EUBRA BIGSEA

EUROPE - BRAZIL COLLABORATION OF BIG DATA SCIENTIFIC RESEARCH THROUGH CLOUD-CENTRIC APPLICATIONS

D7.2: GES3 Data Integration

Author(s)	Nádia P. Kozievitch (UTFPR), Ignacio Blanquer (UPV), Luiz Fernando (UFMG), Walter do Santos (UFMG), Andy S Alic (UPV)
Status	Approved
Version	V1.0
Date	22/11/2016

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

EUBra-BIGSEA is funded by the European Commission under the Cooperation Programme, Horizon 2020 grant agreement No 690116.
Este projeto é resultante da 3a Chamada Coordenada BR-UE em Tecnologias da Informação e Comunicação (TIC), anunciada pelo Ministério de Ciência, Tecnologia e Inovação (MCTI)

Abstract: Europe - Brazil Collaboration of BIG Data Scientific Research through Cloud-Centric Applications (EUBra-BIGSEA) is a medium-scale research project funded by the European Commission under the Cooperation Programme, and the Ministry of Science and Technology (MCT) of Brazil in the frame of the third European-Brazilian coordinated call. The document has been produced with the co-funding of the European Commission and the MCT.
The purpose of this report is to describe the software and data that implement the various data integration techniques from the Massively connected society - Smart cities, use case. The data and software pointed out in this document constitute the main content of deliverable D7.2, of type "OTHER".

Document identifier: EUBRA BIGSEA –WP7-D7.2	
Deliverable lead	UFTPR
Related work package	WP7
Author(s)	Nádia P. Kozievitch (UTFPR), Ignacio Blanquer (UPV), Luiz Fernando (UFMG), Walter do Santos (UFMG), Andy S Alic (UPV)
Contributor(s)	Name (Institution)
Due date	31/10/2016
Actual submission date	22/11/2016
Reviewed by	Sandro Fiore (CMCC Foundation), Daniele Lezzi (BSC)
Approved by	PMB
Start date of Project	01/01/2016
Duration	24 months
Keywords	Data Integration, GES3

Versioning and contribution history

Version	Date	Authors	Notes
0.1	10/10/16	Ignacio Blanquer (UPV)	TOC
0.2	18/10/16	Nádia P. Kozievitch (UTFPR)	
0.3	19/10/16	Ignacio Blanquer (UPV)	Updated TOC and initial data
0.4	24/10/16	Luiz Fernando (UFMG), Walter (UFMG)	Updated data sources info, section about data acquisition
0.5	24/10/16	Andy S Alic (UPV)	Added Brussels
0.6	8/11/16	Ignacio Blanquer (UPV)	Section 5
0.7	21/11/16	Sandro Fiore (CMCC), Ignacio Blanquer (UPV), Nádia P. Kozievitch (UTFPR), Walter (UFMG)	Comments from reviews completed

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

1 EXECUTIVE SUMMARY

EUBra-BIGSEA project aims at developing a set of cloud services empowering Big Data analytics to ease the development of massive data processing applications. EUBra-BIGSEA will develop models, predictive and reactive cloud infrastructure QoS techniques, efficient and scalable Big Data operators and a privacy and quality analysis framework, exposed to several programming environments. EUBra-BIGSEA aims at covering general requirements of multiple application areas, although it will showcase in the treatment of massive connected society information, and particularly in traffic recommendation.

The project starts with the analysis of the use case scenarios that will be used for demonstration, but considering those requirements in broader way. EUBra-BIGSEA is an API-centric project whose main objective is to create a sustainable international (European and Brazilian) cooperation activity in the area of cloud services for Big Data analytics. In particular, T7.2 aims at improving efficiency and throughput of data scientists and data curators.

The Acquisition and Engineering of Georeferenced Environmental, Stationary, Streaming and Social (GES³) data (Task 7.2) is related to the Use Case 1 - (UC1) - Data Acquisition (D7.1). In particular, these data come from sources that are related to urban traffic and cover four main data types: stationary data, dynamic spatial data, environmental data, and social network data. Despite that the EUBra-BIGSEA pilot has been initially planned over the data of the city of Curitiba, where the pilot case is being constructed, the EUBra-BIGSEA framework will be applicable (at least partially) to other scenarios.

Therefore, the data integration covers the general problem of mechanisms for collecting, cleaning, transforming and integrating all the listed data sources, in order to understand the dynamics of traffic and transportation public services in Brazilian cities. Note that some of them are official (original data from municipalities) and some of them are informal (Google/OpenstreetMaps). This document also covers several issues among their integration: missing entities (new streets and parks at a new district), difference among sources (Google has an average of 60% of the available bus stops in Curitiba, GTFS has not the same bus lines), non-structured, tabular, and spatial data, as long as different precision, accuracy and consistency. In summary, issues related to data quality, entity matching, and data mining. Several pre-processing steps will be applied within this phase, so the final user can interact (query) an integrated data source.

After a description of all data sources, the integration process has gone through the following steps: first, data sources within the same theme were integrated (such as official sources). Second, we performed an integration along different data types (such as stationary and dynamic spatial data). Third, we identified their issues as data quality, entity matching, or data mining problem. Finally, we identified mechanisms to improve their integration and quality for the final user.

The data integration serves to guide the Descriptive Models of GES³ Data (Task 7.3) and Predictive models based on GES³ Data (Task 7.4), as long as the Routes for people Use Case (Task 7.5). The process is also of interest for the rest of the platform as high-level needs that require to be addressed through the platform functionalities.

1 INTRODUCTION

1.1 Scope of the Document

This document describes the data integration process as well as the data sources description. The document also describes practical integration scenarios.

1.2 Target Audience

The document is mainly intended for internal use, although it is publicly released.

At internal level, WP7 members will find in the document the data issues that will be addressed by EUBra-BIGSEA, as well as the expected results at the level of data integration. Technical developers from WP3, WP4, WP5, and WP6 will find the scenarios they have to address in their developments. The project success will be also measured in the degree of fulfilment of such issues.

At external level, Data Scientists (referred here as developers of data-analysis-intensive applications) could evaluate if the issues and solutions addressed are similar to those they have, considering the possibility of adopting the technology. Municipality stakeholders can evaluate and visualize issues already cited in a particular scenario, and possible solutions. Application developers could also understand the kind of problems that could be addressed using the EUBra-BIGSEA components.

The information of this document will be periodically updated in the internal wiki, and the final outcome of the implementation of such requirements will be analysed in deliverable D7.6 (Validation of requirements), by the end of the project.

1.3 Structure

The document is structured into 5 sections. Section 2 describes the nature of the GES^3 Data. Section 3 describes the Data sources used in the project. Section 4 describes the problem of Data acquisition and Section 5 describes the problem of Data integration.

2 GES³ DATA

2.1 Concept of GES³ Data

EUBra-BIGSEA must deal with multiple sources and types of data. These data sources are related to urban traffic as well as static and dynamic information. Despite that they are related to the city of Curitiba, where the pilot case is being constructed, the EUBra-BIGSEA framework should be applicable to other scenarios.

Four main data types are identified:

1) **Stationary data**. This data is related to long-living data that describes the topology of the traffic network of the city, the street map, relevant city spots (such as bus stops and bus terminals), and other geographic information that is relevant to understand the location of the components present within the urban mobility scenario.

2) **Dynamic spatial data**. This data contains georeferenced information of the vehicles and users, valid for a specific point in time.

3) **Environmental data**. This data provides information about the environmental conditions and the weather forecasts that are also relevant for understanding citizens' mobility.

