HARMONY

Holistic Approach for Providing Spatial & Transport Planning Tools and Evidence to Metropolitan and Regional Authorities to Lead a Sustainable Transition to a New Mobility Era

D3.2 - Transport and Spatial Data Warehouse Technical Design

@Harmony_H2020

#harmony-h2020

https://www.linkedin.com/company/harmony-h2020/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Part of: CIVITAS

THE CIVITAS INITIATIVE IS CO-FINANCED BY THE EUROPEAN UNION

SUMMARY SHEET

PROJECT

Project Acronym:	HARMONY
Project Full Title:	Holistic Approach for Providing Spatial & Transport Planning Tools and Evidence to Metropolitan and Regional Authorities to Lead a Sustainable Transition to a New Mobility Era
Grant Agreement No.	815269 (H2020 – LC-MG-1-2-2018)
Project Coordinator:	University College London (UCL)
Website	www.harmony-h2020.eu
Starting date	June 2019
Duration	42 months

DELIVERABLE

Deliverable No. - Title	D3.2 - Transport and Spatial Data Warehouse Technical Design
Dissemination level:	Public
Deliverable type:	Demonstrator
Work Package No. & Title:	WP3 - Data collection tools, data fusion and warehousing
Deliverable Leader:	ICCS
Responsible Author(s):	Efthimios Bothos, Babis Magoutas, Nikos Papageorgiou, Gregoris Mentzas (ICCS)
Responsible Co-Author(s):	Panagiotis Georgakis (UoW), Ilias Gerostathopoulos, Shakur Al Islam, Athina Tsirimpa (MOBY X SOFTWARE)
Peer Review:	Panagiotis Georgakis (UoW), Ilias Gerostathopoulos (MOBY X SOFTWARE)
Quality Assurance Committee Review:	Maria Kamargianni, Lampros Yfantis (UCL)

DOCUMENT HISTORY

Version	Date	Released by	Nature of Change
0.1	02/12/2019	ICCS	ToC defined
0.3	15/01/2020	ICCS	Conceptual approach described
0.5	10/03/2020	ICCS	Data descriptions updated
0.7	15/04/2020	ICCS	Added sections 3 and 4
0.9	12/05/2020	ICCS	Ready for internal review
1.0	01/06/2020	ICCS	Final version

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

TABLE OF CONTENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Explanation
AV	Autonomous Vehicle
BLOB	Binary Large OBject
CRISP-DM	Cross-industry standard process for data mining
CRUD	Create - Read - Update - Delete
Dx.y	Deliverable x.y
ETL	Extract-Transform-Load
GML	Geography Markup Language
GTFS	General Transit Feed Specification
HTTP	Hypertext Transfer Protocol
ITF	International Transport Forum
JSON	JavaScript Object Notation
KDD	Knowledge Discovery Databases
LUTI	Land Use Transport Interaction
MaaS	Mobility-as-a-Service
MB	Megabyte

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

MDM	Multi-Dimensional Data Modelling
MS	Model Suite
NoSQL	non Structured Query Language or non relational
ODBC	Open Database Connectivity
OECD	Organisation for Economic Co-operation and Development
REST	REpresentational State Transfer
SEMMA	Sample, Explore, Modify, Model, Assess
SQL	Structured Query Language
SSD	Solid State Drives
TSDW	Transport and Spatial Data Warehouse
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
WPx	Work Package X
3Vs	Volume, Velocity, Variety

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

EXECUTIVE SUMMARY

Within HARMONY, multidisciplinary data are combined and integrated for land-use and transport modelling and planning. Data are managed in a Transport and Spatial Data Warehouse (TSDW) that stores secondary and primary data for analysis, feeds with data the HARMONY simulators and stores the data generated by them. This deliverable provides the technical design of the TSDW addressing requirements of the components of the HARMONY model suite. In this respect, the following design elements of the TSDW are considered in this deliverable:

- The specification of the required data schemas for storing raw and processed transport and spatial data. Starting from the conceptual architecture of the HARMONY Model Suite (MS) (deliverable D1.3) a thorough record of data produced, consumed and exchanged between the simulators has been compiled. This set of data has been analysed in order to derive a harmonized data model where the main data entities of the model suite have been defined together with their properties.
- The initial approach of the TSDW infrastructure which covers the following points:
  - Processing and storage capacity. Based on the identified data structures and the expected volume and data processing functions, preliminary figures of the required processing and storage capacity have been identified.
  - Communication protocols interacting with external data providers and consumers.
  - Interfaces for read/write access and import/export of transport and spatial data from external sources and outputs of the HARMONY MS.
  - Data processing and management pipeline including the selection of data storage and processing technologies.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

1 Introduction

1.1 Overview

One of the main objectives of the HARMONY project is to combine and integrate multidisciplinary data required for transport modelling and planning. The objective is mainly addressed in Work Package 3 (WP3) “Data collection tools, data fusion and warehousing” where a Transport and Spatial Data Warehouse (TSDW) that stores secondary and primary data for analysis, feeds with data the HARMONY models and stores data generated by them, is being developed. This deliverable is the second of WP3 and is the outcome of task T3.2 “Technical design of HARMONY data warehouse and harmonisation”.

The main goal of T3.2 is to provide the initial technical design of the HARMONY TSDW which will allow to proceed with the technical implementation of the warehouse as well as its population with data and information to be used by the HARMONY models. The design mainly addresses requirements of the components of the HARMONY model suite. In this respect, the following design elements are considered and addressed in this deliverable:

- The specification of the required data schemas for storing raw and processed transport and spatial data. Starting from the conceptual architecture of the HARMONY Model Suite (MS) (deliverable D1.3) a thorough record of data produced, consumed and exchanged between the simulators has been compiled. This set of data has been analysed in order to derive a harmonized data model where the main data entities of the model suite have been defined together with their properties.
- The initial approach of the TSDW infrastructure which covers the following points:
    - Processing and storage capacity. Based on the identified data structures and the expected volume and data processing functions, preliminary figures of the required processing and storage capacity have been identified.
    - Communication protocols interacting with external data providers and consumers.
    - Interfaces for read/write access and import/export of transport and spatial data from external sources and outputs of the HARMONY MS.
    - Data processing and management pipeline including the selection of data storage and processing technologies.

1.2 Requirements of the TSDW

The HARMONY TSDW is being developed to meet the needs and functional/non-functional requirements of the simulators within the MS. The set of requirements was derived during the definition of the conceptual and technical architecture of the HARMONY MS. Based on the identified requirements, the TSDW should:

- Be able to accommodate data for transport supply and demand, simulated travelers (agents) as well as census data mainly used for strategic/ long term simulations.
- In terms of types of data, the HARMONY TSDW should support the storage, processing and retrieval of:
    - Structured and semi-structured data in the form of CSV, JSON, Excel
    - Operational data including simulation workflow logs
    - Binary Large OBjects (BLOBS), including shapefiles
- Provide capabilities for geospatial information storage and retrieval.
- Allow interested third parties to access the stored outputs of the HARMONY MS simulators.
- Rely on open data standards as much as possible and provide standardized access interfaces.
- Support data access and retrieval for components implemented in varying programming languages and which rely on different technologies (i.e. a software-agnostic model suite)
- Provide adequate capacity and storage as well as be scalable to support large and increasing volumes of data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269
Part of: 2020 CIVITAS

1.3 Conceptual Approach

The conceptual approach of the HARMONY TSDW is presented in Figure 1. The TSDW is unique in its kind as it aims to support modelling and simulation of land-use and transport at different levels (Strategic, Tactical and Operational). It provides the necessary input data for simulating a range of scenarios involving new mobility services and stores the output data / simulation outcomes which will support transport and planning-related decisions. Essentially, it is a platform that integrates systems able to store data inputs and outputs, data harmonization and transformation tools and the required interfaces that provide access to these data in formats that are required by the simulators or data analysis applications through a set of APIs and/or direct database connectors.

