OptEEmAL

OPTIMISED ENERGY EFFICIENT DESIGN PLATFORM FOR REFURBISHMENT AT DISTRICT LEVEL

Optimised **E**nergy **E**fficient Design Platfor**m** for Refurbishment a**t** District **L**evel
H2020-WORK PROGRAMME 2014-2015 – 5. Leadership in enabling and industrial technologies
H2020-EeB-05-2015: Innovative design tools for refurbishment at building and district level

D1.3: Requirements and specification of geo-clustering data sets access module

WP1, Task 1.3

August 2016 (m12)

Deliverable version: **D1.3, v1.5**

Dissemination level: **Public**

Author(s): **Patricio Moreno[1], Giuliana Mangiapia[2], Ciro Caterino[2], Mariano Folla[2], Alvaro Sicilia[3], Maxime Pousse[4,] Pedro Fernandez-Orellana[7],**

([1]ACC, [2]ES, [3]FUNITEC, [4]NBK, [5]TEC, [6]TUC, [7]UTRC-I)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 680676

Document History

Project Acronym	OptEEmAL		
Project Title	Optimised Energy Efficient Design Platform for Refurbishment at District Level		
Project Coordinator	Miguel Á. GARCÍA-FUENTES (miggar@cartif.es) Fundación CARTIF		
Project Duration	1st September 2015 – 28th February 2019 (42 Months)		
Deliverable No.	D1.3. Requirements and specification of geo-clustering data sets access module		
Dissemination Level	PU		
Status		Working	
		Verified by other WPs	
		Final version	
Due date	31/08/2016		
Work Package	WP1 – Stakeholders' involvement trough IPD design strategy and input data process		
Lead beneficiary	ACC		
Contributing beneficiary(ies)	ES, FUNITEC, NBK, TEC, TUC, UTRC-I		
DoA	Task 1.3 – Integration of existing external data sources: geo-clustering techniques		

Date	Version	Author	Comment
12/01/2016	0.1	Patricio Moreno (ACC)	Table of Contents
01/02/2016	0.2	El Hassan Ridouane (UTRC-I)	UTRC contribution
04/02/2016	0.3	Gianluca Mameli (ES)	ES contribution
08/02/2016	0.4	Maxime Pousse (NBK)	NBK contribution
19/02/2016	0.5	Patricio Moreno (ACC)	ACC contribution
29/02/2016	0.6	Patricio Moreno (ACC)	Final intermediate version including missing sections and modification related to internal reviews
06/05/2016	0.7	Gianluca Mameli (ES)	Updated table of contents
01/07/2016	0.8	Giuliana Mangiapia (ES) Ciro Caterino (ES) Mariano Folla (ES)	Redefinition of the document, updated table of contents and content added to introduction, unstructured data service and unstructured data access module functional architecture
03/08/2016	0.9	Pedro Fernandez-Orellana(UTRC-I) Patricio Moreno Montero	UTRC-I contributions

		(ACC) Maxime Pousse (NBK) Giuliana Rosa Mangiapia, Ciro Caterino, Mariano Folla (ES)	ACC contributions NBK contributions ES contributions
09/08/2016	1.0	Giuliana Rosa Mangiapia, Mariano Folla (ES)	Unify contents – draft deliverable
18/08/2016	1.1	Miguel García (CAR)	Review
22/08/2016	1.2	Mariano Folla (ES), Patricio Moreno (ACC)	Document revision and re-writing
29/08/2016	1.3	Gema Hernández (CAR)	Review
29/08/2016	1.4	Giuliana Mangiapia, Mariano Folla (ES), Patricio Moreno (ACC)	Document revision and re-writing
30/08/2016	1.5	Patricio Moreno (ACC)	Minor editing

Copyright notices

Table of Content

List of Figures

List of Tables

Abbreviations and Acronyms

Acronym	Description
DPI	District Performance Indicator
DQR	Data Quality Rate
ECM	Energy Conservation Measure
EPW	Energy Plus Weather (data format suitable for energy plus energetic simulation)
JSON	Java Script Object Notation
LCA	Life Cycle Assessment
LOD	Linked Open Data
NUTS	Nomenclature of Territorial Units for Statistics
OptEEmAL	Optimised Energy Efficient Design Platform for Refurbishment at District Level.
OWL	Web Ontology Language
QOTF	Query On The Fly
REST	REpresentational State Transfer
RDF	Resource Description Framework

Executive Summary

This document is the final version of the D1.3 “Requirements and specification of geo-clustering data sets access module”. Since in the intermediate version high level requirements have been presented, in this deliverable more specific requirements related to the technical side have been developed. From this point of view the objective is to go further and deeper into the concepts explained in the intermediate version document in order to fully specify the requirement set. This deliverable provides detailed information about the technical integration of geo-clustering data sets access module with the rest of the platform, as the main objective of this document is to provide all the concepts needed to support the development of this module in D1.4. For this purpose, the deliverable describes how the platform can answer automatically to climate and socio-economic queries, focusing on the following three aspects:

- The gathering, processing and integration in the tool of contextual data and metadata.
- The definition of the functional architecture for the geo-clustering data access module.
- The definition of the criteria to be used to assess data quality and the minimum requirements for data quality.

The document has been structured to accomplish the above mentioned goals. Since the contextual data that have to be gathered by the geo-clustering data sets access module can be structured or unstructured data, the sections 2 and 3 introduce the services used for this purpose. In particular, section 2 presents the services that will be used to gather information from external databases related to two aspects that are going to be of paramount importance for the calculation of the DPIs: climate and socio-economic data, while section 3 presents the service which provides a search engine for gathering unstructured information from the web to support the platform users in their refurbishment projects. Section 4 specifies functional and non-functional requirements both for structured and unstructured data as also data requirements coming from the simulation. Section 5 is devoted to explain the high level design of the geo-clustering data sets access module named “Geo-clustering connector” which is going to be in charge of gathering the contextual data. The Geo-clustering connector consists of two different components: Geo Connector that gathers data from well-structured data sources, and Unstructured Data Connector that acts as an interface between the OptEEmAL graphical user interface and the Unstructured data service. The section presents the functional architectures of both the components and their integration into the OptEEmAL platform. Finally, section 6 introduces the Quality integration test needed to make possible the calculation of energy (for climate data) and socio-economic (for energy prices and income levels) DPIs with a sufficient level of quality.

1 Introduction

1.1 Purpose and target group

The purpose of this document is to set the requirements that the external data gathering system is going to have. This system is going to be divided into two well differentiated parts; one for the structured data that is going to be called “Geo Connector” and other for the unstructured data that will be called “Unstructured Data Connector”. The two parts are so differentiated that will not share any software part but they are described together since both of them are components that will gather information from external sources. The geo connector will be in charge of retrieving information related to climate, weather, energy prices and incomes, while the unstructured data service will provide the user with complementary information to be deployed at the objective-definition stage. The information gathered by the first component will be used both for the DPI calculation as for the identification of more adequate ECMs depending on the location; while the complementary information obtained by the second component will cover several topics that could influence the objective definition.

To make the description of both components understandable, this document starts with a description of how the information will be gathered using the Geo Connector for the structured part, describing the Urls that have to be formed to call the REST service. It continues with a description of the functionality that will be provided by the “Unstructured Data Service” since, in this case, a prototype of the software is already developed and thus no description of the architecture or the internal details of it is needed (nevertheless it was done in the first iteration) but it needs the whole picture to integrate the module into the rest of the platform.

Next, the specification of the requirements is done dividing them into “functional” (related to functionalities) and “non-functional” (related to performance and behaviour of the system) for both structured and unstructured data. An evaluation of the data that can be obtained using the methods described in this document is done to ensure the usability of the sets.

1.2 Contributions of partners

The following Table 1 depicts the main contributions from participant partners in the development of this deliverable.

Table 1: Contribution of partners

Participant short name	Contributions
ES	Main contributor of Section 1 (Introduction), 3 (Unstructured data service), 4 (Specification of Requirements), 5 (Design of the geo-clustering connector) and 7 (Conclusions)
ACC	Main contributor of Section 1 (Introduction), Section 2 (Structured data service), 4 (Specification of Requirements) and 7 (Conclusions)
UTRC	Contributor of Section 2.2 (External data sources to be accessed)
NBK	Main contributor of Section 6 (Quality integration test)
FUNITEC	Contributor of Section 5 (Design of geo-clustering connector)
TUC	Contributor of Section 6.3 (Application to structured data sources)

TEC	Contributor of Section 6.3 (Application to structured data sources)

1.3 Relation to other activities in the project

The following Table 2 depicts the main relationship of this deliverable to other deliverables developed within the OptEEmAL project and that should be considered along with this document for further understanding of its contents.

