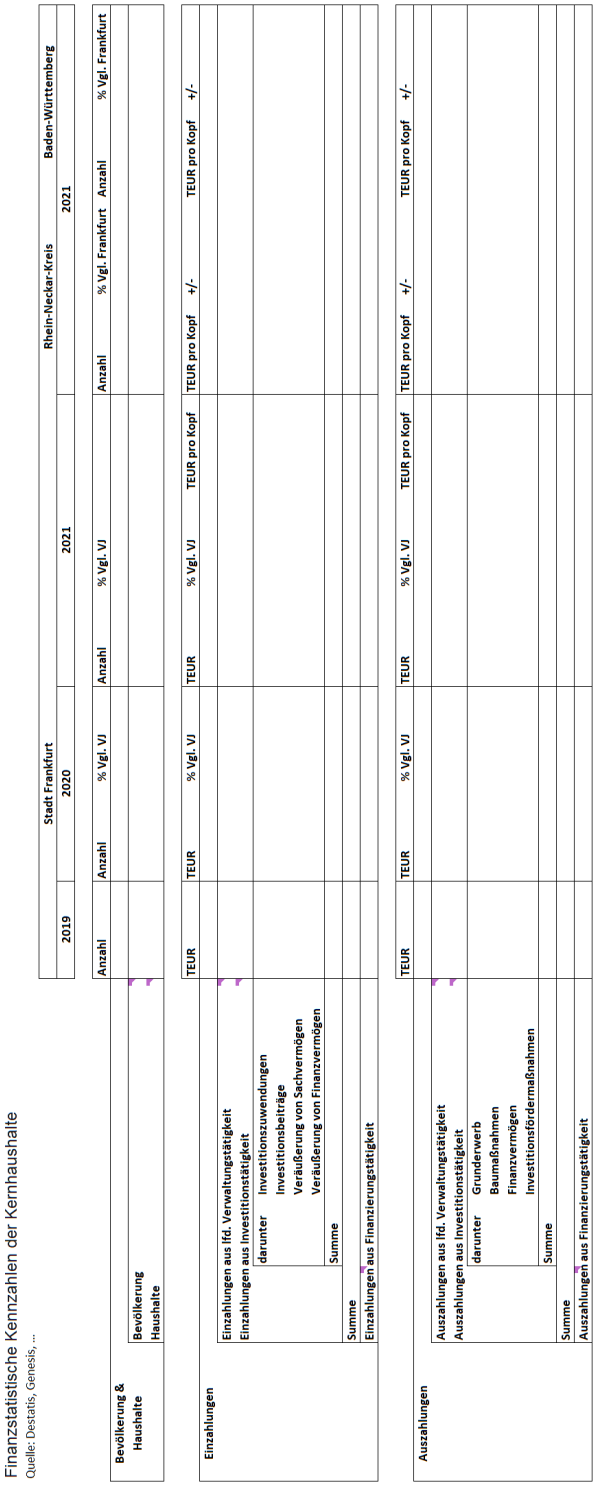


Figure 31: Exemplary Statistical Evaluation Sheet

Based on the functionality of "location sharing" within the Element platform, the bot can obtain the location of users, if they decide to proactively share their location with the bot through a message. The evaluation can therefore be displayed directly for the municipality in which the user is located.

5.2 CHARACTERIZATION OF DATA

5.2.1 Data Sources

Data from the statistical offices of the federal states and the federal government

The central goal of the EU SALTED project, to provide high-quality open data with economic potential, is naturally based on corresponding objectives and efforts at the levels of the states and regions of all European countries.

Accordingly, relevant strategies are also being implemented in Germany by the statistical offices of the federal states and the Federal Government. Here, a collaborative approach is being pursued by the federal states: Based on the definition of an overarching data strategy and a project portfolio derived from it, the statistical offices of the federal states take on the implementation of individual subcomponents in independent projects, which are then integrated into an overarching open data infrastructure.

Of course, there are Open Data projects in the federal states and their subordinate administrative levels of the regions and municipalities that are not strictly integrated into the centrally defined project portfolio. In these projects, however, care is usually taken to ensure that data can be automatically merged by way of "harvesting".

The focus of the content of the data provided by the statistical offices is on data that forms the basis of administrative action or that results from administrative processes, i.e. data that can be attributed to the context of open government.

The data portal in which data is compiled and published within the network of the statistical offices of the federal states and coordinated by the Federal Statistical Office is the GENESIS portal:

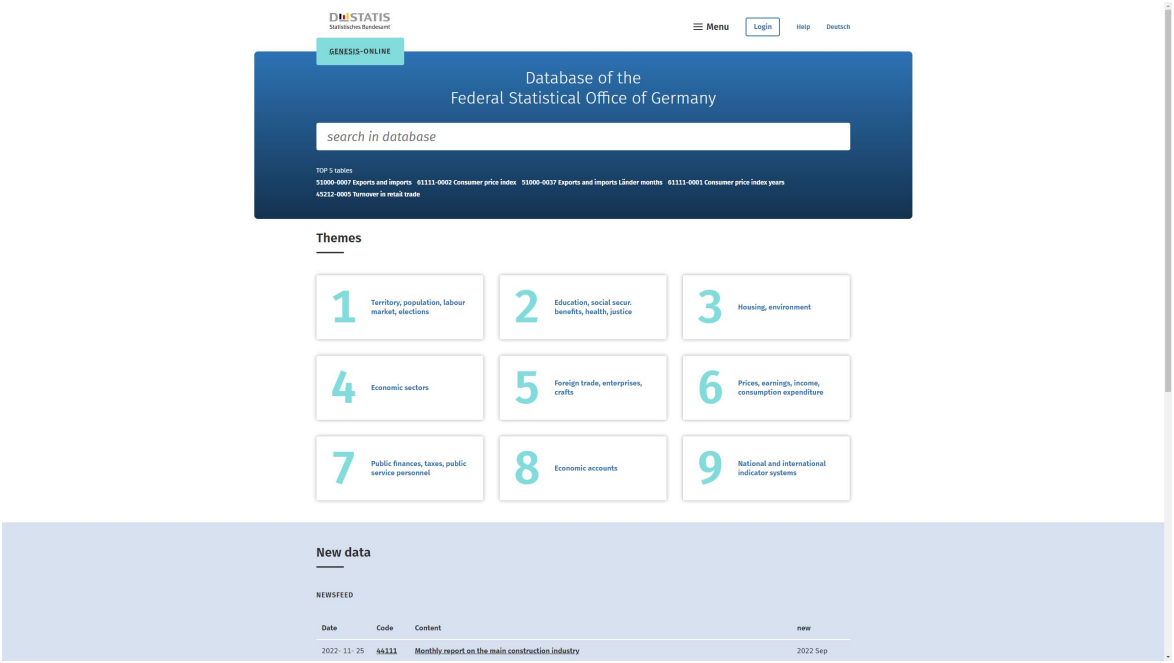


Figure 32: GENESIS platform[60]

One of the tasks of the SALTED project is to prove that the development and enhancement of already openly available data is also possible through SALTED and generates added value.