In the proposed approach, special considerations have been put in place to support the messaging middleware that has been proposed in HARMONY Model Suite Technical Architecture. Note that based on the current understanding, data at rest will be stored in the TSDW while data in motion will be exchanged through the interfaces defined in HARMONY's architecture using the messaging middleware described in D2.1.

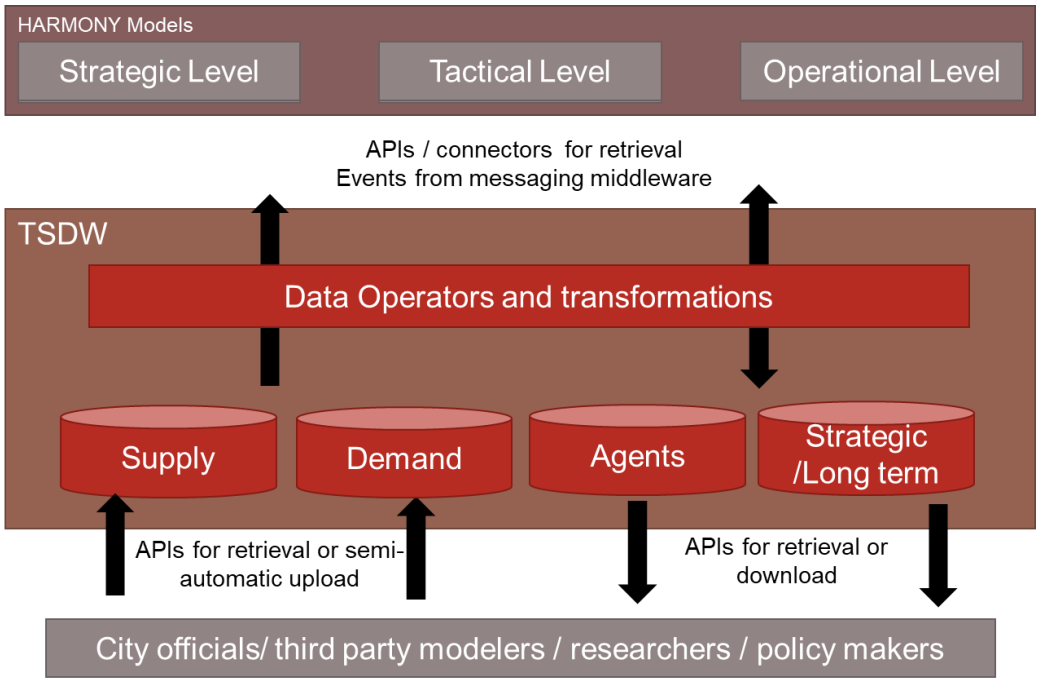

Figure 1: HARMONY TSDW conceptual approach.

Based on HARMONY's approach, modelling is performed at three levels namely the Strategic, the Tactical and the Operational. Specific modelling and simulation components are implemented at each level generating simulation outputs which are used either by other components or are stored for further analysis by transport modellers. In more details:

- The Strategic Level operates at longer time horizons, which usually corresponds to one to five years. The simulators of this level are responsible for generating (i) disaggregate household and firm population and the locations for different types of activities such as employment, housing, and education, (ii) aggregate commodity flows between employment sectors, (iii) future employment including services, health, and educational activities, and (iv) long-term mobility choices of individuals (agents) including car-ownership or subscriptions to different mobility services.
- The Tactical Level operates at mid-range time horizons, which usually corresponds to a day or week. The simulators of this level are responsible for generating (i) disaggregated travel demand in the form of agents' daily activity schedules (trip-chains) and (ii) disaggregated demand in the form of freight vehicle tours (i.e. trucks, vans, freight bikes, etc.) and their corresponding trips.
- The Operational Level operates at the shortest time horizon facilitating within-day simulations disaggregated to steps lasting a fixed number of seconds, or minutes. The simulators of this level

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

are mainly responsible for (i) loading the travel demand into multi-modal networks, while simultaneously capturing travellers' dynamic schedule re-evaluation choices due to supply conditions, (ii) managing the operations of traditional and new mobility services for passengers and freight (e.g. AVs, MaaS, crowdshipping), (iii) generating within-day performance indicators such as travel times, average speeds, traffic volumes, impedance measures, environmental and consumption measures and others.

The components of each level have been identified in the HARMONY technical architecture and are shown in Table 1. In the following sections the components are referenced with their ID provided in the first column of Table 1. Further details per component are provided in deliverables D2.1 and D1.3.

Table 1: The components of the HARMONY model suite.

ID	Component
Strategic Level	
S.1	Demographic forecasting module
S.2	Regional economy module
S.3	Land Use Transport Interaction model (and land supply constraints and Dev sub-model)
S.4	Vehicle ownership module/Mobility Subscription Models
S.5	Spatial interaction freight model
S.6	Aggregate to disaggregate population and employment translator (Synthetic population generator)
Tactical Level	
T.1	Tactical freight simulator
T.2	Tactical passenger simulator
T.3	Day-to-day learning module
Operational Level	
O.1	Traffic flow simulator/ Multi-class Network Model /Traffic Assignment Model
O.2	Freight service controller system
O.3	Passenger service controller system
O.4	Within-day re-evaluation models
O.5	Energy and emission model
O.6	Noise model

1.4 Structure of the deliverable

The remainder of this deliverable is structured as follows. Section 2 provides an overview of the main data schemas which are required by the simulators and will be managed by the TSDW. Section 3 focuses on data management and processing pipelines and describes different approaches and theories, concluding on the envisaged approach that will be followed in the HARMONY TSDW. Section 4 provides an overview of the TSDW infrastructure design and its various elements, including the selection of the database systems to be used, the communication protocols and interfaces, the envisaged deployment approach as well as the non-functional features, including compute-storage-memory requirements, security and privacy related aspects. The document concludes in section 5.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

2 HARMONY Main Data Schemas

In this section, the main data entities of the HARMONY MS are described along with their properties. The HARMONY MS data entities were identified through a structured process, in collaboration with the transport engineers who develop the models of the HAMRMONY MS. More specifically, a template was defined asking different partners to provide the data inputs and outputs of the modelling components they develop. The process was initiated as part of the technical architecture and was extended in order to be able to gather further information related to the TSDW. The template was distributed as an online google sheet[1] and the information that partners provided is the basis for the TSDW. The information that was provided include the name and description of the data, their format and communication frequency, the expected size in MB, an example and whether the data should be retrieved from a database or flow directly between simulators.

The data entities provided by the different partners were subsequently grouped in four main categories:

- Supply data, including information mainly related to the transport network, passenger and freight service infrastructure.
- Demand data, including information mainly related to passenger transport demand for an area, i.e., Traffic Analysis Zone, and freight demand.
- Long-Term data, including information required by the land-use, economic, demographic forecasting and population synthesis models at the Strategic Level.
- Auxiliary data which contain definitions of data entities that are common in the above categories.