Table 2: Relation to other activities in the project

Deliverable Number	Contributions
D1.1	E-Guide on stakeholders' involvement and IPD implementation for the design and execution
D1.2	Requirements and specification of input data process to evaluate users' objectives and current conditions
D1.4	Geo-clustering data sets access module
D2.1	Requirements and specification for the District Data Model
D2.3	Functional architecture of the data repository
D5.2	Functional architecture Interfaces Overall Platform Design

2 Structured data service

2.1 Objectives

The main purpose of this service is to gather information from external databases about two aspects that are going to be of paramount importance for the calculation of the DPIs. The two data categories are:

- Climate data to calculate energetic demands. Weather typical data, which is related to climate and doesn't vary along time, will be deployed since there is software in the platform that is going to make use of it such as Energy Plus. This tool uses hourly weather typical data so that it will be needed to use also weather typical data extracted from a reliable source. Weather and climate are different, while the climate is supposed almost static, weather varies over time (it is dynamic). This is the reason why the usual weather databases are not sufficient for the simulations, the average weather over a year is needed which is not what usual weather websites offer, to have a good idea of what environmental conditions the buildings will have to deal with.
- Energy price data. It will be used to calculate the costs associated with the energy consumption (and thus, also to calculate payback periods). This value varies over time so that a good update and reliable source is needed which differentiates between, at least, gas and electricity prices. If this information is not available, it will be asked to the users, through a graphical user interface.
- Income levels will be used to calculate the energetic poverty rate of the habitants of the district.

There are two possibilities to use structured information inside the platform; to retrieve it "completely on demand", that is, not to store the information in any internal repository, or to store the information to have it "at hand" when it is needed. The second option is preferred because of the speed in which it could be deployed (it is much faster to retrieve the information from an internal database than from an external website and this process will be done several times in the same project), so that the possibility of retrieving the information just one time per project is a desirable feature. The module will put the information in a RDF file format associated with an OWL file that will contain the ontologies.

The information gathered in this process will serve on the one hand for the simulation module in order for the external tools invoked by platform to deploy the necessary data for the DPI calculation. And also for the identification of ECMs that can be directly affected by the data obtained by the geo-clustering service and that can be identified as adequate or not for a certain location.

2.2 External data sources to be accessed

In the table below it is possible to find the two sources that the structured data service is going to deal with. In bold are indicated those values that are mandatory top run the simulations needed in the DPI calculations. Those that are not in bold and are included in the same data source can be retrieved using the same procedure but it is envisioned not to be needed.

Table 3: External services to be used for the structured data gathering system (in bold those mandatory for the simulations)

System name	URL	Application zone and other considerations	Measure(s) offered
European Commission Eurostat GISCO	http://ec.europa.eu/eurostat/web/gisco/geodata The method for extracting the values is depicted in: http://ec.europa.eu/eurostat/web/json-and-unicode-web-services/about-this-service	Eurozone	Administrative units (NUTS, Urban Audit, Countries, communes…)
			Population Distribution /Demography
			Transport networks
			Land cover
			Digital Elevation Model(DD LAEA)
			Aspect Slope,
			Coloured relief
			Hill shade
			Hydrography
			Energy prices (euro)
			Income level (histogram)
Energy Plus Weather website	https://www.energyplus.net/weather	The entire world	**Dry Bulb Temperature (ºC)**
			Relative Humidity (%)
			Barometric Pressure (Pa)
			Wind Speed (m/s)
			Wind Direction (degrees)
			Global Horizontal Solar Radiation (W/m²)
			Cloud Cover (tens)
Ge20	http://www.geoclusters.eu/ge20/	Eurozone KML format	Heating Degree Days
			Cooling Degree Days
			Annual incident energy on a south orientation
			Avg ambient temperature over year
			Average Cooling Seasonal External Air Temperature
			Average Heating Seasonal

			External Air Temperature
			Maximum ambient temperature over year
			Annual Average Ground / Water Temperature
			Average Heating Seasonal Ground / Water Temperature
			Average Cooling Seasonal Ground / Water Temperature
			Average ambient wet bulb temperature over cooling season.
			Average ambient temperature during daylight over cooling season.
			Average solar irradiation during daylight over cooling season on a south oriented plane with a 45° slope.
			Age of construction
			Use residential
			Use residential singe
			Use residential apartment flats
			U value existing –wall
			U value existing –roof
			U value target –wall
			I U value target –roof
			U value target –floor
			Heating System – central
			Heating System - Individual
			Population living in the area at last census.
			Gross domestic product (GDP)
			Gross domestic product in construction
			Employment rate
			Employment in construction
			Labour cost

			Gas prices for household consumers
			Electricity prices for household consumers
			Disposable income of households
			Electricity consumption of households
			National Energy Regulations
			Solar Cooling Incenties
NOAA / NWS's Meteorological Development Laboratory	http://weather.noaa.gov/pub/SL.us008001/ST.opnl/DF.gr2/DC.ndfd/AR.nhemi/ http://weather.noaa.gov/pub/SL.us008001/ST.opnl/DF.gr2/DC.ndfd/AR.nhemi/ Method explained in: http://www.nws.noaa.gov/mdl/degrib/dataloc.php	Northern Hemisphere It is possible to get historical records Grib format (condensed one)	Wind speed 10m kts
			Wind gust kts
			Pressure (MSL)
			Temperature
			Dew Point (2m) Min
			Dew Point (2m) Max
			Precipitation mm/h
			Cloud cover
			Relative humidity (2m)
			Isotherm 0ºC
			Snow depth cm
			CAPE J/Kg
			CIN J/Kg
Dermap	http://www.dermap.com/en/social-gis.html#	Europe	Efficiency Index
			Innovation Index
			Budget Import-Export
			Density of Population
			Price of eggs
			18 years old Students in Europe
			Female Employment
			Population
		World	Public Debt

			Control of corruption (control score, 1 to 10 corruption scale
			Longevity difference in the world
			Life expectancy in the World
			Agricultural land in the World
			Rooms per capita in the World

The list provided in the first iteration of the document about the data sources that are going to be used has been modified because since then, an in-depth study carried out identified some wrong assumptions that were made in the first iteration. The wrong assumptions were:

- Data needed were thought to be weather, but they must be climate related (difference explained in section 2.1).
- Ge20 REST system is not installed neither updated.

The options taken are described in the following sections.

2.2.1 Energy Plus Weather

The Energy Plus website has a database in which it is included representative climate data sets (of each day of the year) for Europe. A regular REST call is required to transform the files where the needed information is contained to the proper format. The aggregation level that can be reached is NUTS 2. The url that is going to be used to do the REST call has to be made using the following steps:

1. There is a part of the url that is common to all cases, it is:

   https://www.energyplus.net/weather_location/europe_wmo_region_6

2. It is needed to add one specific part for the country, it has to be done using the *ISO 3166-1 ALPHA 3* code system:

   https://www.energyplus.net/weather_location/europe_wmo_region_6/ESP

3. The next part of the url is the one related to the region. To obtain it, it will be needed to add a string that is specified in the webpage (just by clicking in the previous link it is possible to see the complete list).

   https://www.energyplus.net/weather_location/europe_wmo_region_6/ESP/ESP_San_Sebastian.080270_SWEC

4. In the last step it is needed to add the format in which we want the data. The most useful for us will be the epw, which is a plain text file with a well stablished format.

   https://www.energyplus.net/weather_location/europe_wmo_region_6/ESP/ESP_San_Sebastian.080270_SWEC.epw

Once the url is formed, the file can be retrieved using a simple REST call like:

wget
https://www.energyplus.net/weather_location/europe_wmo_region_6/ESP/ESP_San_Sebastian.080270_SWEC.epw

2.2.2 Eurostat

Based on D1.2 section 6.3 the minimum socio-economic data required is energy cost and income level per inhabitant. One of the services identified in section 3.1 is http://ec.europa.eu/eurostat

which provides energy price data as required in KW/hour € per country and income level in inhabitant income € per NUTS-2.

To retrieve data programmatically Eurostat provides several REST services that can be called following the format depicted in Figure 1:

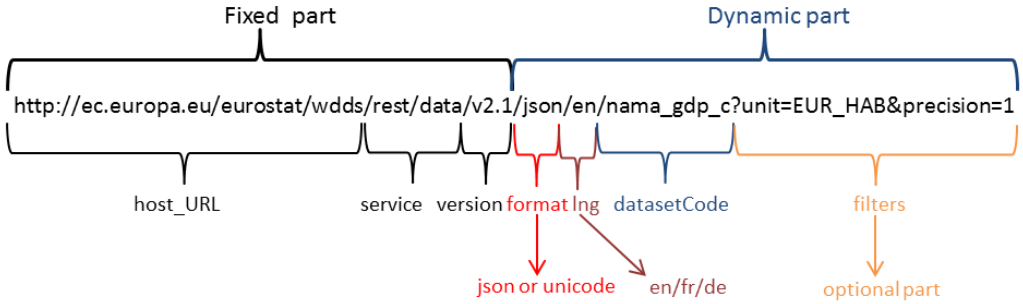

Figure 1: Eurostat REST request

A detailed explanation of this request can be found at link below:

http://ec.europa.eu/eurostat/web/json-and-unicode-web-services/getting-started/rest-request.