The foreseeable added value of opening up data from the above-mentioned open data portals for the SALTED platform consists of the following points:

- As part of the creation of a SALTED Smart Data Model, new entities are defined or existing entities are enriched in terms of content - in this use case, for example, "local authorities" as units of public administration.
- Various data sets from the Open Data Portal of the Statistical Office can be linked with each other in the sense of Linked Data; they can also be merged with data sets from other sources on the SALTED platform.[61]
- Data sets are more easily findable and usable through a unified data model (NGSI-LD) and distribution platform (Scorpio broker).

5.2.2 Limitations in collecting the data

The data portal of the Federal Statistical Office only allows manual data queries on the basis of anonymous access - i.e. no automated processes.

[60] https://www-genesis.destatis.de/genesis/online/data?operation=sprachwechsel&language=en

[61] Kybeidos is are currently checking the availability of information on the activities of municipal enterprises, i.e. enterprises that are majority publicly owned and that perform key tasks of general interest - namely the provision of services and infrastructure in the areas of mobility, energy and water supply, etc. For these public enterprises, information on aspects of company law, e.g. on the public corporations holding shares, can be obtained from the federal government's statistics portals. For these companies, however, data from the smart city context is also of great operational importance due to the nature of their business. The possibility of linking information across domains on the basis of SALTED is therefore of high interest here.

For the regular and automated retrieval of data (harvesting), individualized access (named account) is required, which is, however, usually provided free of charge.

5.2.3 Data Licence

As a rule, the statistical offices of the federal states and the Federation make data available free of charge under the licence "Data Licence Germany - Attribution - Version 2.0"[62]. The formulation in which the name is to be used is specified for each office that publishes data.

Under the conditions of this licence, the data can be used freely for commercial and non-commercial purposes.

5.3 TYPE OF STATISTICAL DATA

The data portals of the federal states in Germany and the central data portal of the Federal Statistical Office provide open data on a broad range of topics.

It is certainly true that these data are already used intensively by companies and institutions that have relevant economic interests or a public or political mandate.

The use of the data by civil society groups, however, is currently still hampered by hurdles in the form of required technical skills. The development of this data through SALTED can possibly contribute to removing these hurdles.

In the following sections, examples are given of topics for which open data is provided by statistical offices and in which added value can be realized by linking this data with data from the smart city context.[63]

The "axes" on which socio-economic data are collected are those of social structures on the one hand and the provision of economic services on the other.

These axes can be outlined as follows

5.3.1 Social data

- Area & Population
- Labour market
- Elections
- Education, Social Services, Health, Law
- Housing & Environment

5.3.2 Economical data

- Economic sectors
- Foreign trade, enterprises, crafts

[62] http://www.govdata.de/dl-de/by-2-0

[63] It should be pointed out once again that these data are already accessible and linkable for tasks of "professional" spatial planning and urban development - but not or only to a limited extent for free and public access.

- Prices, earnings, income and consumption
- Public finances, taxes, personnel

5.3.3 Data to be processed in the SALTED project

From the universe of socio-economic data, in SALTED a small section is considered, which is very relevant in the context of bringing digital technologies into cities - against the background of the following reflection:

Data on the economic performance of communities is of great importance in connection with almost all transformation processes that are demanded of cities and municipalities - among other things, with regard to ensuring that subsidies and support for future investments are granted in a targeted and fair manner that is also transparent and comprehensible for civil society.

We have therefore decided to integrate data from the area of financial statistics within the framework of the SALTED project, to implement a data enrichment pipeline for these as an example and to make the data available to users from civil society via a messaging platform.

5.4 SOCIOECONOMICAL DATA CRAWLING MECHANISM

5.4.1 Search & Collection

The GENESIS data portal, from which financial statistics data are to be obtained, provides an API that implements both the functions of searching and retrieving metadata and the functions of querying data. The description of this API is available at: https://www-genesis.destatis.de/genesis/misc/GENESIS-Webservices_Einfuehrung.pdf

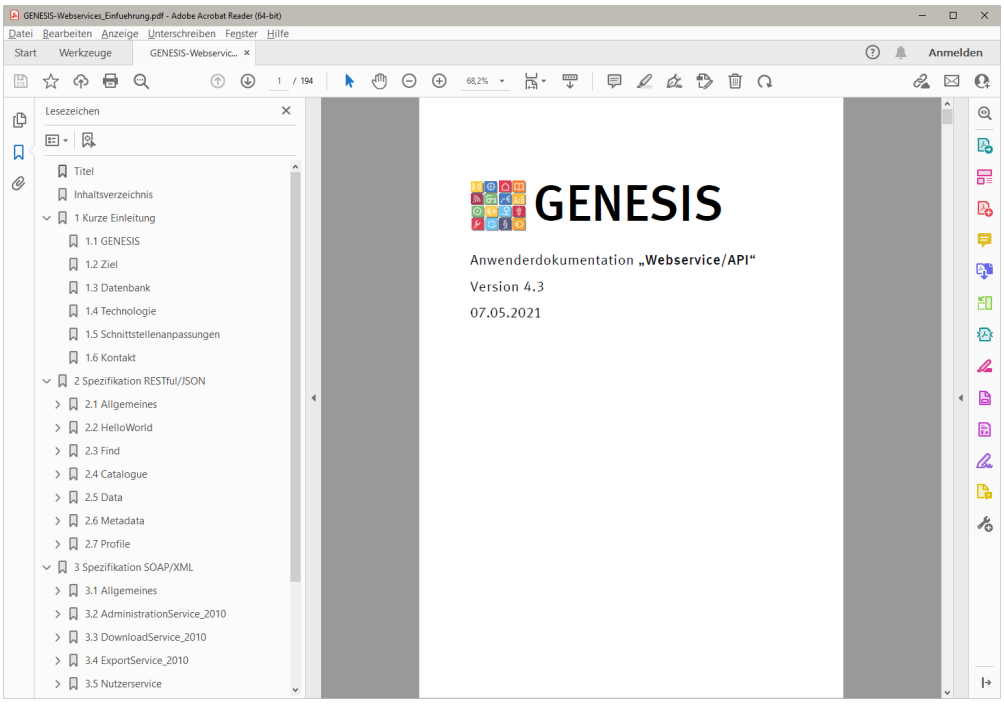


Figure 33: Documentation of the GENESIS API

Compared to data portals like CKAN, GENESIS has similar functions for querying data and metadata, but does not provide a query language like SPARQL and no functions for cataloguing new content; in GENESIS, the latter is a function only available to the back office of the statistical offices.