In the following sections a description of the identified data entities is provided, including their description, their properties, which HARMONY components use the related entity as input and which provide it as output. The components are referenced with their ID provided in the first column of Table 1. Where applicable, the relationships between the data entities are identified and reported. Note that not all data entities described below will be stored in the TSDW. Depending on the needs of the simulators and the requirements for reporting simulation outcomes, a selection of data to be stored in the TSDW will be performed during the integration phase of the project. Based on the current understanding data at rest will be stored in the TSDW while data in motion will be exchanged through the interfaces defined in HARMONY's architecture using the messaging middleware (see HARMONY deliverable D2.1).

2.1 Supply Data

Entity: Road_and_public_transport_network

Description: Road network information including links for public transport, lanes, turning information and segments. Commonly represented in shape files.

Input to: O1, O2, O3.

Properties

- Links_for_road_network:
  - Links are contiguous stretches of single directional roads
- Links_for_public_transport_network:
  - includes bus links, rail links and walk links. If a bus/train-line serves bus-stop/train station a and then eventually bus-stop/train station b, there is an link (a,b) in the public transport network representing this bus-line service
- Lanes
  - Part of segments that are designated for use by a single line of vehicles
- TurningInformation
  - Turning path is a path connecting specific lanes of two connected links.

[1] https://docs.google.com/spreadsheets/d/1fAjPR7cjxrNiOq7CXF-TmN8RiHNUO8URV0SMhVpsmmw/edit#gid=1119022090

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- Nodes
  - The end points of links. Node type include sink, source, intersection, merging, diverging.
- Connectors
  - Connection between lanes of two consecutive segments within the same link
- Segments
  - Sub-divisions of a link. These are typically based on changes in the link geometry, such as change in the number of lanes
- Polyline
  - A sequence of points defining the shape of road network constructs such as segments, lanes, links and turning paths
- Intersection Control Data
  - Information related to the control of intersections

Entity: Public_Transport_Data

Description: Follows the General Transit Feed Specification GTFS format. A GTFS feed is composed of a series of text files collected in a ZIP file. Each file models a particular aspect of transit information: stops, routes, trips, and other schedule data.

Input to: O1

Properties

- Agency
  - Information about the transit agency
- Stops
  - Information about individual locations where vehicles pick up or drop off passengers.
- Routes
  - Information about a transit organization's routes. A route is a group of trips that are displayed to the rider as a single service.
- Trips
  - Information about scheduled service along a particular route. Trips consist of two or more stops that are made at regularly scheduled intervals
- Stop_Times
  - Lists the times that a vehicle arrives at and departs from individual stops for each trip along a route.
- Calendar
  - Defines service categories. Each category indicates the days that service starts and ends as well as the days that service is available
- Calendar_Dates
  - Lists exceptions for the service categories defined above.
- Fare_Attributes
  - Defines fare information for a transit organization's routes.
- Shapes
  - Provides rules for drawing lines on a map to represent a transit organization's routes
- Frequencies
  - Provides the headway (time between trips) for routes with variable frequency of service.

Entity: Parking_spot

Description: As the titles states this entity refers to parking spots information. Information regarding the presence of EV chargers is also included.

Input to: O3

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Properties

- Type
  - Categorical: on street (PCU parking, Freight bay, Bus bay, Taxi stand, Double parking) or off-street (Bus terminal, Parking lot inside and outside buildings for freight and passenger vehicles)
- Coordinates
  - Parking spot coordinates
- Segment_id
  - Road network segment
- Capacity
  - Number of parking slots available
- Number_of_EV_chargers
  - Number of EV chargers in the parking spot.

Entity: On_demand and Sharing_service

Description: Information about sharing services such as car sharing and bike sharing

Input to: O1

Properties

- Type
  - Sharing service type such as car-sharing and bike-sharing, taxi, TNCs
- Stations
  - Including station codes, station names and gps co-ordinates of the sharing service stations
- Fleet
  - Including ids and gps co-ordinates of the sharing service fleet
- Area covered by the service (when free floating)

Entity: Vehicle_fleet_characteristics

Description: Characteristics of available freight fleets

Input to: O1

Properties

- Mode
  - References the Transport_Mode which is described in the Auxiliary Data
- Fleet_size
  - Number of vehicles included in the vehicle fleet of the different services provided by the Freight controller
- Capacity
  - Capacity of the vehicles included in the vehicle fleet of the different services provided by the Freight controller
- Speed
  - Speed of the vehicles included in the vehicle fleet of the different services provided by the Freight controller

Entity: Terminal_locations

Description: Freight terminals locations

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Input to: O1

Properties

- Terminal_id
  - Unique if of the terminal
- Coordinates
  - Lat/long format

Entity: Major_future_transport_infrastructure_development

Description: Locations of future transport infrastructures

Input to: O1

Properties

- Infrastructure_type
  - Including drones landing points, flying taxi landing points, new metro stations, etc.
- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Major_Infrastructure_Development
  - Shapefile with major rail, metro and road planned developments,

Entity: Free-flow_road_travel_time

Description: Road travel time in free-flow conditions

Input to: O1

Properties

- link_id,
  - Reference to Network_Link
- travel time
  - Road travel time in free-flow condition

Entity: Historical_Road_link_travel_times

Description: Travel times in the different links of the network at each simulation step

Input to: O3

Properties

- link_id,
- downstream link_id,
- simulation step start_time,
- simulation step end_time,
- travel time

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Entity: Commodity_Flow

Description: Matrix with commodity flows between NUTS 3 region, Sum of commodity flows by goods type between regions.

Input to: T1

Properties

- region_id
  - Reference to region in the road network
- good_type
  - References to Good in the Auxiliary Data
- sum_of_flows
  - Indicates the flows to the region

2.2 Demand Data

Entity: Trip_OD_Matrix

Description: Origin-Destination data representing movement by mode of transport through geographic space.

Input to: O1

Properties

- id
- origin_zone_code
  - Reference to a zone in the network model
- destination_zone_code
  - Reference to a zone in the network model
- Mode of transport
  - Reference to a Transport_Mode
- Number_of_trips
  - The number of trips from origin to destination

Entity: Customer_demand

Description: Customer demand in terms of package size for last mile delivery, relevant for the operational freight controller.

Input to: O2

Properties

- Customer_id
- Package_size
  - Size of the requested package
- Coordinates
  - Lat/long format

Entity: Vehicle_route

Description: Location and routes of the individual vehicles in the network

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Input to: O2

Properties

- VehicleID
- List_of_street_ids
  - Reference to the property Links_for_road_network of the Road_and_public_transport_network entiry

Entity: Freight_demand

Description: Represents the demand for freight and is relevant for the operational freight controller.

Input to: O2

Properties

- Tour_id
- Trip_id
- Destination
- Coordinates
- Vehicle_type
  - Reference to Vehicle_type, e.g. cargo bike, etc, in the Auxiliary Data

Entity: Traffic_flow_assignment_by_link_by_mode

Description: Traffic volumes on each link by mode.

Output of: O1

Properties

- Link_code
- Mode
  - Reference to Transport_Mode in the Auxiliary Data
- Traffic_volume

Entity: Travel_time_assignment

Description: Origin-Destination travel time between zones by mode of transport

Output of: O1

Properties

- Origin_zone_code
- Destination_zone_code
- Mode
  - Reference to Transport_Mode in the Auxiliary Data
- Travel_time

Entity: Travel_cost_assignment

Description: Origin-Destination travel costs between zones by mode of transport.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Output of: O1, S2

Properties

- Origin_zone_code
- Destination_zone_code
- Mode
  - Reference to Transport_Mode in the Auxiliary Data
- Travel_cost

Entity: Traffic_speed_assignment

Description: Traffic speed on each link by mode resulting from the assignment of traffic

Output of: O1

Properties

- Link_code
- Mode
  - Reference to Transport_Mode in the Auxiliary Data
- • Traffic_speed

Entity: Fleet_schedule

Description: The trip chain of each vehicle of a freight fleet.