Sections "Retrieving Energy Prices" and "Retrieving Income Level" below explain in more detail which datasets will be used and how to infer all the parameters needed for the request. The list of datasets available in Eurostat is defined in http://ec.europa.eu/eurostat/data/database.

2.2.2.1 Retrieving Energy Prices

To retrieve energy prices is necessary to know the country where the design is made; this information should be defined already in OptEEmAL. The query format for countries requires the following ISO 3166-1 alpha-2 code representation. Country codes for OptEEmAL case studies are: Spain ES, Turkey TR, Unite Kingdom UK, Italy IT, and Sweden SE.

The datasets to query energy price are:

1. Gas Price for domestic consumers– nrg_pc_202
2. Electricity Price for domestic consumers – nrg_pc_204

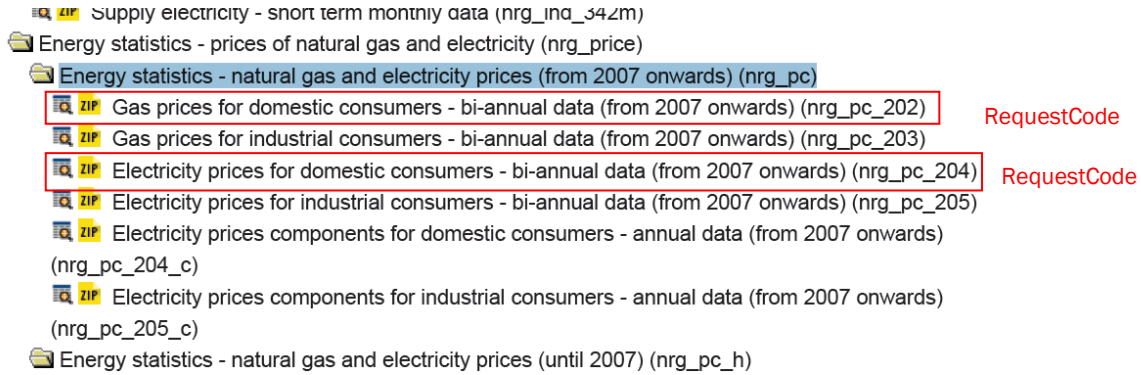

Figure 2: Capture of datasets list - Energy Statistics

The simplest possible query where only the dataset is specified (no filters applied) is:

http://ec.europa.eu/eurostat/wdds/rest/data/v2.1/unicode/en/nrg_pc_204

http://ec.europa.eu/eurostat/wdds/rest/data/v2.1/unicode/en/nrg_pc_202

The response received follows the next format:

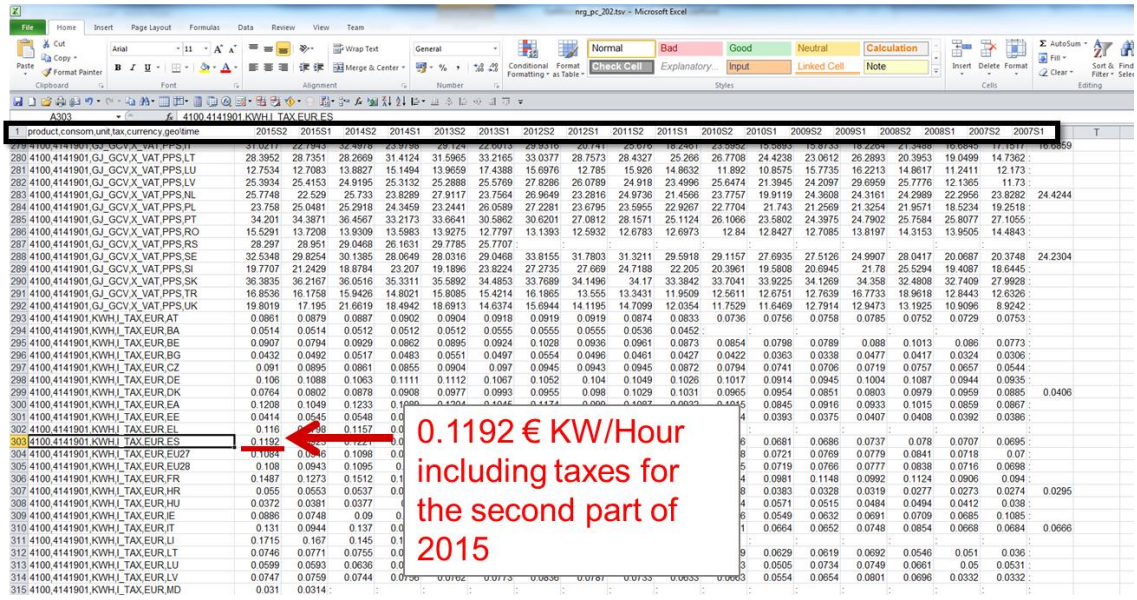

Figure 3: Eurostat gas price response

2.2.2.2 Retrieving Income Level

As per income level, Eurostat provides a wide range of data for income base on different metrics. The one that best fits OptEEmAL requirements is “income of households by NUTS 2 regions”.

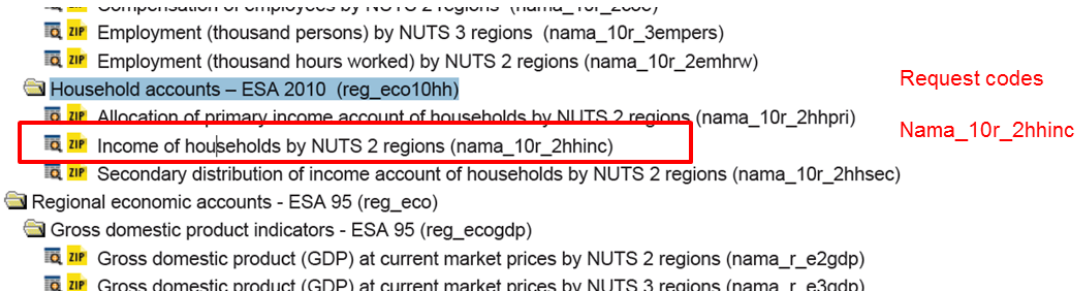

Figure 4: Capture of datasets list - Energy Statistics

To filter this data, the NUTS 2 identifier based on building location is necessary. Eurostat provides information that relates address with NUTS level per country. These data can be retrieved from http://ec.europa.eu/eurostat/tercet/flatfiles.do . The specific file per country can be retrieved programmatically being defined the country code. As an example, the following figure shows how the NUTS 2 code can be obtained for one of the case study (Historic city District, Santiago de Compostela, Spain).

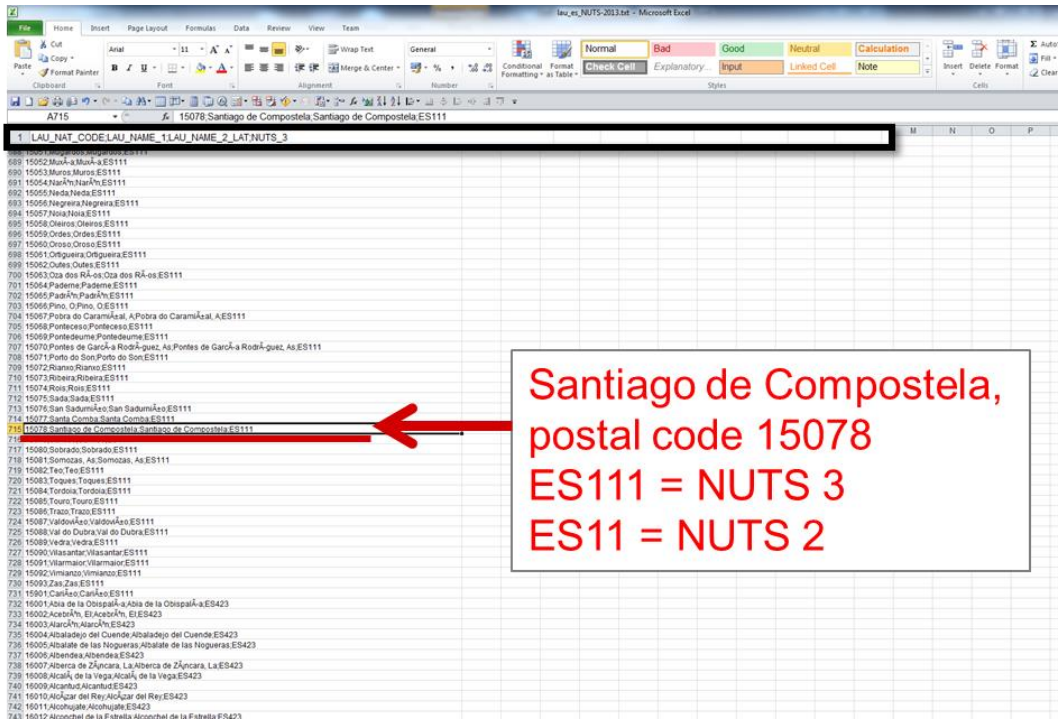

Figure 5: File obtained from Eurostat and codes

http://ec.europa.eu/eurostat/tercet/download.do?file=pc_es_NUTS-2013.txt

Further information about NUTS can be obtained from:

http://ec.europa.eu/eurostat/web/nuts/correspondence-tables/postcodes-and-nuts

The simplest possible query where only the dataset is specified (no filters applied) is:

http://ec.europa.eu/eurostat/wdds/rest/data/v2.1/unicode/en/nama_10r_2hhinc

The response received follows the next format:

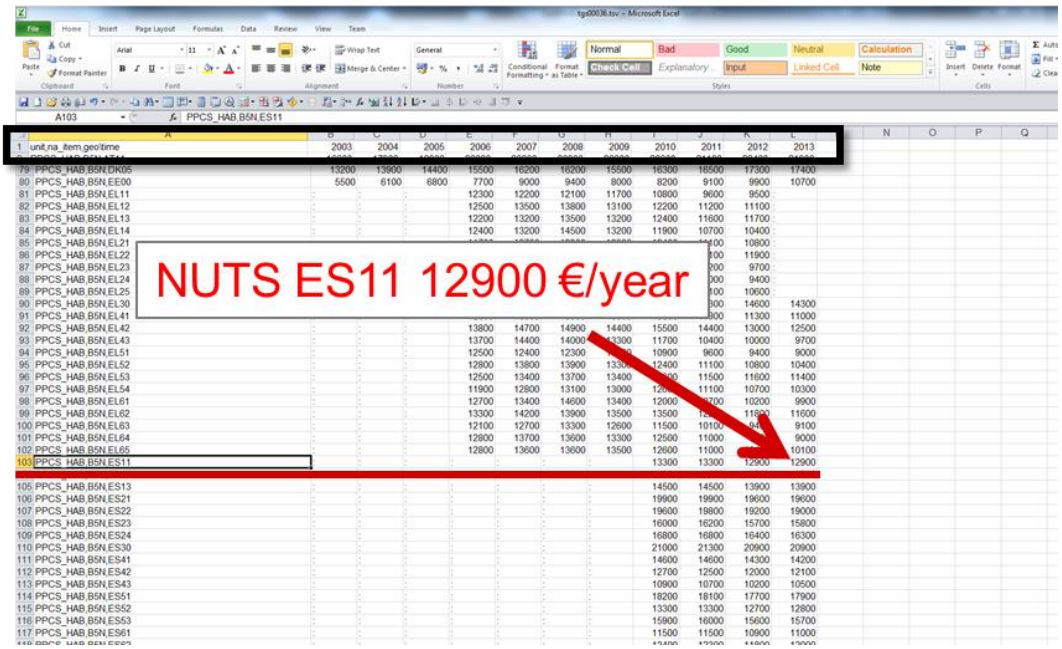

Figure 6: File obtained from Eurostat and codes

2.3 Conclusions for structured data gathering module

There is some confusion about what weather and climate data are, in the description of actions of this project too. This confusion is due to the fact that both concepts are related to the same phenomenon, the environmental conditions. To make a proper simulation, it will be needed to use data that is representative of a specific geographical location. Being representative does not mean to use real weather but to use data that is probabilistically the median of the measures, not the average. This concept is better called as “climate data” using the common definition of it but clearly it is on the border of the two concepts. Fortunately, there is a service set of representative weather data for each day of the year and with the data format needed for the simulation since it is provided by the site that hosts the software, www.energyplus.net.

All socio-economic data required can be retrieved from Eurostat programmatically as defined in this section. It is also possible to get energy prices that can be retrieved at country level as required in D1.2 by the DoA, income level per inhabitant is extracted at NUTS 2, what provides a better granularity than the required.

3 Unstructured data service

In order to support the end-users of the platform in their refurbishment projects, this service, named *Query On The Fly (QOTF)*, provides a search engine for unstructured data that allows them to find useful information in the web. This information will be deployed at the objective definition stage as an external service to aid the user in their decision-making process. The search engine helps users of the platform to identify the correct information in different contents that will be retrieved from the Web to support the evaluation of the current scenario also using a set of semantic metadata to filter the information. It will be possible, for example, for a user to insert in the search box the term "average U-value" and find 100 related documents then, using the semantic metadata as a filter, it will be possible to select a specific "location" reducing to 25 the number of related documents, then to add an "organization" and so on until the user finds exactly the 2-3 useful documents he was looking for.

The query on the fly search allows the user to experience a new way of surfing the web. By applying semantic analysis on the fly to search engines results (Google, Bing and Yahoo), the user is able to filter and explore the vast quantity of results near real-time, finding the information faster and with increased accuracy. With this tool it will be possible, for example, to find quickly the prices of the houses in a particular area or the average U-value in that area.

A district analysis can be performed with the following steps:

1. The user inserts a query in the search box, selects one or more search engine, the language of the information he would like to search and how many documents to consider in the analysis.
2. The QOTF acquires the results from the search engines, analyzes the snippets (produced by the search engines) and represents them semantically.
3. In the left column of the GUI the main semantic tags (topics, entities, relationships, emotions, etc.) have been published. Now, the user can explore the data of interest, select one of the tags to add a new constraint to the search or hide some results he is not interested in.
4. The user can also select one of the tabs at the bottom of the page (Topics, People, Domain Entities, etc.) for another view of the content that can offer a useful and alternative visualization of the information.
5. The user can extend the search to other semantic engines at any time by simply clicking on the specific check box in the search bar. In this scenario new data will be acquired and indexed.

3.1 Usage examples

To have a better understanding of the utility of the unstructured data service, some examples have been reported. Section 4 of D1.2 contains a basic description of the six case studies considered in the project, selected to validate the OptEEmAL platform. These six case studies reflect a decision-making scenario into which the OptEEmAL platform can be used to ease decision making. In the earlier phase of the project it was necessary to gather information about all the case studies regarding the initial situation of the district, the goal of future platform users, all available data, etc. In addition to these data, it can be useful to exploit the presented QOTF service, to increase knowledge about the district, gathering via web, news and topical information related to the district.

In the following figure is reported a usage example of the QOTF for the case study 1. "Cuatro de Marzo"; as shown, the QOTF results report information about "alternative energy" or "renewable energy" related to the analysed district.

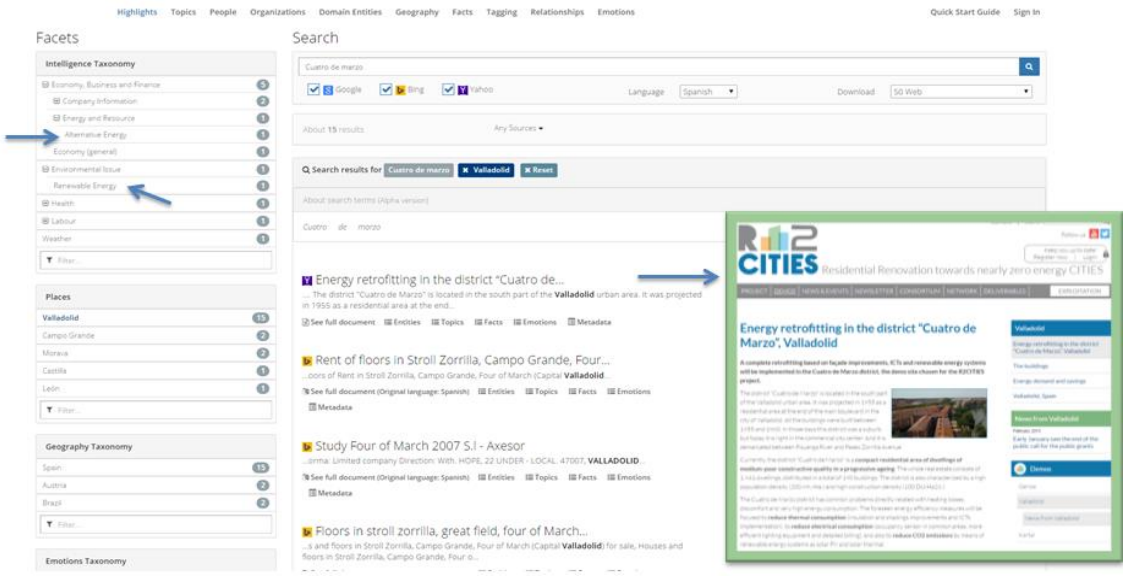

Figure 7: QOTF results for Cuatro de Marzo District

In the figure below is reported a usage example of the QOTF for the case study 6. "Sneinton District – Nottingham"; the QOTF results report information about "critical infrastructures" and "civil infrastructures" related to the analysed district.