Examples of data requests and processes to get them:

Login & Authentication

Request:

https://www-genesis.destatis.de/genesisWS/rest/2020/helloworld/logincheck?username=TheUsername&password=ThePassword&language=de

Response:


```
{"Status":"Sie wurden erfolgreich an- und abgemeldet!",
"Username":"TheUsername"}
```


Find Request

Request:

https://www-genesis.destatis.de/genesisWS/rest/2020/find/find?username=Username&password=ThePassword&term=FEU&category=all&pagelength=16&language=de

Response:


```
{     "Ident":
      {"Service":"find","Method":"find"},
      "Status":
      {"Code":0,"Content":"erfolgreich","Type":"Information"},
      "Parameter":
      {"username":"********************",
      "password":"********************",
      "term":"FEU",
      "category":"Alle",
      "pagelength":"16",
      "language":"de"},
      "Cubes":
      [
      {     "Code":"71811BJ001",
            "Content":"Jahresabschlüsse kfm. b. Extrahh., sonst.
            öff. FEU, Öffentliche Fonds, Einrichtungen und Unter
            nehmen, FEU mit Angaben zum Anlagevermögen, FEU mit An
            gaben zu öff. Zuweisungen u. Zuschüssen, Deutschland
            insgesamt, Jahr",
            "State":"vollständig mit Werten",
            "Time":"2008-2020",
```

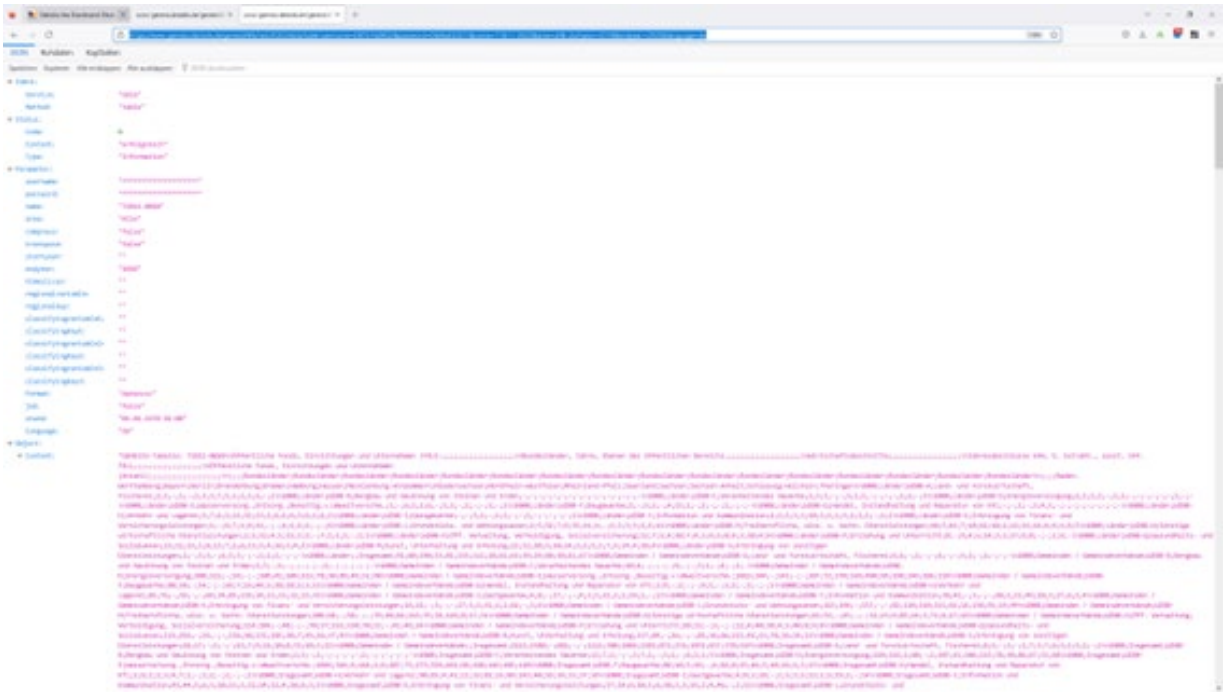
```
            "LatestUpdate":"21.11.2022 09:20:11h",
            "Information":"false"
        },
        {       "Code":"71811BJ002",
            "Content":"Jahresabschlüsse kfm. b. Extrahh., sonst.
            öff. FEU, Öffentliche Fonds, Einrichtungen und Unter
            nehmen, FEU mit Angaben zum Anlagevermögen, FEU mit An
            gaben zu öff. Zuweisungen u. Zuschüssen, Deutschland
            insgesamt, WZ2008 (Abschnitte), Jahr",
            "State":"vollständig mit Werten",
            "Time":"2008-2020",
            "LatestUpdate":"21.11.2022 09:25:13h",
            "Information":"false"
        }
        ]
}
```


Data Request

Request:

Erreur ! Référence de lien hypertexte non valide.