Output of: O2

Properties

- Tour_id
- Trip_id
- Destination
- Coordinates
- Vehicle_Type

Entity: Dynamic_OD_matrix_per_vehicle_type

Description: Indicates movements of different vehicle types.

Input to: S3

Properties

- Origin
- Destination
- Vehicle_type_id
  - Reference to Vehicle_type in the Auxiliary Data

Entity: Skim_matrix

Description: Provides impedances between zones.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Input to: S3

Properties

- sm_record_id
- origin_zone_id
  - Reference to a zone in the network model
- destination_zone_id
  - Reference to a zone in the network model
- Distance
  - Distance between the zones
- distance_units
  - ["km", "uk_mile" , "us_mile" , …]

Entity: Realized_Trip_service_information

Description: The details of a trip performed by an Agent.

Output of: O3

Properties

- Trip ID
- Trip Origin
- Trip Destination
- Taxi ID
- Rq time
- Rq location
- Dp Departure time
- Dp Arrival time
- Dp Origin
- Dp Destination
- Dp delay
- Avg.Dp.Speed
- Pu time
- Pu location (Origin)
- Wt time
- Usage
- Do time
- Do location (Destination)
- Travel-time
- Delay
- Mode

Entity: Daily_Activity_Schedule

Description: A chain of activities performed by an Agent.

Output of: T2

Properties

- Agend_id
- Tour_number
- Tour_type (e.g. work, education, etc.)
- Stop_number
- Stop_type

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- Stop_location
- Previous_location
- Previous_zone
- Stop_zone
- Time_use
- Mode
- Arrival_time
- Departure_time
- **Previou**s_stop_departure_time

Entity: Agent_Perceived_Travel_Times

Description: Pre-day expected agent travel times.

Input to: T3

Properties

- agent_id
- Transport_Mode_code
  - Reference to Transport_Mode in the Auxiliary DAta
- Origin
- Destination
- time

Entity: Traffic_Analysis_Zone

Description: Traffic analysis zone defined by Land Transport Authority

Input to: T3

Properties

- Zone id
- Area
- Population
- Shops
  - Number of shops in the TAZ
- Parking_rate
- Resident_workers
- Employment
- Students

Entity: Agent_parking_spots

Description: Indicates places which are used by an agent as parking spots.

Input to: T2

Properties

- agent_id
  - Reference to an Agent
- Spot_type
  - Parking at home or work

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- Price
  - Parking price
- Spot_location
  - Coordination of the parking spot

Entity: Agent_Vehicle_fleet

Description: Vehicles owned by agents

Input to: T2

Properties

- **agent_id**
- **vehicle_id**
- **Vehicle_type_id**
  - **Reference to Vehicle_Type from the Auxiliary Data**
- **Primary_driver_id :**
  - **allocation of a vehicle to a specific person in the household**

2.3 Strategic/Long Term Level Data

Entity: Agent

Description: Representation of a person/traveler.

Input to: T2

Properties

- Id
- Household_income_category_id
  - Reference to Household_Income_category the household incomescale variable of income (or categories)
- Residential_tenure_id
  - Reference to Residential_tenure
- Housing_type_id
  - Reference to Housing_type
- Annual_household_kms
  - Scale variable of annually driven kms
- Driver_licence
- PT_subscription
- Other_mobility_service_subscription
  - e.g. MaaS
- PT_or_other_pass_pricing
  - Scale variable of annually/monthly price

Entity: Student

Description: Subclass of Agent who is a student.

Input to: T2

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Properties

- id
- agent_id
  - Reference to Agent
- is_school_sudent :
  - Status indicating whether an agent participates in school activity
- school_location
  - Coordinates of the Agent's school
- school_type_id
  - Reference to School_Type in the Auxiliary Data
- is_university_student
  - Status indicating whether an agent participates in university activity
- university_location
  - Coordinates of the university
- iniversity_type_id
  - Reference to University_Type in the Auxiliary Data

Entity: Worker

Description: Subclass of Agent that participates in the labor force

Input to: T2

Properties

- Id
- Agent_id
  - Rereference to Agent
- Occupation_industry_id
  - Reference to Occupation_Industry in the Auxiliary Data
- Work_location
  - Coordinates of the Agent's work location
- Work_shift_category_id
  - Reference to work_shift_category in the Auxiliary Data
- Work_duration
  - Work duration in minutes
- Worker_temporal_ flexibility
- Worker_spatial_flexibility

Entity: Employment

Description: Indicates number of people aged 16 to 74 employed per occupation, industry, wage level, floorspace.

Input to: S1

Properties

- Geography
  - The geography or area the data refer to.
  - Reference to Geography in the Auxiliary Data
- Occupation_Type
  - Reference to OccupationType in the Auxiliary Data
- Industry_Type
  - Reference to IndustryType in the Auxiliary Data
- Wage

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- Reference to Wage
- Floorspace
  - the average floorspace (in m²) per Full-Time Equivalent (FTE) member of staff. It is used as a measure of employment density
- Employment_value
  - number of people aged 16 to 74 employed per occupation, industry, wage level, floorspace

Entity: Wage

Description: Estimates of annual household income for the four income types per area.

Input to: S1, S6

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Financial year
  - The Financial year the data refer to
- Total_annual_income
  - The total annual household income
- Net_annual_income
  - The net annual household income
- Net_income_before_housing
  - The Net annual household income (equivalised) before housing costs
- Net_income_after_housing
  - The Net annual household income (equivalised) after housing costs

Entity: Retail_Activities

Description: Retail activities disaggregated by zone, floorspace, retail sales, employment, sales/expenditure flows.

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Employment
  - Reference to Employment
- Floorspace
  - the average floorspace (in m²) per Full-Time Equivalent (FTE) member of staff. It is used as a measure of employment density
- Retail_sales
- Sale_flows
- Expenditure_flows

Entity: Land_Use_Data

Description: Indicates land use e.g. industrial, commercial, office, residential, correlated with activity data.

Input to: S1

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Land_use_type
  - Reference to LandUseType in the Auxiliary Data
- ActivityData
  - **Economic** activity data

Entity: Population

Description: Usual resident population, resident population in employment, usual workplace population.

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Resident _Population
  - Usual resident population
- Employed_Resident_Population
  - resident population in employment
- Usual_Workplace_Population
  - usual workplace population

Entity: Population_Projection

Description: Estimated future resident population; a subclass of the Population entity

Input to: S1

Properties

- PopulationId
  - Reference to Population entity data
- projection_year
  - the future year the estimated data refer to

Entity: Income

Description: Indicates the mean and median income

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Mean_income
  - mean equivalised household disposable income

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- Median_income
  - median equivalised household disposable income

Entity: Urban_density

Description: Indicates the urban density per area

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Density
  - people per sq. km

Entity: Firm_Data_Industry

Description: Firm data by industry

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Industry
  - Reference to IndustryType in the Auxiliary Data
- Productivity
- Earnings
- Foreign_investment
- Research_Development

Entity: Housing_Tenure

Description: Housing tenure household percentages by type

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Owner_occupied
  - percentage of houses occupied by their owner
- private_renting
  - percentage of houses occupied by private renting
- social_renting

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- percentage of houses occupied by social renting

Entity: House_Prices

Description: Indicates the average price paid for houses per area

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- House_Prices
  - Average Price Paid

Entity: House_Rents

Description: Indicates the average rent prices per area

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- House_Rents
  - Average rent prices

Entity: General_Topology

Description: Indicates the general topology of the area, e.g. urban area; land/coastline geography; rivers and waterbodies.