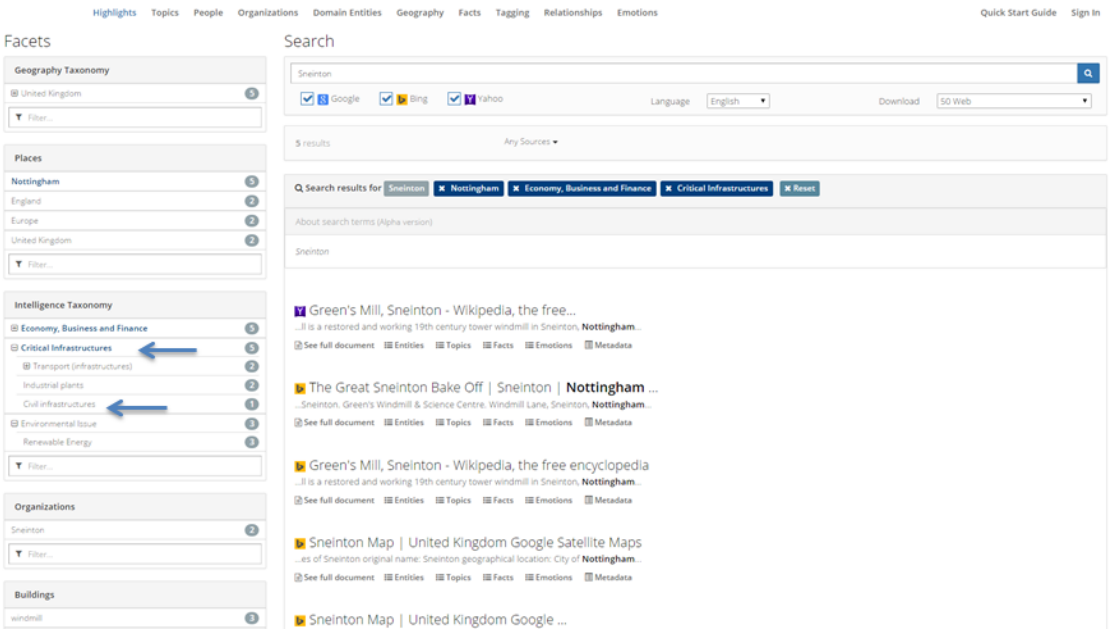

Figure 8: QOTF results for Sneinton District

4 Specification of Requirements

4.1 Data requirements coming from the simulation

Since weather data will be retrieved exclusively from the EnergyPlus weather database, they will comply with the EnergyPlus Weather (.epw) format file, an ASCII, csv format file containing the hourly or sub-hourly weather data needed by the EnergyPlus simulation tool. However, these data are going to be used by other simulation tools as well, where the weather format file may differ. For instance, CitySim requires the weather data to be provided according to the Climate (.cli) format file, starting with a header that contains the city, and its geographical position, followed by the meteorological data for a year, organized by day, month and hour.

Hence, as mentioned above, in the attempt to store the data retrieved in a standard (unified) format, the weather data will be converted to an RDF format file, to be stored in the contextual repository through the corresponding connector of the communication logic layer (see figure 9). Once the information has been stored in the repository, the data management module maps the information in order to create one instance of the baseline scenario. The weather data retrieved are part of the simulation data models generated by the data integration component and stored in the project repository.

This information will be retrieved by the simulation model input generator module which will be in charge of configuring the simulation files and launching the simulation tools. Concerning the weather data retrieved, the simulation files configuration refers to the respective weather files generation, where certain transformation rules (refer to D4.4: Requirements and design of the simulation model input generator module, section 6.2.2) will be introduced to convert RDF format files to EPW, cli or other format files.

Table 4 below provides the relationships between data from EnergyPlus Weather and the ECM catalogue as an example of how the structured data will be used by the platform. Also, after the table, the relationship between EUROSTAT energy prices and the OptEEmAL economic assessment is provided.

Table 4: Relation between the data obtained by the EnergyPlus weather database and the simulation of the Energy Conservation Measures of the OptEEmAL ECM Catalogue

	Dry Bulb Temperature (°C)	Relative Humidity (%)	Barometric Pressure (Pa)	Wind Speed (m/s)	Wind Direction (degrees)	Global Horizontal Solar Radiation (W/m²)	Cloud Cover (tens)
A1.1-Ventilated Facade	X	X	X	X	X		
A1.2-ETICS – External Thermal Insulation Composite System	X	X	X	X	X		
A2.1-Protected Wall – Internal lining	X	X	X	X	X		
A3.1-Cavity wall – Air chamber insulation	X	X	X	X	X		
B1-Roof	X	X	X	X	X		
C1-Slab	X	X	X	X	X		
D1 – Window replacement	X	X	X	X	X	X	X
Biomass boiler							
ST-Flat collector						X	X
ST-Tube collector						X	X
PV-Mono-crystalline						X	X
PV-Multi-crystalline						X	X
Wind turbines					X		
Geothermal-horizontal	X	X					
Geothermal-vertical	X	X					
High efficient boiler							
Condensation boiler							
Cogeneration							
High efficient chiller (electricity)							
High efficient heat Pump							
Energy storage-water tank	X	X					
Energy storage-Phase change materials units	X	X					
Reduction of distribution losses	X	X					
System Scheduling							
Optimal start-up shut down							
Weather compensation							
Load following							
Optimization-based control							

Regarding the economic assessment, the database of Eurostat is going to be used to define the Energy Price (EP) economic parameter. In this way, the following equation will make possible the calculation of the initial and refurbished operational stages economic impact:

$$B6\_ec = \left(\frac{ED_b}{\rho} - RE_k\right) \cdot \frac{EP_y}{FU}$$

Where:

- B6_ec → Baseline or refurbished building operational energy use stages economic impact (€)
- ED_b → Baseline or refurbished building operational annual energy demand
- ρ → Performance of the energy generation system (%)
- RE_k → Renewable energy generated by refurbishment strategies
- EPy → Price of the energy source (€/kWh)
- FU → Functional Unit of the case study

The first factor refers to the energy demanded from the external net (without internal energy renewable sources) and the second to the price the energy has. Taking this into account, this equation can be identified with the one provided in D2.2 for the District performance indicator called ECO 01:

$$Op.energy\ cost = Energy\ consumption\ (kWh/year)\ \ x\ energy\ cost\ (€/kWh)(per\ fuel)$$

4.2 Functional requirements definition

4.2.1 Structured Data

Table 5: Functional requirements for structured data

REQUIREMENT ID	DESCRIPTION
FR1	The system has to be able to set up a url like it is described in the section Energy Plus Weather for the place in which it is located the district.
FR2	The system has to be able to set up a url like it is described in the section Eurostat for the place in which it is located the district.
FR3	The system has to be able to retrieve data from the urls formed using the FR1 and FR2.
FR4	The system has to be able to format data gathered using the FR3 for its utilization by the scenarios evaluation in a RDF format.

4.2.2 Unstructured Data

Table 6: Functional requirements for unstructured data module

Requirement ID	DESCRIPTION
FR1	The unstructured data system has to be able to extract semantic metadata from inside web sites.
FR2	The unstructured data module has to be able to be connected to the rest of the OptEEmAL platform in order to offer its results.

FR3	The unstructured data module should store the result of the searches into its repository.
FR4	The platform should show the users the results of the searches.

4.3 Non-functional requirements definition

4.3.1 Structured Data

Table 7: Non-functional requirements for structured data

REQUIREMENT ID	DESCRIPTION
NFR1	The system has to have high band connectivity to http://ec.europa.eu/eurostat as well as to https://www.energyplus.net.
NFR2	The system has to be able to process data gathered in a reasonable amount of time (less than 1 min).

4.3.2 Unstructured Data

Table 8: Non-functional requirements for unstructured data

REQUIREMENT ID	DESCRIPTION
NFR1	The system must be connected to Internet.
NFR2	The system must show the results of a search in few seconds (less than 1 minute).

5 High level design of the geo-clustering connector

The geo-clustering connector which is going to be in charge of gathering the contextual data consists of two different components:

- **Geo Connector**. This component gathers data from well-structured data sources, described in section 2.2.
- **Unstructured Data Connector**. This component acts as an interface between the OptEEmAL graphical user interface and the unstructured data service (*QOTF service*) described in the section 3. Through this component the users of the OptEEmAL platform can visualize the graphical user interface of the QOTF service and use its search tools.