4) **Social network data**. Refers to immediately consumed data that can provide information about sentiments and unpredictable events.

The first three types are available here[1] (the stationary data used sample from 2015) and samples of the last two data types are available here[2]. Note that additional non official sources were considered, such as GTFS[3] and OpenStreetMaps. The following subsections present the data description of each of these data sources.

2.2 Stationary Data

Although the project is already collecting data from different cities (such as New York and Valencia), the project's main use case is centred in the municipality of Curitiba, Paraná, Brazil. The datasets used for stationary data came from official sources (IPPUC[4], the Municipality of Curitiba[5], along with data from URBS[6]) and non official sources (Open Street Map and Google Maps). Formats vary from shapefiles to csv. General statistics from the mobility data in Curitiba are presented in Table 1 (official data from URBS). Details are listed below.

Table 1 - General statistics from URBS

Description	Quantity
Official Number of Bus Vehicles	1,500
Official Number of Daily Passengers	1,620,000
Official Number of Bus Routes	250

1 https://bigsea.owncloud.lsd.ufcg.edu.br/owncloud/index.php/s/UFKZhHGdxvWzO8w?path=%2FStreamingDataBus - –Last visited on 04/04/2016.

2 http://data.ctweb.inweb.org.br/ - Last visited on 04/04/2016.

3 https://developers.google.com/transit/ - Last visited on 04/04/2016.

4 http://ippuc.org.br/geodownloads/geo.htm. –Last visited on 04/04/2016.

5 www.curitiba.pr.gov.br/dadosabertos/ –Last visited on 04/04/2016.

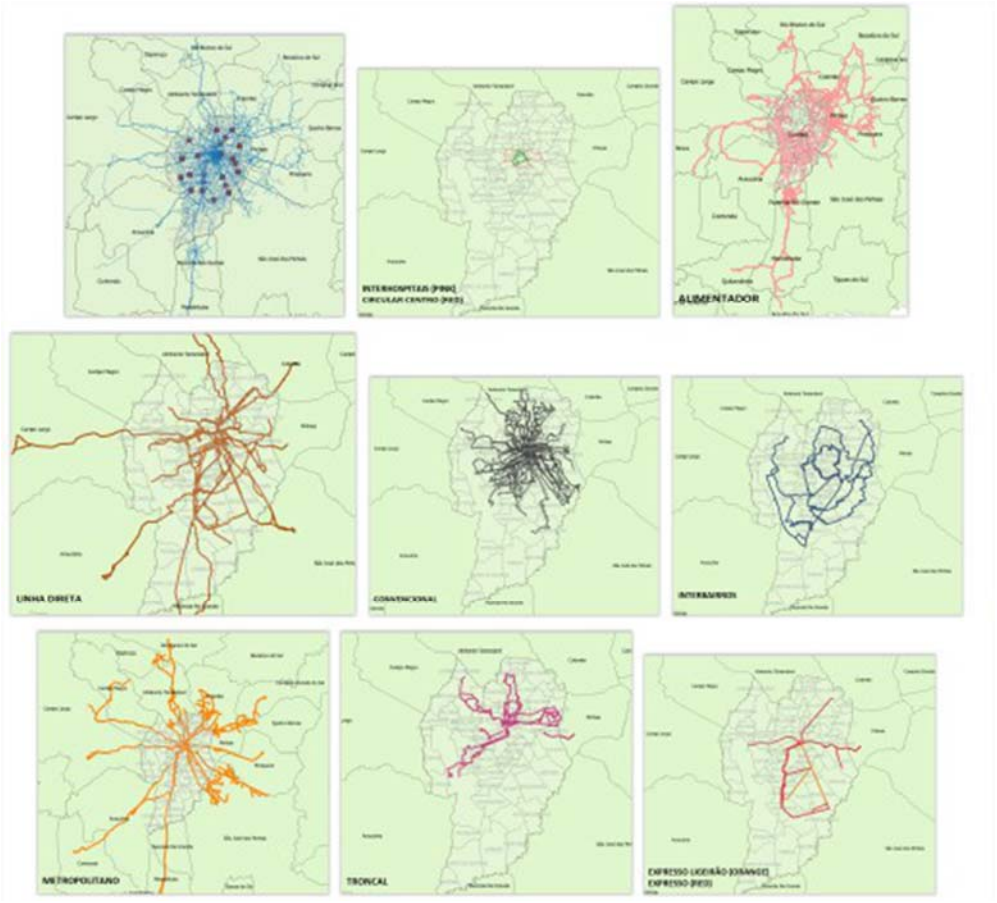
6 https://www.urbs.curitiba.pr.gov.br/ –Last visited on 04/04/2016.

Number of Bus Terminals	23
Number of Tube Bus Stations	342
Average Bus Routes KMs by day	480,000
Average Age of the Fleet	7
Number of roads in Curitiba	9,135

Stationary (or static) data might include geo-referenced data from bus lines, bus stops, terminals, and streets that do not change often.

Bus Routes. The city has an average of 482 bus routes distributed within 11 categories (Metropolitano, Linha Direta, Expresso Ligeirão, Interbairros, Expresso, Troncal, Convencional, Turismo, Circular Centro, Interhospitais, Alimentador), as shown by coloured lines in Figure 1. The categories *Alimentador* and *Convencional* have the majority of lines, with 265 and 65 units, respectively.

Figure 1 – Representation of bus lines (blue) and bus terminals (red) on upper left, followed by different bus categories

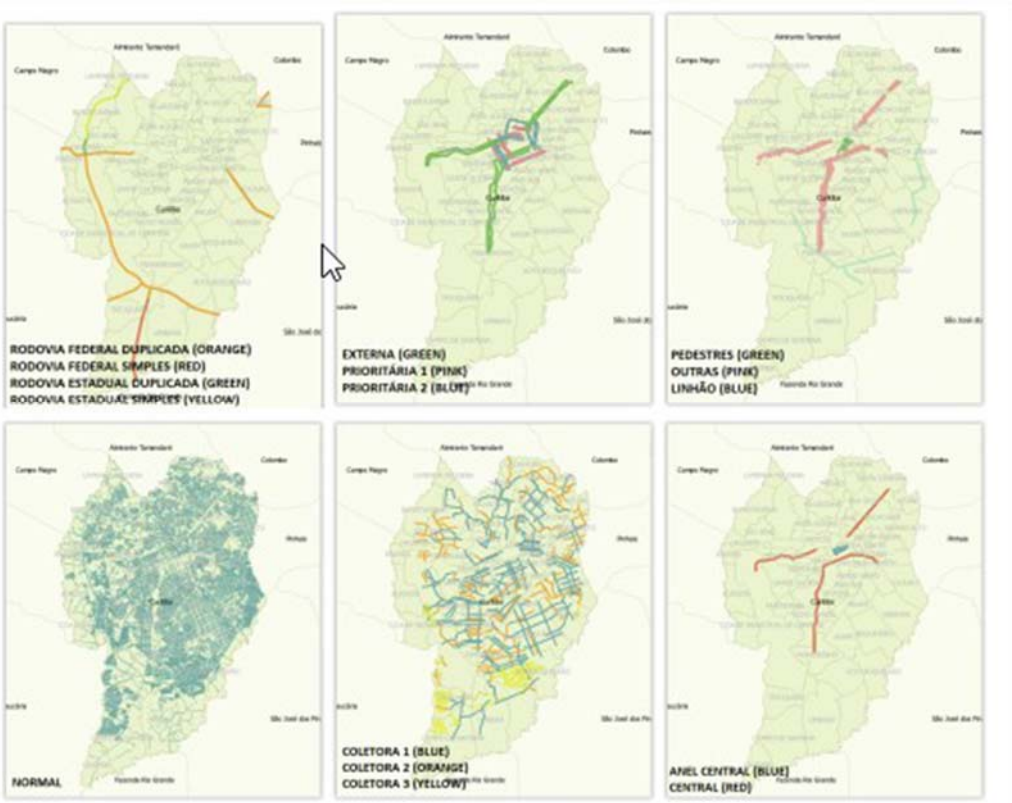


Bus Stops. The city has an average of 9,940 bus stops detected within the stationary data. Curitiba has an additional bus stop category named Tube Bus Stations (in portuguese, Estação Tubo). The tube stations (officially 342) are bus stops which look like tubes, for specific bus routes, such as Expresso and Linha Direta. The districts named CIC and Centro have the majority of them, with a total of 1,628 and 667 units each one.