Response:




```
{           "Ident":
    {       "Service":"data","Method":"table"},
            "Status":{"Code":0,"Content":"erfolgreich","Type":"Information"},
            "Parameter":{"username":"********************","password":
            "********************",
            "name":"71811-0020",
            "area":"Alle",
            "compress":"false",
```

```
            "transpose":"false",
            "startyear":"",
            "endyear":"2020",
            "timeslices":"",
            "regionalvariable":"",
            "regionalkey":"",
            "classifyingvariable1":"",
            "classifyingkey1":"",
            "classifyingvariable2":"",
            "classifyingkey2":"",
            "classifyingvariable3":"",
            "classifyingkey3":"",
            "format":"datencsv",
            "job":"false",
            "stand":"01.01.1970 01:00",
            "language":"de"
        },
        "Object":
        {       "Content":"GENESIS-Tabelle: 71811-0020\nÖffentliche Fonds, Einrichtungen und Unternehmen
                (FEU):;;;;;;;;;;;;;;;;;;;\nBundesländer, Jahre, Ebenen des öffentlichen
Bereichs,;;;;;;;;;;;;;;;;;;;\nWirtschaftsabschnitte;;;;;;;;;;;;;;;;;;\nJahresabschlüsse kfm. b.
Extrahh., sonst. öff. FEU;;;;;;;;;;;;;;;;;;;\nÖffentliche Fonds, Einrichtungen und Unternehmen
(Anzahl);;;;;;;;;;;;;;;;;;;\n;;;;Bundesländer;Bundesländer;Bundesländer;Bundesländer;Bundesländer;Bun
desländer;Bundesländer;Bundesländer;Bundesländer;Bundesländer;Bundesländer;Bundesländer;Bundesländer;
Bundesländer;Bundesländer;Bundesländer\n;;;;Baden-
Württemberg;Bayern;Berlin;Brandenburg;Bremen;Hamburg;Hessen;Mecklenburg-
Vorpommern;Niedersachsen;Nordrhein-Westfalen;Rheinland-Pfalz;Saarland;Sachsen;Sachsen-
Anhalt;Schleswig-Holstein;Thüringen\n
2008;Länder;WZ08-A;Land- und Forstwirtschaft, Fischerei;2;1;-;1;-;1;3;3;7;1;1;1;3;2;-
;1\n2008;Länder;WZ08-B;Bergbau und Gewinnung von Steinen und Erden;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-
\n2008;Länder;WZ08-C;Verarbeitendes Gewerbe;2;3;1;-;-;3;1;1;-;-;-;-;2;2;-;3\n
2008;Länder;WZ08-D;Energieversorgung;2;1;2;1;-;1;1;-;-;-;-;-;-;1;-;-\n
2008;Länder;WZ08-E;Wasserversorg.,Entsorg.,Beseitig.v.Umweltverschm.;1;-;6;1;2;6;-;3;3;-;2;-;-;3;-
;2\n2008;Länder;WZ08-F;Baugewerbe;1;-;5;2;-;4;23;1;-;2;-;-;1;-;-;-\n
2008;Länder;WZ08-G;Handel, Instandhaltung und Reparatur von Kfz;-;-;1;-;3;4;1;-;-;-;-;-;-;-;-;-
\n2008;Länder;WZ08-H;Verkehr und Lagerei;5;7;4;2;11;35;13;1;4;4;6;3;6;1;4;1\n
[...]
;Insgesamt;3600;2570;400;780;123;305;1579;469;1821;3036;1315;354;1079;551;591;636\n
__________\n
© Statistisches Bundesamt (Destatis), 2022\nStand: 11.12.2022 / 14:37:46\n",
            "Structure":
        {"Head":
           {"Code":"71811","Content":"Jahresabschlüsse kfm. b. Extrahh., sonst. öff.
            FEU","Type":"Statistik","Values":null,"Selected":null,"Structure":
           [
           {"Code":"FEU001","Content":"Öffentliche Fonds, Einrichtungen und Unternemen",
            "Type":"Merkmal","Values":null,"Selected":null,"Structure":null,"Updated":"see parent"
           }
           ],"Updated":"see parent"},"Columns":
           [
           {"Code":"DLAND","Content":"Bundesländer","Type":"Merkmal","Values":"16",
            "Selected":"16","Structure":null,"Updated":"see parent"
           }
           ],
           "Rows":
           [
           {"Code":"WZ08BA","Content":"WZ2008 (Abschnitte)","Type":"Merkmal","Values":"19",
            "Selected":"19","Structure":null,"Updated":"see parent"
           }
           ],"Subtitel":
           {"Code":"JAHR","Content":"Jahr","Type":"Merkmal","Values":"0","Selected":"0",
            "Structure":null,"Updated":"see parent"
           },
           "Subheading":
           {"Code":"KRPEB4","Content":"Ebenen des öffentlichen Bereichs","Type":"Merkmal",
            "Values":"2","Selected":"2","Structure":null,"Updated":"see parent"
           }
        }
    },"Copyright":"© Statistisches Bundesamt (Destatis), 2022"
}
```


5.4.2 Data Format

The data under consideration for socio-economic data will be delivered at fixed cut-off dates in the form of complete data sets of one or more federal states. Stream processing functions are therefore not to be implemented.

5.5 EMPLOYED DATA MODELS

The final publication of the structures in which the financial statistics data in question will be provided is expected to take place in the course of the first quarter of 2023.

In the project, a structure was developed from relevant sources that are likely to be a very good approximation of the future data structure; these sources are:

- “Guide to Budget Structure in the New Local Government Budget and Accounting System“
  Ministry of the Interior, for Digitalization and Local Authorities Baden-Württemberg
  https://im.baden-wuerttemberg.de/fileadmin/redaktion/m-im/intern/dateien/pdf/Leitfaden_zur_Haushaltsgliederung_Stand_2010.pdf
- "Hessian Municipal Statistics"
  Hessian State Statistical Office
  https://statistik.hessen.de/publikationen/hessische-gemeindestatistik
- “Key financial statistics of the core budgets of the municipalities and municipal associations 2013 - 2018 (hessen.de)"
  Hessian State Statistical Office
  https://statistik.hessen.de/sites/statistik.hessen.de/files/2022-06/lii1-j18.pdf

The specific used data models and the mapping to NGSI-LD will be developed starting in December 2022 and will form the basis for the elaboration of the following chapters.

5.5.1 KeyPerformanceIndicator

The financial indicators of the statistical offices are structured multi-dimensionally. Each individual report usually consists of 3 or more dimensions, which can also be structured in rows or columns on several levels. They are therefore not atomic, but complex indicators.

Objectively - i.e. especially with regard to the attributes - the existing Smart Data Model that best describes complexly structured financial indicators is the "KeyPerformanceIndicator" Data Model.