As explained in the section 2.1, only the information coming from the structured part is going to be used in the optimization and simulation process performed by the platform, Therefore, these two components will be integrated into the platform separately as it is explained in the following section.

5.1 Integration into the OptEEmAL platform

The structured data retrieved from the **Geo Connector** have to be inserted and stored into the platform to be used in the simulations. Therefore, the **Geo Connector** component will be integrated into the Data Insertion Module of the platform, described in D5.2 section 3.1.1, as it is shown in the following diagram which explains the Data insertion and diagnosis process described in D5.2 section 3.1.

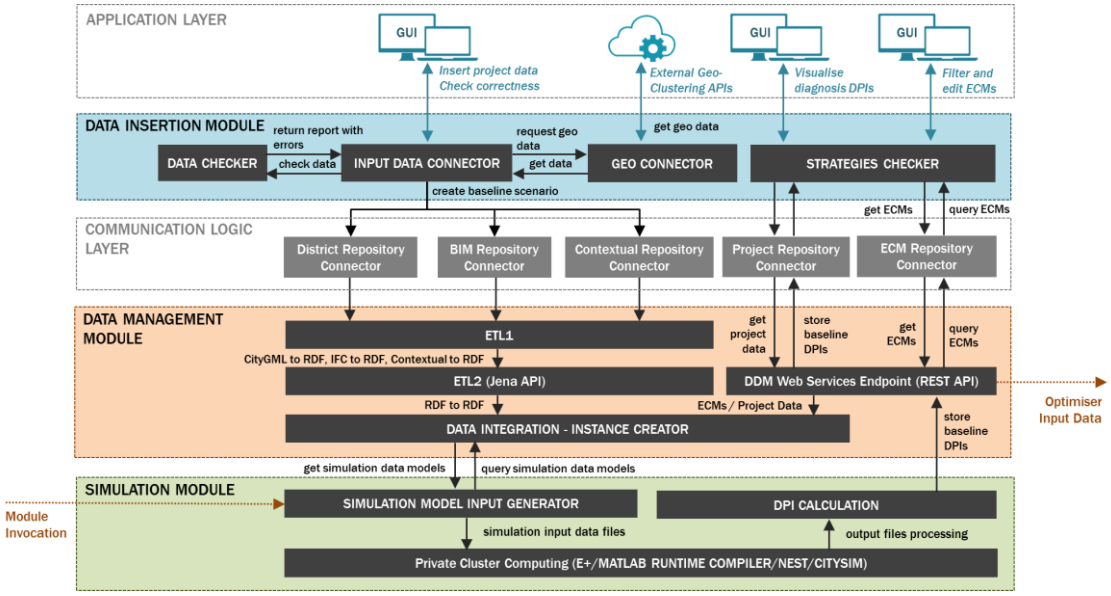

Figure 9: Geo Connector integration

The Unstructured data will not be used in the simulations, but it will be semantically analysed by a dedicated service to be shown to users of the platform in a useful way, as explained in section 3. This service will be developed as an external service (named QOTF service) and will provide users with a search engine through which they can search for useful information in the web. Therefore, the **Unstructured Data Connector** will be a component of the graphical user interface of OptEEmAL platform that will allow users to visualize the graphical user interface of the QOTF service, as described in D5.2 section 5.1.

The following diagram shows the integration of the Unstructured Data Connector into the graphical user interface of OptEEmAL platform and the process for gathering useful information from the web.

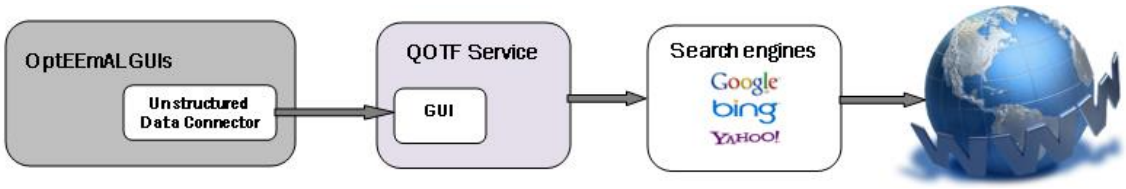

Figure 10: Unstructured Data Connector integration

The following sections explain the functional architecture designed for the Geo Connector component and the Unstructured data service (QOTF service), as the Unstructured Data connector will be a simple web page of the OptEEmAL GUIs (or even more simply a link/button in the project dashboard described in D5.2 section 5.3). Since the Unstructured data service will be an external service, it will be designed more deeply and developed in D1.4.

5.2 Functional architecture of the Geo Connector component

The Geo Connector component has the aim to gather the structured contextual data from external data sources, described in section 2.2. The data sets chosen as data sources allow requesting the available data through REST Services. Since the data retrieved from these services have to be stored in an RDF Store, that is the Context Repository described in D2.3 section 3.5.3, the Geo Connector has a specific sub-component (**Mapper**) that converts these data to RDF format, as it is illustrated in the following figure that shows the functional architecture of the Geo Connector.

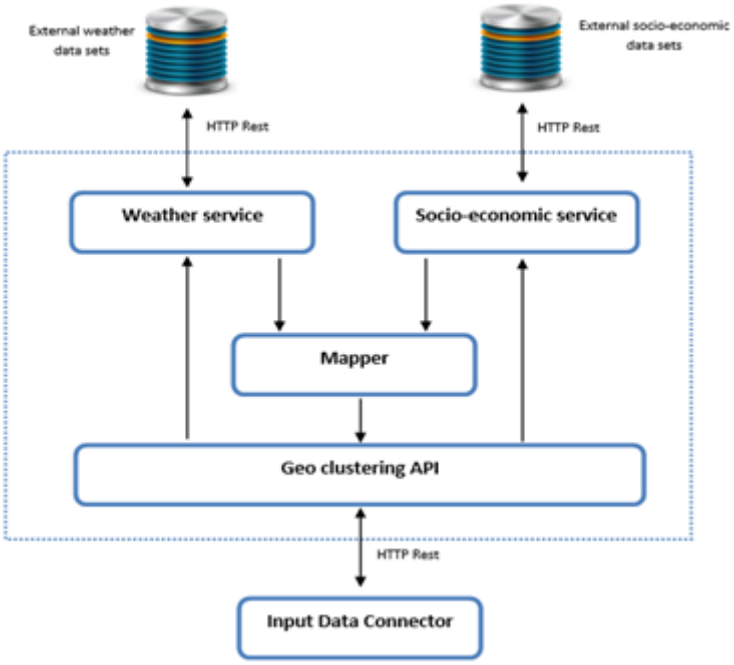

Figure 11: Functional architecture of the Geo Connector component

The Geo Connector component will technically be composed of the following sub-components:

- **Weather service**: this component allows requesting the weather data to the External weather data service (EnergyPlus) through its REST services. The REST calls to these services are explained in detail in section 2.2.1.
- **Socio-economic service**: this component allows requesting the socio economic data to the External socio-economic data service (Eurostat) through its REST services. The REST calls to these services are explained in detail in section 2.2.2.
- **Mapper**: this component parses the data retrieved from the external services in RDF format.

- **Geo clustering API**: this component provides the Application Programming Interface that will be used by the Input Data Connector component of the Data Insertion module, as it is depicted in the figure 9.

A class diagram for this component can be found in the D5.2 section 3.2.1.2, while its interaction with the others components of the Data insertion module of the OptEEmAL platform is described in the related sequence diagram in section 3.3.1 of the same document.

5.3 Functional architecture of the Unstructured data service

The following figure shows the functional architecture of the Unstructured data service.