Bus Terminals. The city has 23 terminals (buses) and one terminal which also uses trains, as shown by red points in Figure 1. The oldest one is named Guadalupe, from January 1st, 1956. The districts CIC and Boqueirão concentrate their majority, with 3 and 2 units each one.

Streets. The city has 9,135 streets, divided among 17 categories (Anel Central, Central, Coletora 1, Coletora 2, Coletora 3, Externa, Linhão, Normal, Outras Vias, Pedestre, Prioritária 1, Prioritária 2, Rodovia Estadual Duplicada, Rodovia Estadual Simples, Rodovia Federal Duplicada, Rodovia Federal Simples, and Setorial). Each category might present a distinct speed limit and objective in the city, as listed in Figure 2. The districts CIC and Sítio Cercado have their majority, with 1217 e 524, each one.

Figure 2 – Official Representation of streets, by each category



The official stationary database included street layout (streets, future streets, blocks, boardwalks, cemeteries, government edifications of art, kerbs, parks, squares), hydrography (water parting, lakes, rivers), legal limits (districts, cities, states), transportation (bus lines, bus stops and bus terminals).

2.3 Dynamic Spatial Data

The datasets used for dynamic spatial data came from official sources (URBS) and non-official sources (Generic Transit Feed Specification - GTFS[7]). Formats vary from text files to zip files. Dynamic data from URBS include data from bus lines, and user cards, transmitted at an average frequency of 5 minutes. In particular, the examples listed here use data between October 19th, 2015 and October 26th, 2015.

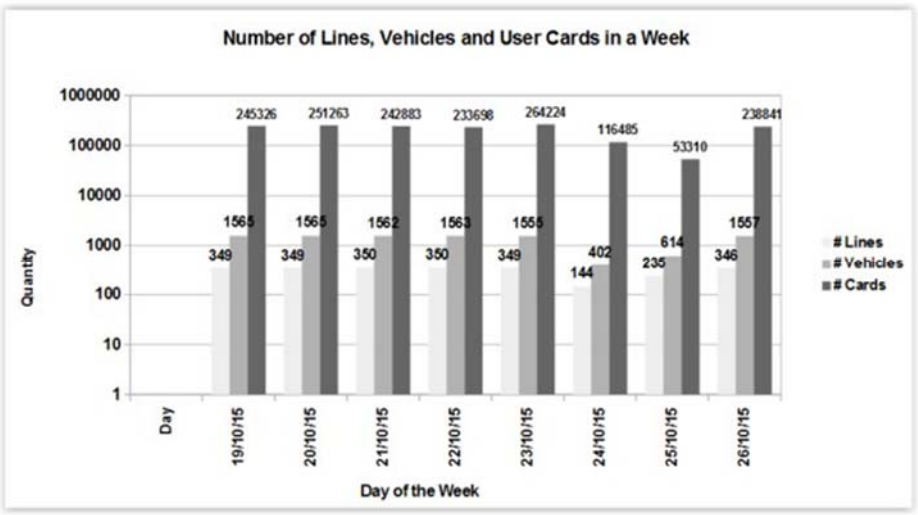
User Cards. This set of data represents the commuting information comprising trip data per user card: the vehicle id, line code, the user card number, and the date of the trip. Within the date range selected, it was detected 349,729 different user cards, 269 different bus lines were present, divided among 1,522 vehicles. Part of the vehicles and lines represented here have geolocated data, presented within the following set.

[7] https://developers.google.com/transit/ –Last visited on 04/04/2016.

Vehicles and Respective Bus lines. This set of data represents the daily itinerary, having as data the vehicle id, line code, the datetime, and the latitude and longitude. Within the data range selected, it was detected 1,623 distinct vehicles and 703 different bus lines.

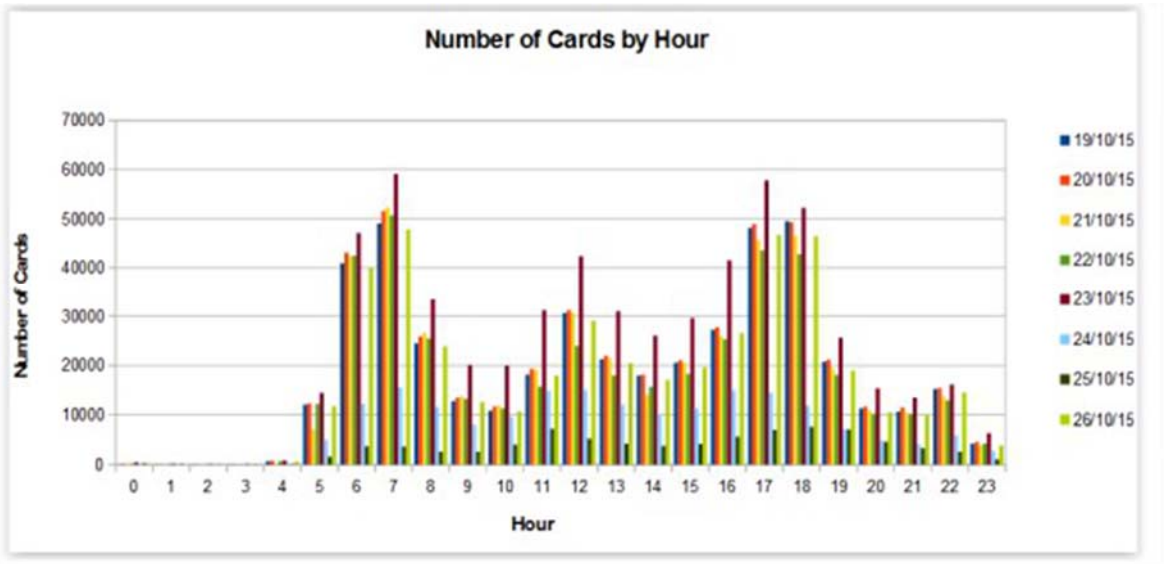
An initial analysis of dynamic data from user cards (time range listed above) indicated some trends: (i): Over the week of data analyzed, there is an average of 350 bus lines, an average of 1,500 vehicles, and an average of 250,000 user cards (Figure 3) per day. Sundays presented the lowest number of user cards. (ii): The analysis of user cards by hour indicated that the peak hours for traffic are 7 AM, 5 PM, and 6 PM (Figure 4). Friday (October 23rd, 2015) presented the highest amount of different user cards (264,224), and Sunday (October 25th, 2015) presented the lowest amount of user cards (53,510). (iii): The top 5 bus lines which had the majority of the user cards were during the time range were ”Oper s/Linha”, ”Op. Contingência”, ”Sistema Araucária”, ”Interbairros IV” and ”Interb. II Anti H”. The first one indicated bus operating without a specific line. The second one indicated extra buses to the already existing lines. The third one connects the metropolitan region to Curitiba. The last two lines operate within the Interbairros category. These lines connect several districts.