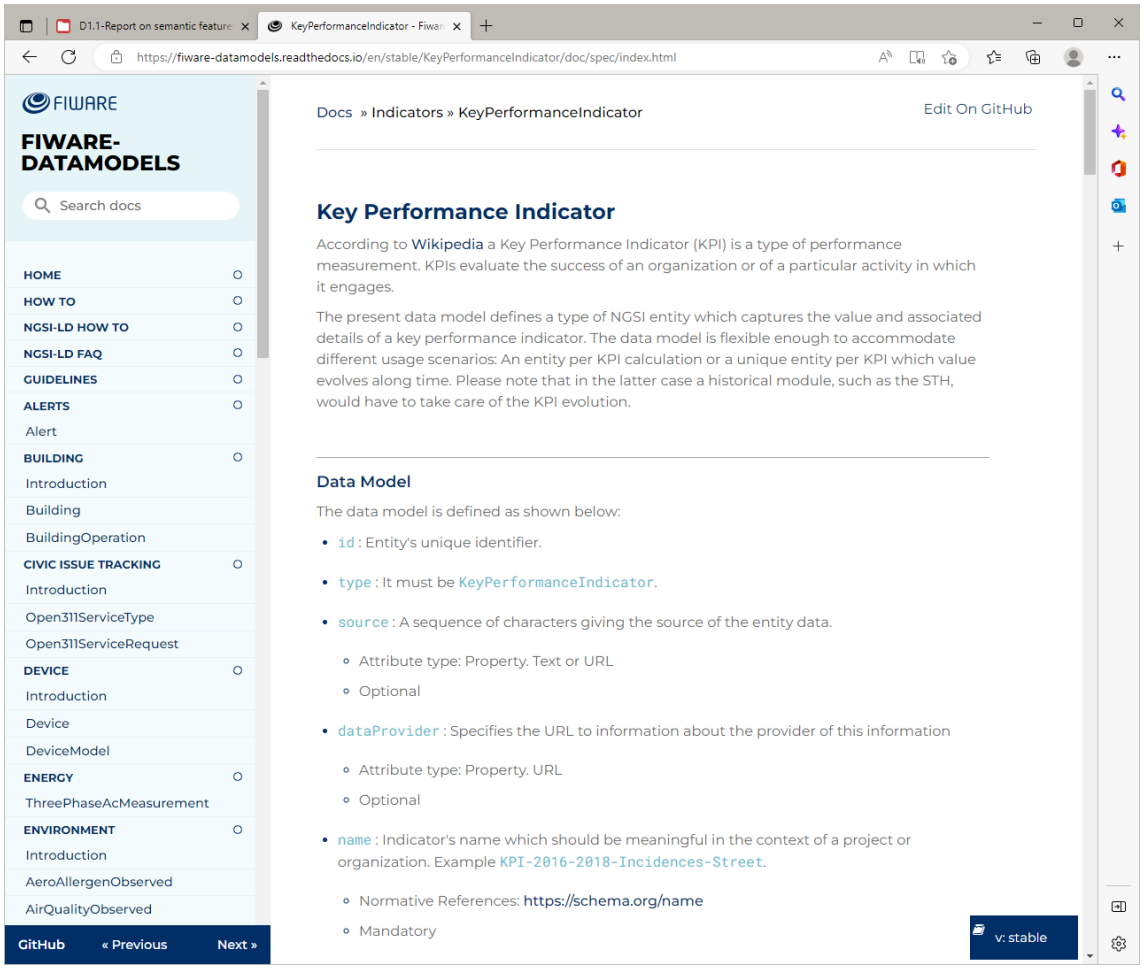


Figure 34 KeyPerformanceIndicator data model[64]

In particular, the attributes - "organisation", "provider", "calculationFrequency" etc. - are exactly those with which financial ratios are to be described. Only "value" in this case is not a single value, but a data vector consisting of possibly several key and data values.

The structure of these complex "value" attributes is described by metadata of the individual data sources, cf. e.g. the following report structure with id "71811-0001":

64 https://fiware-datamodels.readthedocs.io/en/stable/KeyPerformanceIndicator/doc/spec/index.html

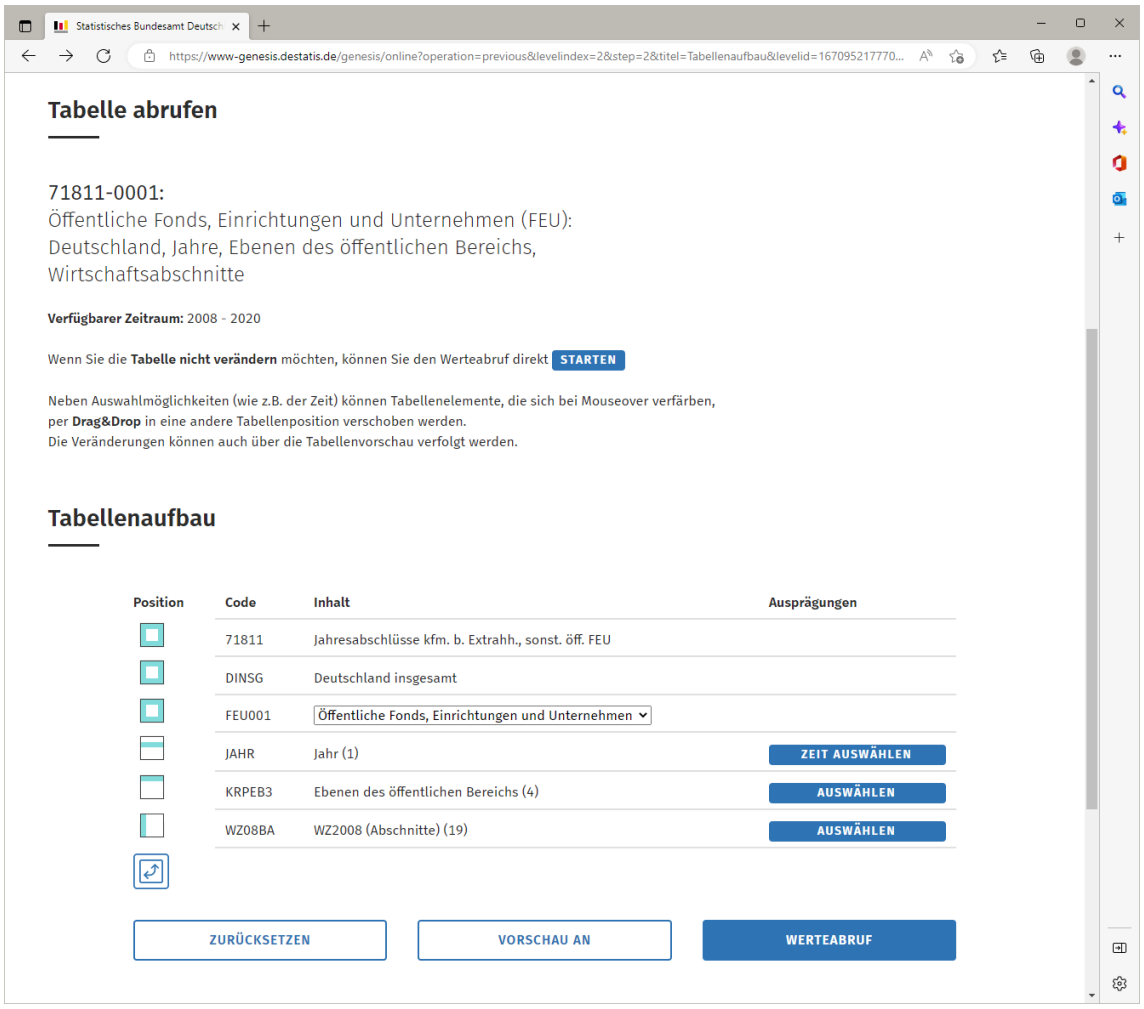


Figure 35 metadata example for one value attribute[65]

The challenge in the representation of complex indicators in the KeyPerformanceIndicator data model is therefore to provide or reference the metadata so completely that the complex structure can be resolved clearly and easily.