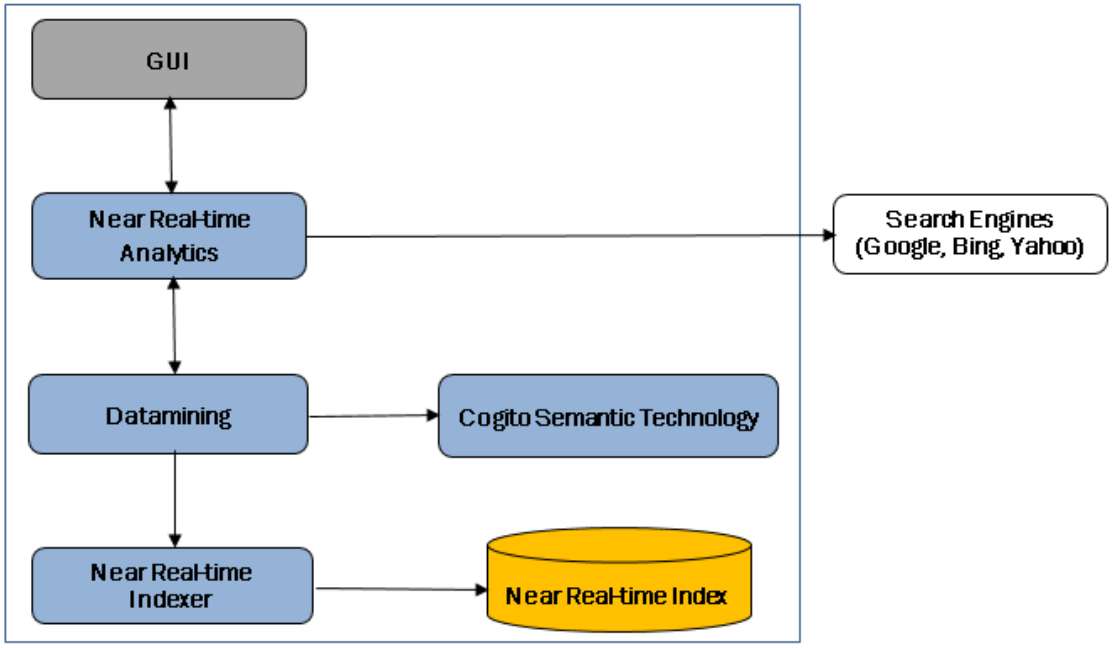

Figure 12: Functional architecture of the Unstructured data service

The component with grey background is the Graphical User Interface through which users can perform their searches and visualize results. The components with blue background are those that provide the functionalities of capture, analysis, manipulation, and indexation of unstructured information. In the meantime, the component with orange background stores some metadata of the retrieved information in order to support the near real time features.

The process is the following: The **Near Real-time Analytics** component is the entry point for user's requests; in a matter of seconds it acquires content, analyses it with **Cogito Semantic Technology** and joins together the many results (also from different search engines). To do this, it uses the **Datamining** component to elaborate the request of document analysis. The results are indexed and stored through the **Near Real-time Indexer** component.

The Unstructured data service is composed of the following components:

- **GUI**: the graphical user interface allows users to experience a new way of surfing the web. By applying semantic analysis on the fly to search engines results (Google, Bing and Yahoo), the user is able to filter and explore a vast quantity of results near real-time, finding the information faster and with increased accuracy. As a part of the faceted navigation capability, the user has the collected information grouped in different tabs (Topics, People, Domain Entities, etc.), allowing to have a different view of the content that can offer a useful and alternative visualization of the information.

- **Near Real-time Analytics:** this component is designed to expose the Query-on-the-fly to users. In a matter of seconds, Query-on-the-fly acquires content, indexes it with Cogito Semantic Technology and joins together many results (also from different search engines).
- **Datamining**: this component elaborates the requests of document analysis and uses the Cogito semantic analysis features to discover useful knowledge in the search results.
- **Cogito® Semantic Technology:** this component represents the Expert System's software platform for document analysis. Cogito® via its ESSEX service exposes a set of simplified interfaces which allow for coordinated access to the base analysis capabilities, categorization and extraction. These capabilities are explained in detail in the first iteration of D1.3 section 3.2.
- **Near Real-time Indexer**: this component indexes the analysed results. It will be implemented with Apache Solr, that is an open source enterprise search platform, written in Java, from the Apache Lucene project. Its major features include full-text search, hit highlighting, faceted search, real-time indexing, dynamic clustering, database integration, NoSQL features and rich document (e.g., Word, PDF) handling. Providing distributed search and index replication, Solr is designed for scalability and Fault tolerance.
- **Near Real-time index**: This component represents a repository for storing the index generated by the indexer.

6 Quality integration test

Data quality is an essential parameter while using datasets from different origins. Indeed, an assessment of the quality of those external data is needed to ensure that the envisaged data does not reduce the quality of the analysis. This is well reflected in the ISO 14044 standard related to Life Cycle Assessment (LCA) which defines data quality as "characteristics of data that relate to their ability to satisfy stated requirements" [1]. In the frame of the OptEEmAL platform, as geo-clustering techniques will be used to complement the data model, it is essential to ensure that those data have the ability to satisfy stated requirements before their integration in the platform. In OptEEmAL, the requirement for the geo-clustered data is to make possible the calculation of energy (for weather data), economic (for energy prices) and social (for average incomes) DPIs with a sufficient level of quality.

Data quality is usually defined through a set of different indicators or criteria [2] [3] whatever the application domain is. For instance, in public health (where data quality is of crucial importance), a review of data quality assessment studies shows that up to 49 different criteria can be used to assess data quality [3]. In the frame of the OptEEmAL platform, it will be necessary to define the relevant criteria which will be used to assess data quality. Also, in order to make the data quality assessment procedure transparent and easier to apply in the platform, a quantitative assessment method for data quality is required.

As a consequence, the following overall methodology is proposed to ensure the data quality of geo-clustering data:

- To define relevant criteria for the OptEEmAL platform and for each of this criterion, elaborate the relevant assessment matrix.
- To set a minimum requirement of data quality.
- To define a process to check the quality of the envisaged geo-clustering data sources (both for structured and unstructured geo-clustering data sources) and validate if they comply with the minimum requirement.

6.1 Criteria used to assess data quality

The number of criteria used to assess data quality is very different depending on the application sector as well as on the data origin. For instance, as indicated above, up to 49 criteria can be used in public health when assessing data quality. When investigating the specific case of Linked Open Data (LOD), it has been revealed that data quality can be represented with a maximum of 26 dimensions [4]. Finally, in LCA, where data quality related issues have been studied for several years and quantitative data assessment is commonly performed, it is widely accepted that input data quality can be evaluated using 5 different criteria [5]. It has to be mentioned that in the field of building energy simulations, data quality is a current area of concerns but is less advanced than in LCA for instance, without any data quality assessment procedure commonly agreed [6].

The analysis of the different criteria used in the different sectors reveals that some common criteria are used to define the quality of a given data or datasets. Those criteria are:

- Geographical Representativeness (GR)
- Time Representativeness (TiR)
- Technological Representativeness (TeR)
- Accuracy (also called "Precision/Uncertainty) (A)
- Completeness (C)
- Reliability/Credibility (R)

In the frame of the OptEEmAL project, two different and already defined geo-clustered data sources will be used: one for weather data and the other one for energy prices and income levels. Considering that these data sources are already known, it is proposed, for efficiency purpose, to perform an a priori manual assessment of these data sources in terms of data quality.

In case that the expected data sources are either not known or expected to change within the project, as it will be the case with unstructured data sources, it will be needed to implement an automatic data quality evaluation process. This is a more time consuming process and it implies the gathering of metadata associated with the data of interest. However, the methodology presented below will be exactly the same.

In addition, it is needed to establish a transparent data quality assessment methodology as it will make possible the comparison between different data sources and also the evaluation, in the future, of potential new data sources and the comparison of them with an already used one.

Considering the abovementioned requirements, the definition of a data quantitative assessment methodology in a transparent manner is needed. To do so, it is proposed to use as a basis the approach for semi-quantitative assessment of overall data quality proposed in the EC Product Environmental Footprint Guide [7]. This approach aims at defining for each criteria used in the data quality assessment process, a scoring and associated definition onto which each different data will be evaluated. Within the OptEEmAL project, it is necessary to adapt this approach for the selected criteria. This work is presented below for the different criteria to be used in OptEEmAL namely:

- Geographical Representativeness (Table 9)
- Time Representativeness (Table 10)
- Accuracy (also called “Precision/Uncertainty) (Table 11)
- Completeness (Table 12)
- Reliability/Credibility (Table 13)

Table 9: Definition of data quality assessment levels for Geographical Representativeness (GR).

Criteria	Definition	Score	Level	Associated values
Geographical representativeness	Degree to which the dataset reflects the true population of interest regarding geography	1	Very good	Local (Municipal data)
		2	Good	Regional
		3	Fair	National
		4	Poor	Continental
		5	Very poor	International; Unknown

Table 10: Definition of the data quality assessment levels for Time Representativeness (TiR).

Criteria	Definition	Score	Level	Associated values
Time representativeness	Degree to which the dataset reflects the specific conditions of the system being considered regarding the time/age of the data	1	Very good	2012 – 2015
		2	Good	2010 – 2012
		3	Fair	2005 – 2010
		4	Poor	1995 – 2005
		5	Very poor	<1995; Unknown

Table 11: Definition of the data quality assessment levels for Accuracy (A)

Criteria	Definition	Score	Level	Associated values
Accuracy	Qualitative expert judgement	1	Very good	Very low uncertainty
		2	Good	Low uncertainty
		3	Fair	Fair uncertainty
		4	Poor	High uncertainty
		5	Very poor	Very high uncertainty

Table 12: Definition of the data quality assessment levels for Completeness (C)

Criteria	Definition	Score	Level	Associated values
Completeness	To be judged with respect to the coverage of each data source and in comparison, to a hypothetical ideal data quality	1	Very good	Very good completeness
		2	Good	Good completeness
		3	Fair	Fair completeness
		4	Poor	Poor completeness
		5	Very poor	Very poor or unknown completeness

Table 13: Definition of the data quality assessment for Reliability/Credibility (R)

Criteria	Definition	Score	Level	Associated values
Reliability/Credibility	Degree to which the data source is coming from a reliable/credible institution	1	Very good	Widely acknowledged institution
		2	Good	Acknowledged institution
		3	Fair	Fairly acknowledged institution
		4	Poor	Poorly acknowledged institution
		5	Very poor	Unknown institution

6.2 Minimum requirement for data quality

After the definition of such assessment tables for each criterion, it will be necessary to define the minimum requirement for data quality. This requirement will be the threshold that will be used to state whether a given data source can be used in the OptEEmAL platform or not.