Figure 3 – Number of Lines, Vehicles, and User Cards in a Week



Note that dynamic data also considered GTFS. A GTFS feed is a collection of at least six, and up to 13 CSV files contained within a .zip file. In contrast to European transit industry exchange standards such as Transmodel or VDV-45X, GTFS only includes scheduled operations that are meant to be distributed to riders. GTFS include data from trips, routes, stop times, stop and calendar, among others.

Figure 4 – Number of user cards by hour

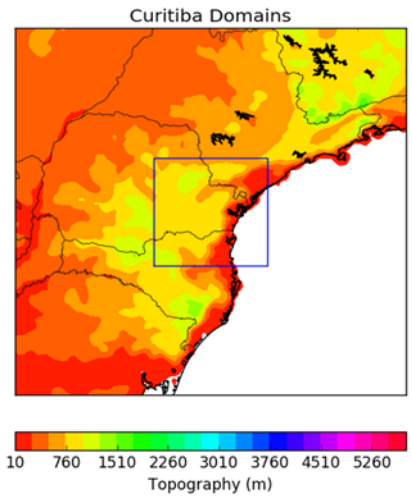


2.4 Environmental Data

Environmental data consists on weather forecasting information for the region around the city of Curitiba generated by the Weather Research and Forecasting Model (WRF). The model was configured with nested grids covering the center-south area of Brazil (Figure 5). The grids have equal size of 99x99 points in east-west and north-south direction, with horizontal resolutions of 36km (outer domain) and 12km (inner domain). Forecasts are produced daily, using as initial conditions the analysis and forecasts from Global models initialized at 00 UTC time. Timespan of the forecasts are typically 72hrs, generating weather information for every point in both domains containing several meteorological variables, such as:

- Precipitation
- Wind magnitude and direction
- Air temperature
- Relative humidity
- Atmospheric pressure
- Instability indexes

Figure 5 – Nested WRF domains, centered over the city of Curitiba



2.5 Social Network Data

Social network data represents an important data source that can provide information about opinions, personal reports, sentiments, unpredictable events and incidents. Regarding traffic and mobility, this kind of data represents a valuable source of real time information although its volume and reliability should be carefully analyzed. We have social information from two social networks: *Twitter* and *Facebook*.

Twitter. The Twitter dataset is collected through Twitter Streaming API[8] producing real time data. The Twitter dataset can be applied to detect traffic events, mobility issues, crowds of people and unpredictable events and arises some NLP, sentiment analysis and spatio-temporal challenges. Data collection follows three different filters: geolocation, users and terms. The geolocated tweets dataset is composed of every message posted within the rectangle that covers the Brazilian territory. The users filter aims every message posted by specific users within a list of target users. These target users are those that post messages about traffic status. Finally, the terms filter defines some target keywords related to traffic. All filters are described here[9]. The estimated data volume is shown on Table 2.

Table 2 – Daily number of records collected by each filter

Filter	Daily volume
Geolocated	800,000
Users	5,000
Terms	90,000

Facebook. Data from Facebook is concerned with events created in the Social Network. Such events can enable analysis about crowds of people that may affect the traffic and the mobility. As the Facebook API[10] does not provide a specific query to search for events within a region, we apply a heuristic to find events as detailed in the following. First, we map the subdistricts of each city. These subdistricts are smaller regions inside the cities. Second, we map the centroid of each subdistrict. We use the IBGE[11] database to perform these two initial steps. Third, for each centroid, we query for existing venues on Facebook located within 2,000 meters radius from the centroid. Finally, we search for events created by these venues. These two final steps are done with the Facebook API. The crawler runs once a day collecting about data about 1,200 events in each execution. We do not avoid event repetitions because some features may vary, such as the event popularity.

8 https://dev.twitter.com/streaming/overview
9 http://data.ctweb.inweb.org.br/
10 https://developers.facebook.com/
11 http://www.ibge.gov.br/home/

3 DATA SOURCES

EUBra-BIGSEA aims at developing cloud services for the management of Big Data Analytic applications. The demonstrative application to be developed on top of EUBra-BIGSEA is focused on traffic recommendation. This use case will use information from multiple sources to provide a recommendation that is not uniquely based on the distance or the time estimated to arrive to destination. Therefore, multiple sources of data are required to perform this prediction and analysis, including social data, points of interest, historical information, climate, etc.

This section describes the data sources identified in the project that are used for the development of the use case and for the storage services in WP4. The consortium uses a shared document [12]

3.1 Brazil

Data provided by online social networks covering the whole Brazilian territory:

Table 3 – Data sources for the whole Brazilian territory

Data	Category	Format	Update	URL
Tweets	Social network data	JSON	Realtime	https://twitter.com/

3.1.1 Top 10 cities

Data provided by online social networks and other services available in the Web are being collected for the top 10 most populated cities in Brazil, which are:

1. São Paulo
2. Rio de Janeiro
3. Brasília
4. Salvador
5. Fortaleza
6. Belo Horizonte
7. Manaus
8. Curitiba
9. Recife
10. Porto Alegre

Table 4 – Data sources generic for several cities in Brazil

Data	Category	Format	Update	URL
Points of interests	Stationary data	JSON	Weekly	https://www.tripadvisor.com/Attractions-*

[12] Data sources characterization: https://drive.google.com/drive/u/0/folders/0B9CyYvZMJ-RSQ0E3cERaQzJtQVk

Events from Facebook	Social network data	JSON	Daily	https://www.facebook.com
News from the web	Social network data	JSON	Realtime	Many
Traffic status	Dynamic spatial data	JSON	30 min	http://www.maplink.com.br/*/*/TransitoAgora
Weather conditions	Environmental data	JSON	1 hour	https://openweathermap.org/
Weather conditions and predictions	Environmental data	JSON	12 hours	https://darksky.net/dev/

3.1.2 Curitiba

Table 5 – Data sources specific to Curitiba

Data	Category	Format	Update	URL
Points of interest	Stationary data	JSON	Weekly	http://transporteservico.urbs.curitiba.pr.gov.br/getPois()
Traffic reports	Dynamic spatial data	JSON	Daily	http://www.curitiba.pr.gov.br/boletimtransito
Bus card	Dynamic spatial data	JSON	Data of 97 days	Not online
Bus vehicles position	Dynamic spatial data	JSON	Data of 97 days	Not online
Bus lines	Stationary data	JSON	Daily	http://transporteservico.urbs.curitiba.pr.gov.br API Method “getLinhas()”
Bus stops	Stationary data	JSON	Daily	http://transporteservico.urbs.curitiba.pr.gov.br API Method “getPontosLinha()”
Bus vehicles	Stationary data	JSON	Daily	http://transporteservico.urbs.curitiba.pr.gov.br API Method “getTabelaVeiculo()”
Bus routes	Stationary data	JSON	Daily	http://transporteservico.urbs.curitiba.pr.gov.br API Method “getShapeLinha()”

Bus paths	Stationary data	JSON	Daily	http://transporteservico.urbs.curitiba.pr.gov.br API Method “getTrechosItinerarios()”
Bus schedules	Stationary data	JSON	Daily	http://transporteservico.urbs.curitiba.pr.gov.br API Method “getTabelaLinha()”
Taxi GPS	Dynamic spatial data	CSV	Data of 3 days	Not online
GTFS	Stationary data	GTFS	Yearly	http://transporteservico.urbs.curitiba.pr.gov.br API Link “Arquivo GTFS”
Bus card	Dynamic spatial data	POSTGIS	Daily	Will be available to the project.
Bus vehicles position	Dynamic spatial data	POSTGIS	Daily	Will be available to the project.
Bus lines	Stationary data	POSTGIS	Daily	Available at the Postgis Database.
Bus stops	Stationary data	POSTGIS	Daily	Available at the Postgis Database.
Bus vehicles	Stationary data	POSTGIS	Daily	Will be available to the project.
Bus routes	Stationary data	POSTGIS	Yearly	Available at the Postgis Database.
Bus paths	Stationary data	POSTGIS	Yearly	Available at the Postgis Database.
Bus schedules	Stationary data	POSTGIS	Yearly	Available at the Postgis Database.