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Topology_Type
  - Types include urban area; land/coastline geography; rivers and waterbodies

Entity: General_Topology

Description: Indicates the municipality/metropolitan area boundaries.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Municipality
  - the municipality the data refer to

Entity: Local_Authority_Boundaries

Description: Indicates the county/district boundaries.

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Authority_ID
  - the County/District the data refer to

Entity: Building_Footprints_Outline

Description: Indicates the outline of building footprints in an area.

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Building_Footprints
  - Shapefile with building footprints; commercial floorspace or building heights; residential units

Entity: Building_ Functions

Description: Indicates the building functions in an area

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- Building_ Functions
  - Shapefile with buildings function data- office, retail, industrial, residential

Entity: Public_Housing_Estates

Description: Indicates the location of large public housing estates

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Public_Housing_Estates
  - Shapefile with the public housing estates

Entity: Future_Urban_Development

Description: Indicates the location of new towns; opportunity areas; areas for densification

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Major_Future_Urban_Development
  - Shapefile with the location of new towns; opportunity areas; areas for densification

Entity: Planning_Zoning

Description: Indicates the zoning restrictions on urban development

Input to: S1

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Planning_zoning
  - Shapefile with the location of zoning restrictions on urban development

Entity: Environmental_restrictions

Description: Indicates the zoning restrictions on urban development for environmental purposes

Input to: S1

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Properties

- Geography
  - The geography or area the data refer to
  - Reference to Geography in the Auxiliary Data
- Year
  - the year the data refer to
- Environmental_restrictions
  - Shapefile with green belt restrictions; national park/reserve restrictions; flooding development restrictions

Entity: Population_by_zone

Description: Indicates total population by zone

Input to: S2

Properties

- Zone
  - Reference to a zone in the network model
- Inhabitants
  - number of inhabitants

Entity: Floorspace_Prices

Description: Indicates floorspace price for houses and commercial buildings by zone

Input to: S2

Properties

- Year
  - the year the data refer to
- Zone
  - Reference to a zone in the network model
- HousePrice
  - floorspace price for houses
- CommercialBuildingsPrice
  - floorspace price for all the commercial buildings
- IndustrialPrice
  - floorspace price for industrial buildings only
- OfficePrice
  - floorspace price for offices only
- ShopPrice
  - floorspace price for shops only

Entity: CarOwnership

Description: Indicates car ownership by zone

Input to: S2

Properties

- Zone
  - Reference to a zone in the network model

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- MotorisationRate
  - motorisation rate

Entity: PTInfraInvestments

Description: Indicates the (public) economic overall effort in the region for public transport infrastructure

Input to: S2

Properties

- Investments
  - (public) economic effort in the overall study area for transport infrastructure

Entity: NationalGDP

Description: Indicates the national gross domestic product

Input to: S2

Properties

- GDP
  - National gross domestic product

Entity: OutsourcingIndex

Description: Indicates the outsourced (out of the region) effort made by industry

Input to: S2

Properties

- zone
  - Reference to a zone in the network model
- ProductiveSector
  - Reference to IndustryType in the Auxiliary Data
- Index
  - Index capturing the % of outsourced (out of the region) effort made by industry

Entity: TourismSupply

Description: Indicates the tourism supply per zone

Input to: S2

Properties

- zone
  - Reference to a zone in the network model
- Bedplaces
  - total bedplaces by zone

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Entity: TourismDemand

Description: Indicates the tourism demand per zone

Input to: S2

Properties

- zone
  - Reference to a zone in the network model
- Year
  - the year the data refer to
- Volume
  - total volume of tourists by zone
- Revenue
  - total tourism revenue by zone

Entity: NetDisposableIncome

Description: Indicates the available income of households to invest, save, or spend per zone

Input to: S2

Properties

- zone
  - Reference to a zone in the network model
- Year
  - the year the data refer to
- Income
  - available income of households to invest, save, or spend after taxes and necessities paid and transfers made

Entity: GrossCapitalFormation

Description: Indicates the Regional Gross Fixed Capital Formation

Input to: S2

Properties

- NUTS2Zone
  - Reference to a NUTS2Zone in Auxiliary data
- Year
  - the year the data refer to
- GFCF
  - Regional Gross Fixed Capital Formation

Entity: EmploymentIndustry

Description: Indicates Employment by Industry

Input to: S2

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Properties

- zone
  - Reference to a zone in the network model
- EmploymentEntity
  - It incorporates all the properties of the Employment entity
- IndustryType
  - Reference to IndustryType in Auxiliary data
- Employment_value
  - number of jobs for each sector

Entity: HouseholdsNumber

Description: Indicates the number of households per analysis zone

Input to: S6

Properties

- Zone REF Zone
- Year
  - The year the data refer to
- Number

Entity: HouseholdComposition

Description: Indicates the mean number of persons in a household per zone

Input to: S6

Properties

- Zone
- Year
  - The year the data refer to
- Persons

Entity: HousingType

Description: Indicates the housing types per zone

Input to: S6

Properties

- Zone
- Year
  - The year the data refer to
- Housing_type

Entity: Household_Vehicles

Description: Indicates the mean number of vehicles per household per zone

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Input to: S6

Properties

- Zone
- Year
  - The year the data refer to
- Vehicles

Entity: PopulationWorkers

Description: Indicates the mean number of workers per zone

Input to: S6

Properties

- Zone
- Year
  - The year the data refer to
- Workers

Entity: PopulationStudents

Description: Indicates the mean number of students per zone

Input to: S6

Properties

- Zone
- Year
  - The year the data refer to
- Students

Entity: PopulationGender

Description: Indicates the gender categories per zone

Input to: S6

Properties

- Zone
- Year
  - The year the data refer to
- Gender

Entity: PopulationAge

Description: Indicates the mean age per zone

Input to: S6

Properties

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- Zone
- Year
  - The year the data refer to
- Age

Entity: PopulationEducation

Description: Indicates the number of people per education category per zone

Input to: S6

Properties

- Zone
- Year
  - The year the data refer to
- Education_type
  - Reference to EducationType
- Value

Entity: PersonalIncome

Description: Indicates the mean personal income per zone

Input to: S6

Properties

- Zone
- Year
  - The year the data refer to
- Income

WorkHoursPerWeek

Description: Indicates the mean hours of work per week per zone

Input to: S6

Properties

- Zone
- Year
  - The year the data refer to
- WeeklyHours

Entity: WeeklyStudingHours

Description: Indicates the mean hours of weekly studying in school per zone

Input to: S6

Properties

- Zone

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- Year
  - The year the data refer to
- WeeklyHours

Entity: MobilitySubscriptions

Description: Subscriptions to mobility services of a person or household

Input to: S6

Properties

- Zone
- Year
  - The year the data refer to
- Mode
  - Reference to Transport_Mode in the Auxiliary Data
- PersonalSubscriptions:
  - Number of personal subscriptions per type (e.g. PT pass) in a zone
- HouseholdSubscriptions
  - Number of household subscriptions per type (e.g. PT pass) in a zone

Entity: ParkingSpaces

Description: Parking spots available to a person or household

Input to: S6

Properties

- Zone
- Year
  - The year the data refer to
- parkingSpots
  - Number of parking spots at home or at work

Entity: Synthetic Household

Description: A household in a virtual population

Output of: S6

Properties

- Zone
- Year
  - The year the data refer to
- householdID
- Persons
  - The number of persons of the synthetic household
- MeanAge
  - The mean age of the synthetic household
- Vehicles
  - Mean number of vehicles per household

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Entity: Synthetic Person

Description: A person in a virtual population

Output of: S6

Properties

- Zone
- Year
  - The year the data refer to
- householdID
  - Household id
- PersonID
- Age
  - The age of the synthetic person
- Gender
  - The gender of the synthetic person
- Education
  - EducationType of the synthetic person
- Vehicles
  - Number of vehicles owned by the synthetic person

Entity: TransShipment_Location

Description: Locations of transshipment/intermodal facilities

Input to: S3

Properties

- ts_location_id
- Coordinates

Entity: Household_Travel_Demand_Survey

Description: Mode choices of households from travel surveys

Input to: S3

Properties

- Survey_id
- Household_id
- Number_of_vehicles
- Number_of_parking_spaces
- Number_of_trips
- Mode_Choice

Entity: Cost_per_vehicle_type

Description: Costs per vehicle ride.