As of 2022-12, this aspect is still the subject of the development of the present concept.

[65] https://www-genesis.destatis.de/genesis/online?operation=previous&levelindex=2&step=2&titel=Tabellenaufbau&levelid=1670952177709&acceptscookies=false#abreadcrumb

5.6 DATA CURATION

5.6.1 Features of the curated Socioeconomical Data

The data on financial statistics of municipalities and the federal states are aggregates of detailed data from the accounting of municipalities. In the process of publishing the data, offsets and adjustments of payments between different regional levels take place.

When curating these data, it is, therefore, necessary to make transparent how published key figures are defined and what offsets and adjustments have been made.

5.6.2 Data quality dimensions metadata linking

With regard to data published by the statistical offices of the federal states and the Federation in Germany, we are in the fortunate position that the data have already undergone a multi-stage process of quality assurance at the time of publication and meet the highest standards of quality.

Nevertheless, corrections to the data cannot be ruled out as a matter of principle. Important information on the quality and on any corrections that may have been made include:

- The date of publication of data records,
- Errata, if available, and, if applicable, statements communicated after publication on limitations to data quality and usability.

5.7 ALIGNMENT TO THE SALTED ARCHITECTURE

The provision of socio-economic data provided by the German Federal Statistical Office is an example of the indexing and semantic and analytical enrichment of data - which is then published on the Context Broker and, if applicable, on the European data portal.

The data used as an example in this application are made available on an annual basis on the statistics portal GENESIS of the Federal Statistical Office - at the level of territorial authorities of the defined regional levels: the cities and municipalities, the association municipalities, the districts, the regional districts and the federal states.

Through enrichment processes on the SALTED platform, the data are enhanced in the following way:

- Regional unit keys are linked to geographical information so that information can be mapped and linked to other geolocated data.
- The enrichment process will also show how time series aspects and the linking with hierarchically higher or lower territorial units as well as the calculation of derived key figures can be done.

From an architectural point of view, the bot itself is a central service that interacts with users in the communication spaces of the messaging platform. It accesses data published on the Scorpio broker via the intended application interfaces.

The value of this demonstrator lies in the fact that it exemplifies how social media - in this case the messaging platform - can be used as a channel for the provision of data from the context of open government.

Bots are also a way to address new user groups in a low-threshold way for classic data and applications from the smart city context.

6 NATIONAL AND INTERNATIONAL METEOROLOGICAL AGENCIES

6.1 CHARACTERIZATION OF METEOROLOGICAL DATA

6.1.1 Data Sources

SmartSantander

The *SmartSantander* IoT infrastructure[66], already mentioned in Section 2, has thousands of sensors providing several kinds of traffic and environmental data. For this section, the specific types of data that have a direct relation to the theme are the air quality data, and the temperature data. All the sensor observations are obtained through a REST subscription API, resulting in stream data that is collected in an asynchronous way.

Other Spanish cities

Most cities in Spain provide an Open Data portal that allows for the collection of batch data through periodic requests. Since we are driven by some of the use cases of the project, specifically the use case where we are trying to correlate traffic data from Smart Cities and pollution or air quality data from either Smart Cities or meteorological agencies, we have focused our efforts on the cities where we have access to both of these kinds of data. This has resulted in gathering data from Barcelona, Bilbao, Santander, Valencia and Vitoria. The Open Data portals provided by these cities were already shown in Section 2. However, we have also collected historical meteorological data from the Spanish government[67] website.

European agencies

We wanted to expand the scope of the aforementioned use case to the entirety of Europe. For instance, Dublin is one of the key cities that is allowing us to perform research on the correlation between traffic and pollution. For this reason, we have also used data published into the European Environment Agency (EEA)[68] website. Specifically, we have collected very valuable air quality data from one of their Linked Tables[69] that provides more than 4 million entries from several countries and timespans.

6.1.2 Regional availability and limitations in collecting the data

Generally speaking, most of the Open Data portals we have visited don't provide the meteorological data that we were looking for. There were some cases where only the last update of the data was available, but they didn't provide historical data which is a key feature for our research. For this reason, we had to look for data providers with more scope (i.e. the EEA) that actually provided the meteorological data that we needed. Some important European cities are lacking meteorological and air quality data, whether in their Open Data portals, the EEA or the European Data Portal (EDP). This has been a limitation on extending the scope of the research, as it means that either they don't have the data or they are not providing it publicly. However,

[66] https://api.smartsantander.eu/
[67] https://www.miteco.gob.es/es/
[68] https://www.eea.europa.eu/
[69] https://discomap.eea.europa.eu/App/AirQualityStatistics/index.html

we consider that we will be able to carry out successful research with the cities that we have found.

6.2 TYPE OF METEOROLOGICAL DATA

6.2.1 Public Providers

In most cases, meteorological data will be open and available to the public. Cities provide this data through Open Data portals, while other agencies and public entities often have data catalogues redirecting to the same endpoints (in the style of the EDP). In some cases, these agencies may provide data that was previously private or at least not easily accessible, as in the case of the EEA. Sometimes the key limitation when obtaining the data is not that it is private, but rather the way of discovering and/or collecting it even though it is public.

6.2.2 Private Providers

We have, however, found some cases where data gathered by cities was not public. The most common case is Open Data portals that provide real-time data. Once a new batch of data enters their system, the previous batch is made unavailable since there is only one endpoint. Presumably, these cities have a storage system where they keep historical data but it is unavailable to the public. The alternatives for the collection are either find it through an agency or a public entity or contact the provider directly and ask for the data, which is not a very dynamic solution.