[*] This information is not publicly available and requires holding the proper credentials.

3.1.3 Campina Grande

Table 6 – Data sources specific to Campina Grande

Data	Category	Format	Update	URL
Bus Lines, stops and schedules	Stationary data	GTFS	Monthly	Not publicly available yet
Bus vehicles locations over time	Dynamic spatial data	JSON	Continuously	Not publicly available yet

3.1.4 Belo Horizonte

Table 7 – Data sources specific to Belo Horizonte

Data	Category	Format	Update	URL
Taxi GPS	Dynamic spatial data	CSV	Data of 3 days	Not online
GTFS	Stationary data	GTFS	Yearly	http://servicosbhtrans.pbh.gov.br/bhtrans/e-servicos/S43F01-extracao.asp
Bus schedule	Stationary data	JSON	Yearly	http://servicosbhtrans.pbh.gov.br/bhtrans/e-servicos/S43F01-extracao.asp
Bus stops	Stationary data	JSON	Yearly	http://servicosbhtrans.pbh.gov.br/bhtrans/e-servicos/S43F01-extracao.asp
Bus lines	Stationary data	JSON	Yearly	http://servicosbhtrans.pbh.gov.br/bhtrans/e-servicos/S43F01-extracao.asp

3.1.5 Fortaleza

Table 8 – Data sources specific to Fortaleza

Data	Category	Format	Update	URL
Taxi GPS	Dynamic spatial data	CSV	Data of 3 days	Not online
GTFS	Stationary data	GTFS	Yearly	http://dados.fortaleza.ce.gov.br/catalogo/dataset?tags=transporte%20p%C3%BAblico

3.1.6 Rio de Janeiro

Table 9 – Data sources specific to Rio de Janeiro

Data	Category	Format	Update	URL
Bus vehicles position	Dynamic spatial data	JSON	1 minute	http://data.rio/group/transporte-e-mobilidade
GTFS	Stationary data	GTFS	Yearly	http://data.rio/dataset/onibus-gtfs

3.1.7 São Paulo

Table 10 – Data sources specific to São Paulo

Data	Category	Format	Update	URL
Bus vehicle position	Dynamic spatial data	JSON	2 minutes	http://www.sptrans.com.br/desenvolvedores
GTFS	Stationary data	GTFS	Yearly	http://www.sptrans.com.br/desenvolvedores/GTFS.aspx
Bus lines	Stationary data	JSON	Weekly	http://www.sptrans.com.br/desenvolvedores/APIOlhoVivo/Documentacao.aspx
Bus stops	Stationary data	JSON	Weekly	http://www.sptrans.com.br/desenvolvedores/APIOlhoVivo/Documentacao.aspx

3.2 Europe

3.2.1 Valencia

The city of Valencia has several sources of geolocalized data related to the use case. In particular, the information available is:

Table 11 – Data sources from the city of Valencia. (*)All URLs are preceded by admoncatalogo.valencia.es/dataset

Data	Category	Format	Update	URL
Real time traffic data	Traffic intensity	GeoJSON	3 minutes	http://*/estado-trafico-tiempo-real
Bike lanes traffic intensity	Traffic intensity	CSV, GeoJSON	Yearly	http://*/intensidad-de-los-puntos-de-medida-de-bicicletas-espiras-electromagneticas
Bus lines and stops	Transport Maps	ZIP	Yearly	http://*/google-transit-lineas-paradas-horarios-de-autobuses
Bus Stops	Transport Maps	CSV, GeoJSON	Yearly	http://*/paradas-emt
Taxi stops	Transport Maps	GeoJSON	Yearly	http://*/paradas-taxis
Bike lane	Transport Maps	CSV, GeoJSON	Yearly	http://*/carril-bici
Bike stops	Transport Maps	CSV, GeoJSON	Yearly	http://*/aparcabicis
Noise Map	Health	CSV	Monthly	http://*/datos-diarios-ultimo-mes-estaciones-ruido, http://*/mapa-ruido-noche, http://*/mapa-ruido-tarde, http://*/mapa-ruido-lden-24h, http://*/mapa-ruido-dia

Air pollution	Health / Environment	CSV	Monthly	http://*/datos-contaminacion-atmosferica-boulevar-sur-7a, http://*/datos-contaminacion-atmosferica-avda-francia-6a, http://*/datos-contaminacion-atmosferica-viveros-5a, http://*/datos-contaminacion-atmosferica-pista-silla-4a, http://*/datos-contaminacion-atmosferica-moli-sol-3a, http://*/datos-contaminacion-atmosferica-universidad-politecnica-1a
Pollen maps	Health / Environment	GeoJSON	Yearly	http://*/mapa-polen-casuarina, http://*/mapa-polen-ulmus, http://*/mapa-polen-ligustrum http://*/mapa-polen-fraxinus, http://*/mapa-polen-cupressus, http://*/mapa-polen-quercus, http://*/mapa-polen-morus, http://*/mapa-polen-olea http://*/mapa-polen-populus, http://*/mapa-polen-pinus, http://*/mapa-polen-platanus
Fallas (monuments of our main celebration)	Tourism	CSV / GeoJSON	Yearly	http://*/monumentos-falleros
Monuments	Tourism	CSV / GeoJSON	Yearly	http://*/monumentos-turisticos

3.2.2 Brussels

For Brussels there is quite a lot of info publicly available.

Table 12 – Data sources specific to Brussels [More datasets are available here]

Data	Category	Format	Update	URL
Bus lines and stops	Transport Maps/GTFS	CSV	Every few days	https://transitfeeds.com/p/societe-des-transports-intercommunaux-de-bruxelles/527
Bus Stops	Transport Maps/GTFS	CSV	Every few days	https://transitfeeds.com/p/societe-des-transports-intercommunaux-de-bruxelles/527
Bus Scheduling	Transport Maps/GTFS	CSV	Every few days	https://transitfeeds.com/p/societe-des-transports-intercommunaux-de-bruxelles/527

4 GES³ DATA ACQUISITION

4.1 Challenges in data acquisition

Each data source has a different set of requirements. These requirements are associated to the volume of the data, update interval, data format, and availability.

Stationary data might come from different official sources, and changes along the years. In general, the data volume is low, with rare updates.

Dynamic spatial data acquisition brings the challenges of acquiring, storing, processing and querying large volume of data and when the data include user information, the concerns about security and privacy arise. Due to the large volume, some data may be lost if they are not acquired within a time window. Almost all dynamic spatial data acquired by the project is provided by cities government agencies or transportation companies (such as taxi companies). Some data are provided as a dump and do not require a special software, but most of time, an API is provided and in this case, an API consume is implemented. Data source providers require authentication to identify the consumer in order to provide sensitive data or control/limit the number of requests by a time window.