To define this minimum requirement, it is proposed to set an overall minimum requirement. This overall minimum requirement is obtained through the aggregation of each criterion in a common data quality indicator. To do so, it is proposed to also use the methodology proposed in the EC Product Environmental Footprint Guide and presented in [8], which defines the following formula to calculate an aggregated data quality indicator called "Data Quality Rating (DQR)".

$$DQR = \frac{Criterion_1 + Criterion_2 + Criterion_n}{n} \quad (1)$$

With n = number of criteria

The adaptation of this formula in OptEEmAL gives:

$$DQR = \frac{GR + TiR + A + C + R}{5} \quad (2)$$

Then, using a quality scale provided in [8], it is possible to assess the overall data quality level of a given dataset (Table 14).

Table 14: Overall data quality level according to the achieved aggregated data quality indicator

Data Quality Rating	Overall data quality level
=< 1.6	"Excellent quality"
1.6 to 2.0	"Very good quality"
2.0 to 3.0	"Good quality"
3.0 to 4.0	"Fair quality"
> 4.0	"Poor quality"

The final step consists in defining the minimum quality level that will be required for a given dataset to be included in the OptEEmAL platform. Considering the level of detail that the platform is expected to have and provide, we consider here the maximum acceptable value for the DQR to be 3.0. This value represents the limit between datasets of "good" quality and "fair" quality (according to the considered methodology).

6.3 Application to structured data sources

6.3.1 Energy Plus Weather

Data used considered in this section are available on the EnergyPlus Weather website mentioned before in this document (section 2.2.1). Their main characteristics with respect to data quality are indicated below (Table 15).

Table 15: Data quality assessment for EnergyPlus Weather data

Data quality criteria	Value	Associated score (according to the OptEEmAL methodology)
G	Country specific	2
TiR	2015	3
A	Fair uncertainty	3
C	Very good completeness	1
R	Widely acknowledged institution	2

Using equation (2) presented above, the aggregated data indicator for this example is:

$$DQR = \frac{2+3+3+1+2}{5} = 2.2 \quad (3)$$

Consequently, it means that the considered dataset is of "good quality" and can be used in the calculations of the OptEEmAL platform.

6.3.2 Eurostat

Data considered in this section are available on the EUROSTAT portal as indicated before in this document.

6.3.2.1 Energy prices

Their main characteristics of energy prices data from EUROSTAT with respect to data quality are indicated below (Table 16).

Table 16: Data quality assessment process for energy prices data from Eurostat

Data quality criteria	Value	Associated score (according to the OptEEmAL methodology)
G	Country specific	3
TiR	2015	1
A	Fair uncertainty	3
C	Very good completeness	1
R	Widely acknowledged institution	1

Using equation (2) presented above, the aggregated data indicator for this example is:

$$DQR = \frac{3+1+3+1+1}{5} = 1.8 \quad (4)$$

Consequently, it means that the considered dataset is of "very good quality" and can be used in the calculations of the OptEEmAL platform.

6.3.2.2 Income levels

Their main characteristics of income level data from EUROSTAT with respect to data quality are indicated below (Table 17).

Table 17: Data quality assessment process for income level data from Eurostat

Data quality criteria	Value	Associated score (according to the OptEEmAL methodology)
G	Regional	2
TiR	2014	1
A	Fair uncertainty	3
C	Very good completeness	1
R	Widely acknowledged institution	1

Using equation (2) presented above, the aggregated data indicator for this example is:

$$DQR = \frac{2+1+3+1+1}{5} = 1.6 \quad (5)$$

Consequently, it means that the considered dataset is of "very good quality" and can be used in the calculations of the OptEEmAL platform.

7 Conclusions

The OptEEmAL platform needs a large amount of data as input to define the characteristics of the environment of the particular project that will work with it. This contextual data will be gathered using external services that will be accessed using two well differentiated sub-systems that have been denominated:

- Geo Connector.
- Unstructured data connector.

The external services to be used for gathering the structured contextual data have not been chosen among those described in the first iteration of the document. It was thought to use others that were chosen when, later on, it was done the research of the low level characteristics that data has to comply with, and we realized that the sources had to be different ones. This decision was motivated not only because of format issues but also because of aspects such the misunderstanding of the differences between weather and climate data (it is explained in section 2.1).

The aspect that goes beyond the actual state of the art is the use of data coming from external sources for simulation purposes. It is not easy (not to say impossible) to find references of a system that uses them as extensively as OptEEmAL will do. This fact makes of paramount importance to be strict when crossing the needs of the simulation software with the characteristics of the data that can be obtained from a particular data source. Despite the lack of specific requirements coming from other parts of the platform, it was only needed to know the software that will run the simulations at the end of a chain of several processes. Knowing what this software needs, in terms of data format and completeness, we can guarantee to fulfil the requirements imposed by it.

For the unstructured contextual data, it is explained how information from the external service will be extracted, performing queries on the fly regarding the district associated to the project. This explanation did not need to describe the internal aspects of its architecture since its prototype is already developed and just an adaptation will be needed. This adaptation will not have an important impact since data coming from this subsystem will not be stored in the database as data coming from the Geo Connector will be. It will be retrieved and shown to the user instantly on demand.

Once the description is done, the requirements specification follows. Two separated sets were stablished in which the integration with the rest of the platform is contemplated in a low enough level to develop the solution but not as low as to impose hard technical constraints that are not really needed in the development. The size and lack of complexity of the description of the software suggests not going too deep in describing the solution to allow freedom in its development, but it does not mean that the functionality that it has to comply with doesn't have to be described with enough level of detail as it was intended in the specific tables.

The last section, the one called "Quality Integration Test" performs a study of the data quality coming from the Geo Connector for ensuring that data gathered by the means described here will be useful for the platform purposes, mainly for the use of the information by the simulation engines. The score obtained was: "good quality" for climate data and "very good quality" for energy prices and income levels that are close to the top score.

8 References

[01] ISO, "ISO 14044 - Environmental Management - Life Cycle Assessment - Principles, frameworks and guidelines." 2006.

[02] M. Bergdahl, M. Ehling, E. Elvers, E. Földesi, T. Körner, A. Kron, P. LohauB, K. Mag, V. Morais, A. Nimmergut, H. Viggo Saebo, U. Timm, and M. Joao Zilhao, "Handbook on Data Quality and Assessment Methods and Tools," 2007.

[03] H. Chen, D. Hailey, N. Wang, and P. Yu, "A Review of Data Quality Assessment Methods for Public Health Information Systems," *Int. J. Environ. Res. Public Heal.*, vol. 11, pp. 5170 – 5207, 2014.

[04] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality Assessment Methodologies for Linked Open Data: A Systematic Literature Review and Conceptual Framework," *Semant. Web – Interoperability, Usability, Appl.*, vol. 1, p. 33, 2012.

[05] B. Weidema, C. Bauer, R. Hischier, C. Mutel, T. Nemecek, J. Reinhard, C. Vadenbo, and G. Wernet, "Data quality guideline for the ecoinvent database version 3," *ecoinvent reports*, vol. 3, no. 1, 2013.

[06] N. Jain, A. Ramallo Gonzalez, and S. Natarajan, "Understanding the Effect of Aggressive Energy Efficiency Regulation on an Unprepared Building Sector Using Uncertainty Analysis.," in *Building Simulation International Conference*, 2015.

[07] S. Manfredi, K. Allacker, K. Chomkhamsri, N. Pelletier, and D. M. De Souza, "Product Environmental Footprint (PEF) Guide," 2012.

[08] Official Journal of the European Commission, *Commission Recommandation on the use of common methods to measure and communicate the life cycle environmental performance of products and organisations*. Europe, 2013, pp. 1–210.

[09] European Commission, "EUROSTAT," 2016. [Online]. Available: http://ec.europa.eu/eurostat/. [Accessed: 07-Jun-2016].

[10] European Commission, "Electricity prices for domestic consumers - bi annual data (from 2007 onwards)," 2016.