6.3 METEOROLOGICAL DATA COLLECTION MECHANISM

6.3.1 Search

The data we have collected in the section is, as stated previously, available mostly in Open Data portals or public agencies. This narrows the search we have to do as data collectors since it is straightforward to look for the specific datasets we want (i.e. meteorological) and request it from the corresponding endpoints. We have faced the challenge of having to find data that wasn't directly available in some of the cities' Open Data portals. We have tackled this by using the EEA website as a nexus for all the European cities with air quality data and filtering the ones we wanted one by one. In terms of time periods, we are interested in data from April 2019, April 2020, August 2020, April 2021 and April 2022. This will allow us to perform an analysis on months with very different traffic profiles (due to the lockdowns).

6.3.2 Collection

Once we have found the meteorological data, the collection step is normally performed through an HTTP request to the Open Data portal API. We are collecting data from specific time periods, which means that in some cases we collect from the direct links found on the Open Data website. This is easier than making programmatic HTTP requests as long as the data to be collected is limited in the number of files. In the case of the EEA website, after filtering the data from every individual city, we have downloaded a JSON document with a batch of data covering a specific place and timespan, which we can then process the same way as the rest.

6.4 EMPLOYED DATA MODELS

6.4.1 AirQualityObserved

This Smart Data Model[70] describes an observation of air quality conditions at a specific place and time. This is one of the most complete data models we have used since it includes specific properties for a great number of pollutants. We discuss below those that we have used:

- **co**: Carbon monoxide level. Units can be specified using the UN/CEFACT Common Codes (this applies to all the properties below).
- **co2**: Carbon dioxide level.
- **no2**: Nitrogen dioxide level.
- **o3**: Ozone level.
- **pm1**, **pm10** and **pm25**: Level of particulate matter 1 micrometre, 10 micrometres and 2.5 micrometres or less in diameter, respectively.
- **so2**: Sulphur dioxide level.
- **relativeHumidity**: Level of relative humidity in the air.

This data model includes some other properties that can be useful given the right context, such as wind speed, wind direction and many other pollutants. However, we have been focusing on those we have access to, and that are useful for our targeted use cases.

6.4.2 Temperature

This Smart Data Model[71] describes the value of temperature at a specific place and time. It has been adapted from the CIM data models. The most significant properties, or the ones we have used for our meteorological data, are briefly discussed below:

- **value**: Value of the temperature measurement. Default unit: °C.
- **unit**: Can be used to specifically use another unit for the temperature value. An alternative to UN/CEFACT Common Codes.

6.4.3 Mapping of raw data to the NGSI-LD Model

In the case of meteorological data, it is very common to collect the data in either CSV or JSON format. The temperature entities have a more straightforward mapping since they only include the value and unit of measurement. This means they can be mapped to NGSI-LD by way of a JMESpath[72] template. The air quality entities are a bit more complex, as they can be divided into several air pollutants that have to be put together before proceeding to the mapping itself. We have done this by identifying measurements from the same pollution stations taken at the same moment, which can be easily done by matching the timestamps and the station identifiers. Once the measurement has been put together, we can use a template as in the previous case, as long as the template includes all the pollutants. This can be assured by creating a template that

[70] https://github.com/smart-data-models/dataModel.Environment/tree/master/AirQualityObserved
[71] https://github.com/smart-data-models/dataModel.EnergyCIM/tree/master/Temperature
[72] https://jmespath.org/

covers all of the existing pollutants in the AirQualityObserved Smart Data Model. As for CSV files, they can easily be transformed into JSON objects by using the standard *CSV* and *JSON* packages for Python.

6.5 Data Curation

The curation of meteorological data is very similar to that seen in Section 2. Firstly, the curation module detects and tags novelties in the new observation by performing machine learning techniques in the previous datasets. These datasets contain not only the numerical values of the measurements but also spatial and temporal dimensions such as location and time of the year (extracted from the timestamps).

The next step is the addition of metadata to the original NGSI-LD entities. These metadata properties are linked to data quality and include frequency, completeness, precision and accuracy of the new measurement. Several mechanisms can be appended to the data where the results are off the mark. For instance, if the completeness property indicates that there is data missing, we can apply interpolation techniques in order to generate synthetic measurements.

Finally, we also check for errors in the data which are usually derived from the malfunctioning of the sensor generating said data. For instance, a sensor may be providing the same exact value for two days or some of the measurements may be obviously wrong (e.g. a negative CO2 value).

6.6 Ingestion Process

6.6.1 Data streams collection

As explained in Section 2, in order to inject stream data into the Scorpio Context Broker, we use the standard interface for the injection of individual entities. This is done with a POST request to the */entities* endpoint.

6.6.2 Datasets collection

Datasets are the most common way of collecting meteorological data since it doesn't make sense to provide historical data as asynchronous streams. Batches of NGSI-LD entities can be pushed into the Scorpio Context Broker via the */entityOperations/upsert* endpoint. This allows us to push a dataset of both new and existing entities into the broker, who will differentiate automatically between them, creating the new entities and updating the existing ones.

6.7 Interaction

The services provided in this section can implement an additional interface acting as an MQTT client. This way, higher-level services can request specific actions from the injection chain in order to, for instance, collect new data that is required by an external app. The *SmartSantander* injection chain is already collecting all available data in the platform, so it would not make sense to include a closed-loop interface. However, the other collectors based on requests to specific meteorological-related endpoints can receive requests from the control plane to start collecting data from new sources (i.e. a new city), as long as the correct template is also provided in the MQTT message.

6.8 ALIGNMENT TO THE SALTED ARCHITECTURE

Meteorological data is, essentially, one of the key types of data contemplated in the project. As such, it is straightforward to analyze its contribution to the architecture. From the collection of the data, both in batch and in stream mode, to the final injection into the broker, meteorological data follows the architecture presented in Deliverable 2.1 as a by-the-book SALTED injection chain. This can be seen in Figure 36.