Environmental data usually have predictable size as they are associated with a predictable number of regions and aspects. However, data volume can be large and processing them can be very complex due to the scientific nature of the data and metadata-related aspects (e.g. at the level of global and variable specific attributes, etc.). In addition, data should be updated very often as the metadata may change very often.

While dynamic spatial data volume can be huge, it is in general predictable. On the other hand, data from online social networks (OSN) is not only huge but also very difficult to predict. During data acquisition, bursts can occur in the volume of comments in OSN and should be managed by data acquisition software. For example, during weather or sports events or even during TV shows broadcast, the amount of comments published in Twitter can grow from dozens to thousands in a few minutes. To deal with bursts, OSN data acquisition software should use intermediate storage, e.g. a queue, before storing data in the persistent storage. Not all data from OSN is available. There are policies to guarantee user privacy and they must be respected.

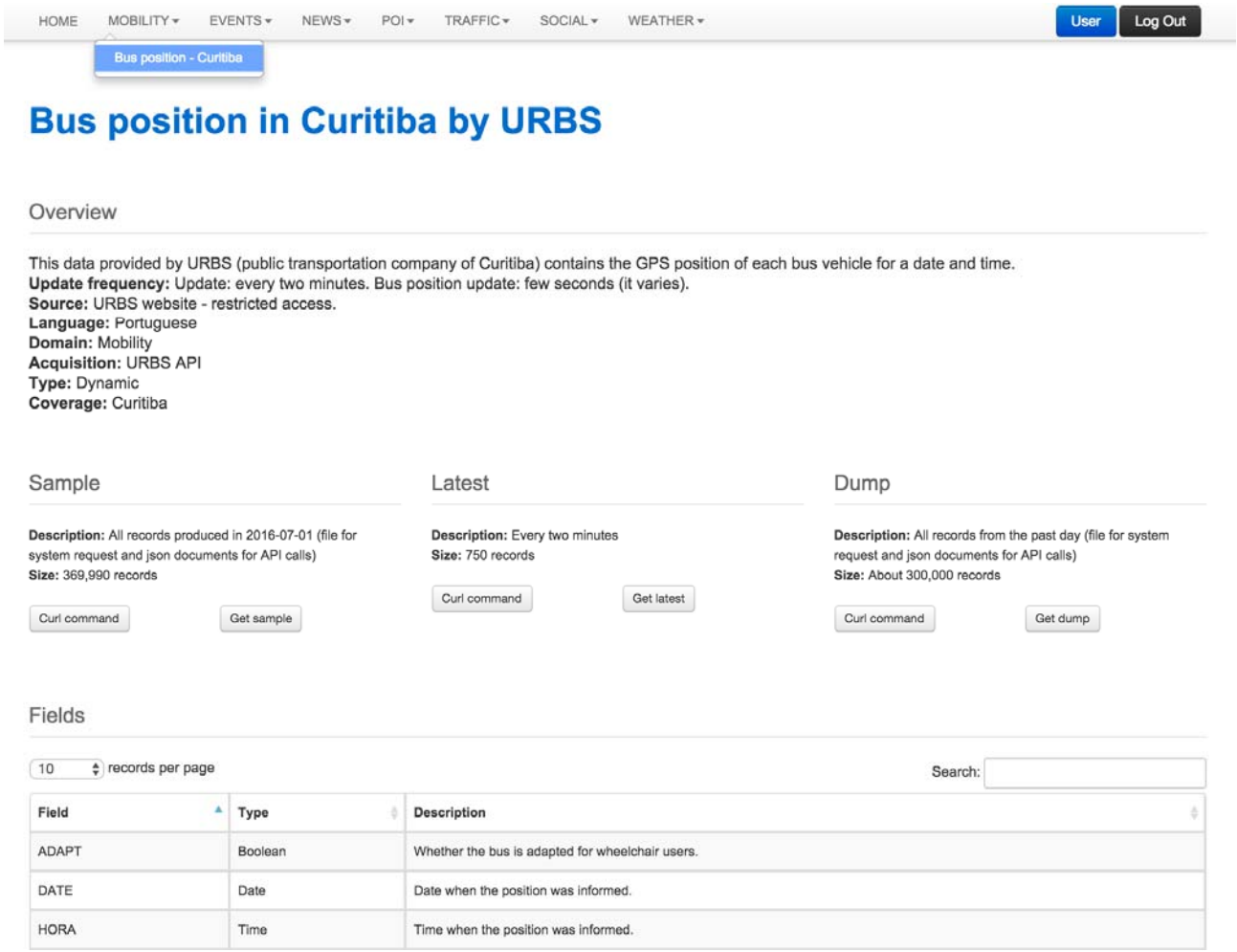
4.2 Data acquisition software

In order to acquire the different data types and data sources, a set of *collectors* were built, along with official data provided by municipalities. A collector is a specialized program that acquires data from a data source, respecting its *contract* (e.g. API specification), request limits and privacy restrictions. A collector also controls the frequency of acquisition, performs basic transformation in the data, deal with bursts and save data in persistent storage.

All collectors are written in *Python* and are available as open-source programs at a public repository (https://github.com/eubr-bigsea/crawlers). A collector either performs requests to an API using the Requests package (http://docs.python-requests.org/) or implements data web crawler data using the *BeautifulSoup* package (https://www.crummy.com/software/BeautifulSoup/). Most of them do not require authentication or the authentication can be obtained by requesting an access token to the data source provider.

Acquired data are made available to all EUBra-BIGSEA partners by an internal API and companion documentation site, available at http://data.ctweb.inweb.org.br (see Figure 6), along with stationary data in a project directory. The documentation site includes datasets regarding: (i) mobility, (ii) events, (iii) news, (iv) point of interests, (v) traffic, (vi) social media, (vii) weather. Figure 6 shows an example of mobility data regarding the bus position in Curitiba by URBS.

Figure 6 – Documentation site providing access to the data sources analys in the project http://data.ctweb.inweb.org.br



5 GES³ DATA INTEGRATION

The integration of data is a complex problem itself giving the different nature of the data sources. Along with the pure syntactic problems due to different formats, more complex problems appear in the inconsistency of geographic references among providers and along time. The acquisition of data has to deal with integrity checks, data quality analysis (that validate the coherence of locations and coordinates) and different formats (which change along the years and along the sources).