Input to: S3

Properties

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- sm_record_id
  - Reference to Skim_matrix
- Vehicle_type_id
  - Reference to Vehicle_Type in Auxiliary Data
- Single_occupancy_cost
- shared_ride_cost
- Shared_ride_persons

Entity: Population_flows_by_type
Description: Indicates population flows from an origin to a destination. **Input to: S3**
Properties • Origin • Destination • Population_Type

2.4 Auxiliary Data Types

Entity: Transport_Mode
Description: Available transport modes
Properties • id, • Description   ○ Including Bus, taxi, car, car sharing, bike sharing, ride hailing, flying taxi, freight vehicles, Cargo Bikes, new mobility modes.

Entity: Emission_Factor
Description: Emission factors by vehicle-, road-, and location type

Properties
• vehicle_type
  ○ the vehicle types used for freight transport (freight vehicles, Cargo Bikes, etc.).
• road_type
  ○ a classification of road types for freight transport.
• Location_type
  ○ a classification of location types for emission calculation.
• Emmision_factor
  ○ Emission factors by vehicle-, road-, and location type

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Entity: OccupationType

Description: Indicates the occupation classification used, e.g. SOC2000[2]

Input to: Strategic/Long Term Models

Properties

- Id
- Type
  - The Occupation type. Example include:
    - Managers_senior_officials: number of people aged 16 to 74 employed as managers or senior_officials
    - Professional: number of people aged 16 to 74 employed as Professionals
    - Associate_professional_technical: number of people aged 16 to 74 employed as associate professionals and technical staff
    - Administrative_secretarial: number of people aged 16 to 74 employed with an administrative and secretarial occupation
    - Skilled_trades: number of people aged 16 to 74 employed with a skilled trades occupation
    - Personal_service: number of people aged 16 to 74 employed with a personal service occupation
    - Sales_customer_service: number of people aged 16 to 74 employed with an occupation on Sales and customer service
    - Process_plant_machine_operatives: number of people aged 16 to 74 employed with an occupation on process, plant and machine operatives
    - Other

Entity: Geography

Description: Indicates the geography (polygon) of the area of interest.

Input to: Strategic Models

Properties

- Polygon
  - A Spatial Data type, which do not have to be limited to classic geometric shapes like Squares, Hexagons, etc. but can also be geographic borders.
- Type
  - The type of geography the data refer to
  - Reference to Geography_Type

Entity: Geography_Type

Description: Indicates the geography type.

Input to: Strategic Models

Properties

- Id
- Description
  - The type of geography the data refer to. It may follow a classification such as the following:
    - 2001 super output areas - lower layer,

[2] https://www.nomisweb.co.uk/census/2001/KS012A/view/2013265925?cols=measures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- 2001 super output areas - middle layer,
- 2001 output areas,
- 2003 CAS wards,
- parishes 2001,
- parliamentary constituencies 2010,
- former metropolitan counties,
- local authorities: county / unitary (prior to April 2015),
- local authorities: district / unitary (prior to April 2015),
- english counties, regions,
- pre-2009 local authorities: county / unitary,
- parliamentary constituencies 1995 revision,
- pre-2009 local authorities: district / unitary, countries
- building

Entity: IndustryType

Description: Indicates the industry types used. A classification such as SIC92[3] may be used.

Input to: Strategic Models

Properties

- Id
- Industry_types
  - The type of industry the data refer to. It may follow a classification such as the following:
    - A_Agriculture_hunting_forestry: integer # agriculture, hunting and forestry industry
    - B_Fishing: integer # fishing industry
    - C_Mining_quarrying: integer # mining and quarrying industry
    - D_Manufacturing: integer # manufacturing industry
    - E_Electricity_gas_water_supply: integer # Electricity, gas and water supply industry
    - F _Construction: integer # Construction industry
    - G_Wholesale_retail_vehicles: integer # Wholesale and retail trade and repair of motor vehicles industry
    - H_Hotels_restaurants: integer # Hotels and restaurants industry
    - I_Transport_storage_communications: integer # Transport storage and communications industry
    - J_Financial_Intermediation: integer # Financial Intermediation industry
    - K_Real_estate: integer # Real estate,renting and business activities industry
    - L_Public_administration: integer # Public administration and defence and social security industry
    - M_Education: integer # Education industry
    - N_Health: integer # Health and social work industry
    - O_P_Q_Other: integer # other community, social and personal service activities, private households with employed persons and extra-territorial organisations and bodies

Entity: LandUseType

Description: Indicates the land use type e.g. industrial, commercial, office, residential

Input to: Strategic Models

Properties

- Id

[3] https://www.nomisweb.co.uk/census/2001/ST039/view/2013265925?rows=c_occpuk11_1&cols=c_indgpuk11

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- Land_use_type
  - The type of land use. It may follow a classification such as industrial, commercial, office, residential

Entity: NUTS2Zone

Description: A subclass of Zone

Input to: S2

Properties

- zone
  - Reference to a zone in the network model
- ZoneCode
  - NUTS2 zone code

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

3 Data processing and management pipeline

The Harmony TSDW will include methods and prototype tools that facilitate data integration from many heterogeneous sources in an automated and standardized way, while ensuring data quality and also provide the technical power to manage large data streams efficiently. HARMONY's data management infrastructure will integrate static data and near real-time data streams under a single scalable data warehouse based on big-data technology supporting wide access to data for metropolitan areas.

Big-data is a broad terminology for extremely large and complex data sets, which cannot be adequately handled by traditional data processing tools and mechanisms. Most common definitions for big-data are Gartner's 3Vs definition for Big Data, De Mauro's variant for the 3Vs and IBM's 4Vs definition:

- Big-data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making (Gartner, Inc., 2013).
- Big-data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value (De Mauro, Greco, & Grimaldi, 2014).
- Big-data provides the ability to achieve superior value from analytics on data at higher Volumes, Velocities, Varieties or Veracities (4Vs). With higher data volumes, a more holistic view of the subject's past, present and likely future can be taken. At higher data velocities, decisions can be grounded in continuously updated, real-time data. With broader varieties of data, a more nuanced view of the matter at hand can be provided. And as data veracity improves, analysts can be confident that they're working with the truest, cleanest, most consistent data (IBM, 2020)

In all definitions, **Volume** represents the ever-augmenting amount of data collected, **Velocity** corresponds to the exponential growth on data acquisition, and **Variety** stands for the growing heterogeneity of data formats and communication protocols that exist to share and spread data. **Veracity** refers to the biases, noise and abnormality in data.

With respect to transportation data all of the big-data characteristics are present, from large quantities of data, captured every day in intervals ranging from hours to seconds, varying from real-time or simulated traffic data, floating-car and GPS data, weather and traffic forecasting and history data among several others.