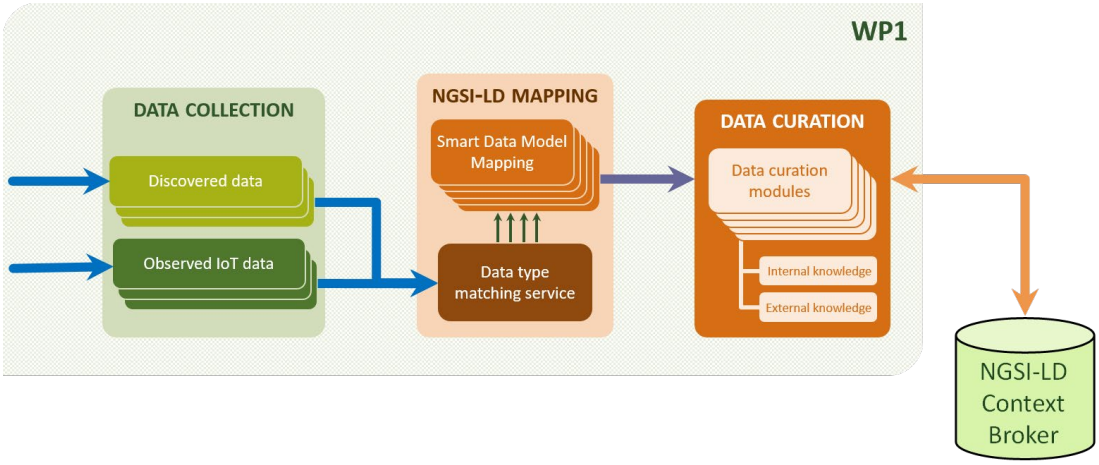


Figure 36: Standard SALTED injection chain

In the case of *SmartSantander*, it is a stream-based injection chain that includes collection, mapping and curation for individual measurements/entities. As for the data extracted from European cities, the only change is that it is collected as a batch, and treated as such through the rest of the injection chain. Therefore, the sum of all meteorological data described in this section is contributing to the architecture as a set of injection chains working in parallel and pushing NGSI-LD entities into the Scorpio Context Broker in both stream and batch modes.

7 CONCLUSIONS

The SALTED project has tackled the problem of defining and using an injection chain reusable by different use cases. The general idea of it, in fact, has been satisfactorily applied and demonstrated along the different examples described in this document. A variety of data sets have been collected and the way to format them according to NGSI-LD defined data models have been described. Some curation procedures have also been identified and proposed in order to improve the quality of the data. These specifications and the related development work resulted in the availability of data that are relevant for the management of urban processes in the sense of the basic FIWARE approach and infrastructure and that can be also integrated and combined in order to exploit the linking and the enrichment of data sets. This is a very relevant result fully aligned with the expected goals of the SALTED project.

Some major points of attention emerged and they may need some consideration for future work:

- The search of sources. The search and identification of existing sources have been essentially left to the users. The reason for this is that they know the problem and the context they want to tackle and most likely, they know the trusted sources of data; or, at least, after a brief analysis, these data can be identified and retrieved by humans. The Social Media Use Case has, however, shown the possibility of searching specific keywords and creating a data set accordingly to them. The use case, in fact, has been built around the possibility offered to the user to specify a few parameters to be used to crawl some social media. In this case, the search for data sets is somehow substituted by the dynamic creation of the data set by crawling. Once the data set has been built, the mapping to the newly defined Data Model can be executed and the data curated.
- Different types of processes for data acquisition have been considered. The aforementioned *crawling* (on-demand) is one possibility that could be generalized in terms of means of requesting the dynamic acquisition of certain data for a certain period of time. *Real-time flow* processing is another type of data management that has been tackled (e.g. the smart city real-time data acquisition). In this case, the mapping to the model is generally not too complicated, but instead, the understanding and the quality of the received data pose some more questions. It is important to have a clear understanding, for instance, of the acceptable data range and the trustiness of the specific source. Interpolation techniques are not necessarily applicable to long sequences of missing data. On the other side, *batch processing* (again in the smart city example) can cope and provide additional insights on how to manage and improve the quality of large data sets.
- The creation of a catalogue of available data. This is somehow related to the first bullet point. In principle, it would be extremely useful to organize the available data into a searchable catalogue in such a way to create a one-stop-shop for large communities of data engineers. However, the current experiences (e.g., the Madrid Open Portal) are not fully exploited by people and there is relevant work to do in the design of the catalogue and then on the management of its structure in order to keep it updated to the actual content. In spite of these difficulties, future work could

comprise an attemptble to create such a well-formed and well-maintained catalogue of NGSI-LD based data models and related data sets so that researchers and practitioners can find the desired data. Artificial Intelligence techniques as well as annotation, enrichment and linking functions provided by the SALTED architecture could provide an added-value in this sector.

The next step in the validation of the SALTED proposal is to actually implement and use these injection chains for building new opportunities to create services. Many involved partners have expressed their interest in further work on the specific injection chains in order to internally build the capabilities to quickly operate and deal with the data without the need to repeatedly curate and transform the data. This is an encouraging stimulus because the project has a very practical spirit and it is eager to “eat its own food”.

8 BIBLIOGRAPHY

[1] "Situation-Aware Linked heTerogeneous Enriched Data D2.1: Report on Data Linking and Enrichment Architecture Work package WP 2 Task Task 2.1."

[2] E. Kovacs, M. Bauer, J. Kim, J. Yun, F. Le Gall, and M. Zhao, "Standards-Based Worldwide Semantic Interoperability for IoT," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 40–46, Dec. 2016, doi: 10.1109/MCOM.2016.1600460CM.

[3] K. T. Jánosi-Rancz and Á. Lajos, "Semantic Data Extraction," *Procedia Technol.*, vol. 19, pp. 827–834, Jan. 2015, doi: 10.1016/J.PROTCY.2015.02.119.

[4] D. Dou, H. Wang, and H. Liu, "Semantic data mining: A survey of ontology-based approaches," *Proc. 2015 IEEE 9th Int. Conf. Semant. Comput. IEEE ICSC 2015*, pp. 244–251, Feb. 2015, doi: 10.1109/ICOSC.2015.7050814.

[5] S. Aral, C. Dellarocas, and D. Godes, "Social media and business transformation: A Framework for research," *Inf. Syst. Res.*, vol. 24, no. 1, pp. 3–13, 2013, doi: 10.1287/isre.1120.0470.

[6] K. Zarei, R. Farahbakhsh, N. Crespi, and G. Tyson, "Dataset of Coronavirus Content from Instagram with an Exploratory Analysis," *IEEE Access*, vol. 9, pp. 157192–157202, 2021, doi: 10.1109/ACCESS.2021.3126552.

[7] M. AbuKausar, V. S. Dhaka, and S. Kumar Singh, "Web Crawler: A Review," *Int. J. Comput. Appl.*, vol. 63, no. 2, pp. 31–36, Feb. 2013, doi: 10.5120/10440-5125.

[8] F. Shi, Q. Li, T. Zhu, and H. Ning, "A survey of data semantization in internet of things," *Sensors (Switzerland)*, vol. 18, no. 1, p. 313, Jan. 2018, doi: 10.3390/s18010313.