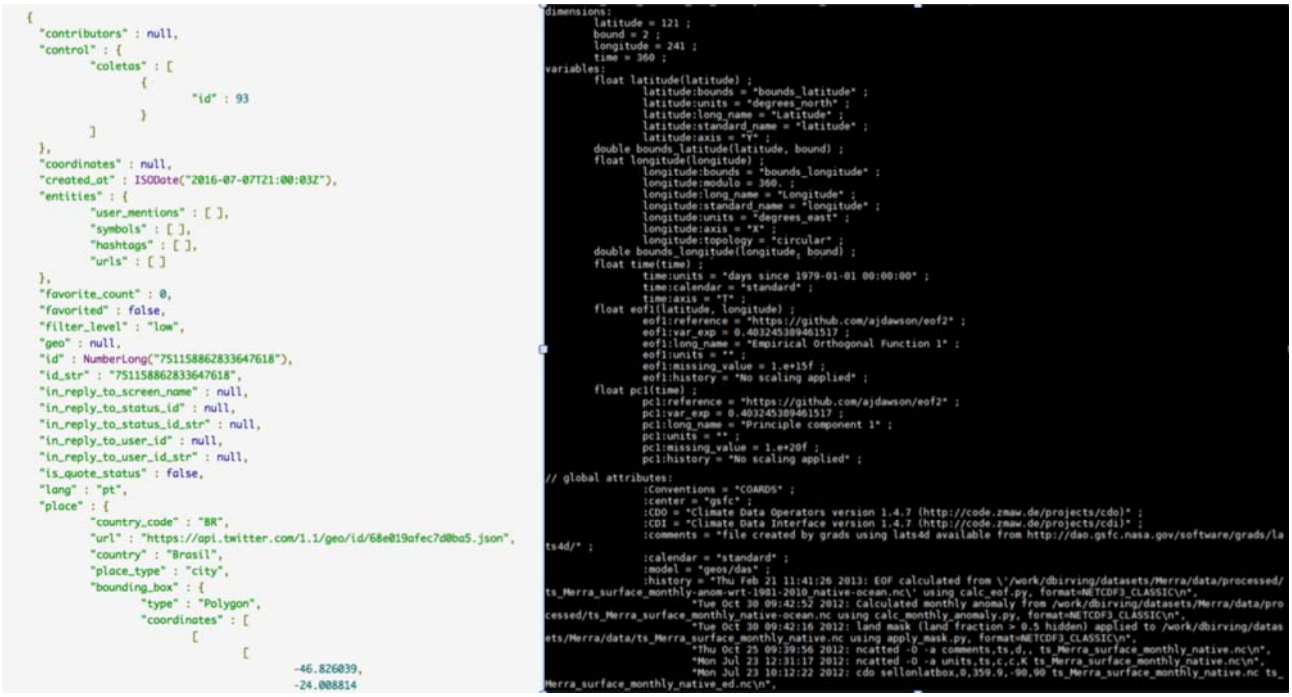
This section presents samples of issues for integrating data.

5.1 Issues integrating data

5.1.1 Different File Formats

Different data types and data from different sources were stored in different file formats. Recurrent ones are CSV, XLS, JSON and Shapefile; others, like NetCDF, are designed for array-based scientific data and need specific support as they include, despites the data part, also a metadata section. Depending on the integration technique or process, there could be additional steps like format conversions (e.g. for instance data could still be converted to SQL for data definition and insertion into a Relational DBMS) and/or pre-processing steps for compliance checks (e.g. like in the case of NetCDF files[13]). It is important to consider the integration process may not necessarily increase the quality of data.

Figure 7 – Geolocated tweets from Twitter (JSON format) and Weather data (NetCDF format)



5.1.2 Different Reference Systems

Data from different sources may use different Reference Coordinate Systems. The data from IPPUC used SAD69 (South American Datum from 1969), but the official standard in the country changed to SIRGAS2000 (Geocentric Reference System of the Americas) in 2015, and data from Global Positioning Systems uses WGS84 (World Geodetic System 1984). One last observation to be made is that, once street names and neighbourhoods change arbitrarily over time, they must be used with caution if considered as integration elements.

With regard to WGS84, it should be noted that starting from 1980s until 1996 there have been at least 4 different realizations: the first one was compatible with NAD83; in 1994, a new one WGS84(G730) no longer compatible with NAD83; in 1996, another one called WGS84(G873); and finally, in 2002, a new one called WGS84(G1150)). This shows the degree of complexity in managing data over time even using the "same" reference system (e.g. for historical data). For weather forecast data several reference systems can be used;

[13] http://cfconventions.org/compliance-checker.html

in this case tools for converting the data from one reference system to another one are available in the scientific community (e.g. CDO - Climate Data Operator[14]).

5.1.3 Different File Structures over Time

Consider the official data source from IPPUC released in 2012 and 2014. Consider also the data was imported into a PostGIS spatial database. There are some inconsistency issues, such as length, area or format of geometries, considering temporal changes. Table 13 presents the differences among areas for some parks available at the data.

Table 13 – Differences of park areas within 2012 and 2014

Name	2012 Area	2014 Area
PARQUE BARIGUI	4047524.92136156	4026454.21975076
PARQUE LINEAR CAJURU	636022.427397878	636022.426214896
BOSQUE DA FAZENDINHA	66157.479513349	70300.2779603284
PARQUE LINEAR DO BARIGUI	393085.586289849	614423.245830638
BOSQUE DO PILARZINHO	94860.7781767845	37799.4421067666
BOSQUE DO TRABALHADOR	427549.946190163	470151.92663829
PARQUE IGUAÇU	48230098.2717326	48834429.3685392
PARQUE ITALIANO	228068.908487592	138987.898075076
BOSQUE ITALIANO	49906.8876625597	138987.898075076
BOSQUE JOÃO PAULO II	58509.8907831777	67415.2223380618
PARQUE MUNICIPAL DO PASSAÚNA	1200924.34956277	1275161.96528643
BOSQUE ZANINELLI	40257.865889851	68543.7057435699

These area differences can be viewed in the map also. Figure 7 shows the areas of Bosque do Pilarzinho in 2012 (yellow area) and 2014 (red area). The source files were also different, presenting different columns along the years, such as the park area.

[14] https://code.zmaw.de/projects/cdo

Figure 8 – Bosque do Pilarzinho with data from 2012 (yellow area) and 2014 (red area)

5.1.4 Issues integrating Official Sources

Official data sources from the use case use data from URBS and IPPUC. Although both are official, there is no standardization in data storage, i.e. each one chooses the most convenient way to store information on the same topic. Consequently, during the integration phase, data quality problems and discrepancies between sources are found which delay the progress of the project, because they require a pre-processing to render them compatible. The most recurrent problems in data integration are listed and exemplified below :

Precision: degree of details that a source chooses to represent an entity in the database. In this type of problem (illustrated in Figure 6), each source considers different attributes to represent the same subject.

Figure 9 – Different precision for bus line data from IPPUC and URBS

```
CREATE TABLE transporte.linha_de_onibus
(
  gid serial NOT NULL,
  objectid numeric(10,0),
  layer character varying(254),
  cd_categor character varying(25),
  categoria character varying(100),
  cd_linha character varying(25),
  linha character varying(100),
  data character varying(25),
  fonte character varying(50),
  seta_senti character varying(50),
  shape_len numeric,
  geom geometry(MultiLineString,4326),
  CONSTRAINT linha_de_onibus_pkey PRIMARY KEY (gid)
)


CREATE TABLE giovane_urbs_linhas
(
  cd_linha character varying(25),
  linha character varying(100),
  somente_cartao character(1),
  categoria character varying(100),
  gid serial NOT NULL,
  geom geometry(LineString,4326),
  CONSTRAINT giovane_urbs_linhas_pkey PRIMARY KEY (gid)
)
```

Accuracy: data accuracy has a different value compared to what is shown to citizens. In this type of problem (illustrated in Figure 9), how the source stores an attribute value which is different to its value in the real world.