Big transportation data is also highly variable, since it still presents lots of inconsistencies, such as intermittent sensor data (traffic, parking spots, etc.) or outdated data from transportation providers (schedules, stops, etc.). These inconsistencies lead to a low veracity ratio, since the quality of data is always changing, even in the same provider. Finally, complexity refers to Volume-to-Variety ratio, and to the difficulties to fuse large amounts of data coming from several different sources.

Taking into account all of the peculiarities of big-data, its collection, treatment and harmonization for future use has to be so that the extrapolation of valuable information and knowledge from big-data surpasses, or at least minimizes the problems raised by big-data characteristics.

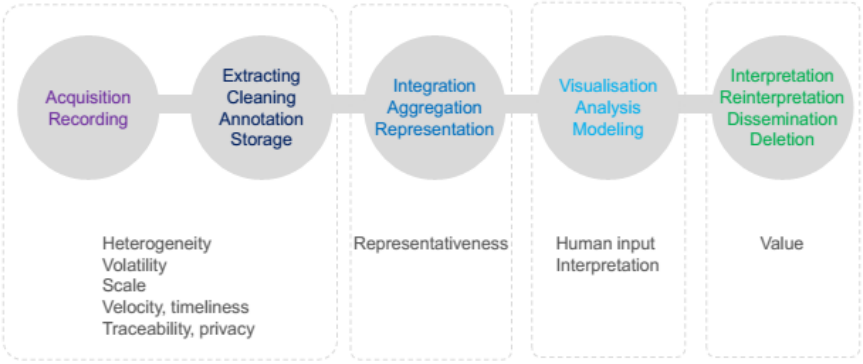

Figure 2: Big Data Lifecycle (OECD/ITF, 2015).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Formally, the lifecycle of big-data, from its collection to its usage as valuable information is comprised by several stages, as shown in Figure 2 (OECD/ITF, 2015).

Beyond data collection, acquisition and recording tasks, data has to be fit for analysis, meaning that, once the data is parsed into relevant fields a series of operations can be performed to clean, transform, map them to interoperable schemas and add meta-data in order to retrieve meaningful value from it. Data analytics represents all the ways in which information can be extracted from a given data set (OECD/ITF, 2015).

3.1 ETL Methodologies

ETL stands for Extract-Transform-Load, representing the process in which data is loaded from one or more source systems to a unified data repository. ETL is a concept introduced in the context of a Data Warehouse, i.e. a central repository for all or significant parts of the data that an enterprise's various business systems collect. However, it is now being adapted to all kinds of data-interacting areas, such as Business Intelligence & Analytics or Data Mining. The steps for a generic ETL process are described below:

- Extract: This step entails data extraction from source systems, making it accessible for further processing. The idea is to extract all the required data from source systems with as little resources as possible.
- Clean: ETL encompasses a cleaning step as a separate step. The cleaning process is one of the most important steps, as it ensures the quality of data, by making identifiers unique, converting null values into a standardized Not Available value or converting numbers such as phone numbers and ZIP codes to a standardized form.
- Transform: The transform step applies a set of rules to transform the data from the source to the target. This includes converting any measured data to the same dimension (i.e. conformed dimension) using the same units so that they can later be joined.
- Load: During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible. The target of the Load process is often a database.

ETL is evolving to support integration across much more than traditional data warehouses. ETL has become the next component of analytic architecture poised for major evolution. Much new data is semi-structured or even non-structured, and constantly evolving data models are making the accepted tools for structured data processing almost useless. ETL can support integration across transactional systems, operational data stores, BI (Business Intelligence) platforms, MDM (Multi-Dimensional Data Modelling) hubs, the cloud, and other big-data platforms, such as Hadoop. ETL software vendors are extending their solutions to provide big-data extraction, transformation, and loading between big-ata platforms and traditional data management platforms. Hence, like the conceptual tipping point that brought to life the term “big-data,” the same scale of evolution with ETL has been reached. The term “Big ETL” describes the new era of ETL processing, defining it as having the majority of the following properties (Caserta & Cordo, 2015):

- The need to process “really big-data” – the data volume is measured in multiple Terabytes or greater.
- The data includes semi-structured or unstructured types – JSON, Avro, etc.
- Interaction with non-traditional data storage platforms – NoSQL, Hadoop, and other distributed file systems (S3, Gluster, etc).

There are several methodologies for managing the lifecycle of ETL, from the ones most widely spread, to the ones created and used within a single company or entity, going through new findings and novel methodologies. Some of the better known, and the ones highlighted here, are the Knowledge Discovery Databases (KDD), the CRoss Industry Standard Process for Data Mining (CRISP-DM), and the Sample, Explore, Modify, Model, Assess (SEMMA) methodologies or models. In the following we analyse the aforementioned open and widely used methodologies, which have been verified in practise, with the aim to select the one that will be adopted within HARMONY.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

1.1.1 KDD

The KDD model is an iterative and interactive model (Piatetsky-Shapiro et al., 1996). It refers to finding knowledge in data and emphasizes the high level of specific data mining method. It consists on the extraction of hidden knowledge according to databases. KDD requires relevant prior knowledge and brief understanding of application domain and goals. There are nine different steps in the KDD methodology (Shafique & Qaiser, 2014):

- Domain Understanding: KDD goals are defined from the customer's point of view and used to develop an understanding about the application domain and its prior knowledge.
- Data Selection: The data is partitioned in subsets of data samples in order to ease the mining processes, both in terms of complexity and performance. This is an important stage because knowledge discovery is performed on all these data subsets.
- Pre-processing: Data cleaning and pre-processing processes to enable completeness and consistency, producing data without any noise and inconsistencies. In this stage strategies are developed for handling noisy and inconsistent data.
- Transformation: This step focuses on transformation of data from one form to another so that data mining algorithms can be easily implemented. For this purpose, different data reduction and transformation methods are implemented on target data.
- Data Mining: The Data Mining step actually encapsulates three distinct steps: Choosing a suitable data mining task, choosing a suitable data mining algorithm and employing the data mining algorithm.
- Interpretation/Evaluation: Interpretation and evaluation of resulting mining patterns. This step may involve extracted patterns' visualization.
- Knowledge Usage: This is the last and final step of KDD process in which the discovered knowledge is used for different purposes. The discovered knowledge can also be used by interested parties or can be integrated with other systems for further usage.

1.1.2 CRISP-DM

CRISP-DM (Azevedo, et al., 2008) was launched in late 1996 by Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR. This model has been refined over the years. CRISP-DM 1.0 version was published in 1999 and is complete and documented. It provides a uniform framework and guidelines for data miners. It consists of six phases or stages which are well structured and defined (Shearer, 2000):

Figure 3. CRISP-DM Process Diagram

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- Business Understanding: This step uncovers important factors including success criteria, business and data mining objectives and requirements as well as business terminologies and technical terms.
- Data Understanding: This step focuses on data collection, checking quality and exploring of data to get insight of data to form hypotheses for hidden information.
- Data Preparation: selection and preparation of final the data set. This phase may include many tasks, such as records, tables and attributes selection as well as cleaning and transformation of data.
- Modelling: Selection and application of various modelling techniques. Different parameters are set and different models are built for same data mining problem.
- Evaluation: Evaluation of obtained models and decision on how to use the results. Interpretation of the model depends upon the algorithm and models can be evaluated to review whether it achieves the objectives properly or not.
- Deployment: Determining possible uses for obtained knowledge and results. This phase also focuses on organizing, reporting and presenting the discovered knowledge when needed.