Figure 10 – Different names for the same bus line

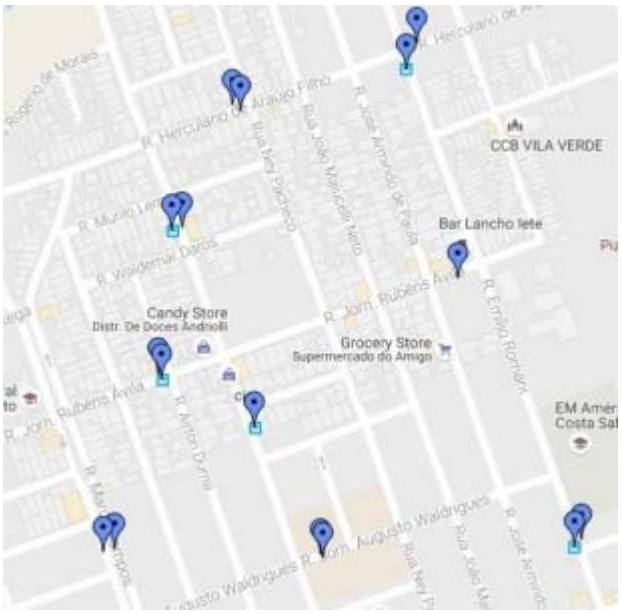
	cd_linha character varying(25)	linha character varying(100)	categoria character varying(100)
1	550	PINHEIRINHO-CARLOS GOMES	EXPRESSO LIGEIRÃO

	cd_linha character varying(25)	linha character varying(100)	categoria character varying(100)
1	550	LIGEIRÃO - PINHEIRINHO / C. GOMES	LIGEIRÃO

5.1.5 Issues integrating Official vs. NonOfficial Sources

As a particular example, consider the integration of bus stops, within official data sources and Google Maps. As you can see in Figure 10, just considering some squares, from the 16 bus stops, only 6 are available at Google Maps.

Figure 11 – Difference of bus stops from URBS and Google Maps



Note that also the GTFS from Curitiba traffic, along with streets available at Google Maps also present different information compared to the official dynamic/static data sets by the project.

5.2 Data Integration software

In the frame of the EUBra-BIGSEA T7.2 task, several components have been developed to access, integrate and curate data from the identified data sources. Those data components are available in the https://github.com/eubr-bigsea organization and are the following:

5.2.1 Data Crawlers

Available in https://github.com/eubr-bigsea/crawlers, they comprise a set of Python programs to crawl information from facebook, globo and google events, bus position and static information from Curitiba, Rio and Sao Paoulo, City Hall information from Curitiba, Points of Interest from Curitiba and TripAdvisor, traffic status and weather.

The code uses a ZABBIX protocol to implement the communication with the sources and stores the results on a MongoDB database. Each module has its own syntax, but some arguments are common to them:

- -s, --server, Name of the MongoDB server (required)
- -p, --persistence, Name of the MongoDB persistence slave (optional)
- -d, --database, Name of the MongoDB database (required)
- -t, --sleep_time, Time sleeping seconds (required)
- -u, --urbs_key, URBS access code (required)
- -f, --control_file, File to control the execution time (required)
- -z, --zabbix_host, Zabbix host for monitoring (required)

5.2.2 Twitter and Facebook integration

Twiter characterization is available in https://github.com/eubr-bigsea/tweets-characterization, it contains three pieces of code to obtain information of geo-located tweets, generate statistics or dump data from different Brazilian cities, such as Belo horizonte, Curitiba, Fortaleza, Manaus, Porto Alegre, Recife, Rio de Janeiro, Sao Paulo, Salvador or all of them in a single interaction.

The module for twitter characterization is written in python and has the following syntax:

python automatic_tweets_json_v1.py -i [1] -o [2] -c [3]

Where the three first arguments are required. The arguments are: [1] is a JSON file of the tweets with the selected fields, [2] is the name of the output file, [3] is the name of the city to be evaluated. Additionally, the user can specify three optional parameters: A .csv File [4] with the ID of each selected user, the initial date [5] and the ending date [6].

The program for extracting statistics expects two arguments:

python statistics.py -i [1] -p [2]

Where [1] is the JSON input file and [2] is a file that owns information about selected users of twitter

Finally, there is a function for dumping specific fields in a MongoDB:

python script_datedump_selectfields.py -s [1] -p [2] -d [3] -c [4] -sd [5] -ed [6] -o [7] test

Where the all the arguments but the second are required and mean: [1] the name of the MongoDB server; [2] the name of the MongoDB persistence slave; [3] the name of the MongoDB database; [4] the name of the MongoDB collection; [5] the date when a project or task is scheduled to begin/start that define the begin of the dump (on the format: "2016-03-30 00:00:00", hour defined by UTC); [6] date when a project or task is scheduled to finish/end that define the end of the dump (in the same format as [5]); and [7] the name of the output file.

Tweeter clustering is available in https://github.com/eubr-bigsea/Tweets-cluster, which provides tools for clustering using K-Means or DBSCAN. The interface of the three Python codes are the following:

- Unsupervised learning at word level, word2vec.py.
- Unsupervised learning at document level, doc2vec.py.
- TF-IDF (*Term frequency – Inverse document frequency*), tf_idf.py.

The code includes a sample set for validation.

Finally, a data gathering web service has been developed to retrieve Facebook events by location. The service is available in https://github.com/eubr-bigsea/facebook-events-by-location.

This implementation uses regular Facebook Graph API calls in a three-step approach to get the events:

1. Search for places in the radius of the passed coordinate and distance (/search?type=place&q=*&center={coordinate}&distance={distance})
2. Use the places to query for their events in parallel (/?ids={id1},{id2},{id3},...)
3. Unify, filter and sort the results from the parallel calls and return them to the client

The service is made available as a Docker microservice that can be built via docker build -t <yourTag> . locally if you like. The API exposes through the basic endpoint /events, the following querying parameters.

Mandatory parameters:

- lat: The latitude of the position/coordinate the events shall be returned for
- lng: The longitude of the position/coordinate the events shall be returned for
- distance: The distance in meters (it makes sense to use smaller distances, like max. 2500)
- access token: The **App Access Token** to be used for the requests to the Graph API

Non-mandatory parameters:

- sort: The results can be sorted by time, distance, venue or popularity. If omitted, the events will be returned in the order they were received from the Graph API

5.2.3 Data visualization

Forked in https://github.com/eubr-bigsea/transit-map, this code shows vehicles (markers) on a map using the public transport timetables to interpolate their positions along the routes (polylines).

A Python API is available in https://github.com/eubr-bigsea/transit-map-api.

6 CONCLUSIONS

Multiple data sources have been inventoried, characterized and analyzed with respect to their format, accuracy and coverage. This data sources are being considered for the development of the applications in the WP7 work package, although they are also of great importance for other WPs, as WP4 and WP5.

The document is a guide for the consortium to locate the data and pieces of software developed in the frame of T7.2, and it will be used as a basis for the T7.3, T7.4 and T7.5. In addition, the document presents initial integration problems already found by the project.

7 GLOSSARY

Acronym	Explanation	Usage Scope
API	Application Programming Interface	WP3/WP4/WP5
DBMS	Data Base Management System	WP7/WP4
GES3	Georeferenced Environmental, Stationary, Dynamic and Social Data	WP7
GTFS	Generic Transit Feed Specification	WP7
IPPUC	Institute of Research and Planning of Curitiba	WP7
JSON	JavaScript Object Notation	WP4/WP5/WP7
K-Means	Data mining clustering algorithm.	WP7
DBSCAN	Density-based spatial clustering of applications with noise	WP7
NAD83	North American Datum 1983	WP7
NLP	Natural Language Processing	WP7
IBGE	Instituto Brasileiro de Geografia e Estatística	WP7
SAD69	South American Datum from 1969	WP7
TF-IDF	Term frequency – Inverse document frequency	WP7
UTC	Coordinated Universal Time	WP7
WGS84	World Geodetic System 1984	WP7