1.1.3 SEMMA

SEMMA (SAS Institute, 2014) model was developed by SAS institute. It has five different phases. It offers and allows understanding, organization, development and maintenance of data mining projects. It helps in providing the solutions for business problems and goals. SEMMA is linked to SAS enterprise miner and basically a logical organization of the functional tools for them.

- Sample: Sampling of data. A portion from a large data set is taken that big enough to extract significant information and small enough to manipulate quickly.
- Explore: Exploration of data. This can help in gaining the understanding and ideas as well as refining the discovery process by searching for trends and anomalies.
- Modify: Modification of data by creating, selecting and transformation of variables to focus model selection process. This stage may also look for outliers and reducing the number of variables.
- Model: Modelling of data. The software for this automatically searches for combination of data. There are different modelling techniques are present and each type of model has its own strength and is appropriate for specific situation on the data for data mining.
- Access: Evaluation of the reliability and usefulness of findings and estimates the performance.

1.1.4 Comparative Analysis of KDD, CRISP-DM and SEMMA

By examining all three data mining process models, it is clear that KDD and SEMMA are almost identical in that every stage of KDD directly corresponds to a stage of SEMMA. SEMMA is directly linked to the SAS enterpriser miner software and CRISP-DM Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR (Shafique & Qaiser, 2014). The CRISP-DM process combines Selection-Preprocessing (KDD) or Sample-Explore (SEMMA) stages into Data Understanding stage. It also incorporates Business Understanding and Deployment stages. An important difference between CRISP-DM and the two other methodologies is that transitions between stages in CRISP-DM can be reversed. This helps a lot when working with real data — any misstep can be fixed without having to finish the whole cycle if someone understands, that chosen target data will not lead to any knowledge. CRISP-DM remains the top methodology for data mining projects, with essentially the same percentage as in 2007 (43% versus 42%) (Piatetsky, 2014). In HARMONY we will be following CRISP-DM methodology as a general structured approach for data management with any modifications that may be needed to accommodate specific needs of the project. At this point we have already identified domain specific data processing objectives and requirements (Business Understanding) and we are in the phase of Data Understanding focusing on data collection, quality checking and exploration.

3.2 Big-data ETL Tools

Unlike traditional ETL platforms that are largely proprietary commercial products, the majority of Big ETL platforms are powered by open source. These include Hadoop (MapReduce), Spark, Flink and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Storm. The fact that Big ETL is largely powered by open source is interesting for several reasons: First, open-source projects are driven by developers from a large number of diverse organizations. Second, one of the most important features of ETL platforms is the ability to connect to a range of data platforms. Instead of waiting for a vendor to develop a new component, new integrations are developed by the community. Third, and perhaps most important, the fact that these engines are open source (free) removes barriers to innovation. Organizations that have a great use case for processing big-data are no longer constrained by expensive proprietary enterprise solutions.

ETL and data integration software is primarily meant to perform the extraction, transformation and loading of data. Once the data is available for example in a data warehouse or OLAP cube, Business Intelligence software is commonly used to analyse and visualize the data. This type of software also provides reporting, data discovery, data mining and dashboarding functionality. Some of the most successful open-source integrated environments with ETL and BI capabilities are listed below:

- Talend Open Studio: Talend Open Studio[4] is a versatile set of open source products for developing, testing, deploying and administrating data management and application integration projects. For ETL projects, Talend Open Studio for Data Integration delivers a rich feature set including a graphical integrated development environment with an intuitive Eclipse-based interface. The advanced ETL functionality including string manipulations, automatic lookup handling, and management of slowly changing dimensions and support for ELT (extract, load, and transform) as well as ETL, even within a single job.
- RapidMiner: RapidMiner[5] is one of the leading data mining software suites. RapidMiner supports all steps of the data mining process from data loading, pre-processing, visualization, interactive data mining process design and inspection, automated modelling, automated parameter and process optimization, automated feature construction and feature selection, evaluation, and deployment. RapidMiner can be used as stand-alone program on the desktop with its graphical user interface (GUI), on a server via its command line version.
- GeoKettle ETL: GeoKettle[6] is a powerful, metadata-driven spatial ETL tool dedicated to the integration of different data sources for building and updating geospatial databases, data warehouses and services. GeoKettle enables the Extraction of data from data sources, the Transformation of data in order to correct errors, make some data cleansing, change the data structure, make them compliant to defined standards, and the Loading of transformed data into a target DataBase Management System (DBMS) in OLTP or OLAP/SOLAP mode, GIS file or Geospatial Web Service.
- Apache NiFi[7]: supports powerful and scalable directed graphs of data routing, transformation, and system mediation logic. Apache NiFi provides a Web-based user interface for seamless experience between design, control, feedback, and monitoring, it is highly and provides dataflow tracking from beginning to end while offering secure interfaces and multi-tenant authorization and internal authorization/policy management

3.3 HARMONY Approach

The HARMONY approach to data processing and management will rely on the CRISP-DM model as described above. At this point the project is in the phase of defining the data processing and data transformation requirements with a focus on answering questions related to how data provided by cities need to be managed and provided as input to the HARMONY simulators. As soon as the requirements are set, the most appropriate big-data ETL tool such as the ones mentioned above will be selected as the basis for developing the TSDW ETL procedures.

[4] https://www.talend.com/products/talend-open-studio/
[5] https://rapidminer.com/
[6] https://live.osgeo.org/archive/10.0/en/overview/geokettle_overview.html
[7] https://nifi.apache.org/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

4 TSWD Infrastructure & Implementation Approach

4.1 Database Choice

At the heart of the persistency layer of the HARMONY TSDW there are database systems for holding the required transport and spatial data. From a technical viewpoint, the selection of the database systems to be used need to support the following:

- Relational data storage and access.
- Geospatial indexing and querying capabilities.
- Flexible data schemas with noSQL mainly for high volume model execution logs (in JSON format) which are used for simulation debugging. These data mainly refer to streaming data from models and are provided by the messaging middleware of HARMONY's technical architecture.
- BLOBs storage for e.g. shapefiles.

4.1.1 Database models

A database system implements a database model that is used to logically structure the data that is being managed. These models determine how a database application will work and handle the information it deals with. There are several types of database models which provide the means of structuring the data, with most popular being the Relational Model. The relational model has been extensively used and is extremely powerful and flexible. However, there have been certain issues or features that these solutions required, leading the developers of database systems to explore a series of different systems and applications called NoSQL databases. Such stems are gaining popularity, with their promise offering additional functionality. Essentially, such NoSQL databases eradicate the strictly structured data schemas and data relations. They work by providing a freely shaped way of working with information, which offers greater flexibility and ease (see Section 4.1.1.3 below).

4.1.1.1 The Relational Approach

The relational model was firstly Introduced in the 1970s and offers a way of structuring, keeping, and using data, relying on formal approaches for data schemas definition and redundancy avoidance while ensuring consistency, such as Entity-Relation diagrams. Relations have the advantage of group-keeping the data as constrained collections in data-tables that contain information in a structured way (e.g. a Person's name and address) and relate the data by assigning values to attributes (e.g. a Person's ID number). Related relational database management systems require defined and clearly set schemas which shape how the data is contained and used.

The relational model has been widely used and over the last decades a number of database systems that implement it have grown to work extremely efficiently and reliably. Furthermore, computer engineers and developers are commonly accustomed to this model and have the experience in working with relational databases. This is a main reason that this model has become the main choice when software applications are developed.

Some popular relational database management systems are:

- SQLite: An embedded relational database management system.
- MySQL: The most popular and commonly used RDBMS.
- PostgreSQL: The most advanced, SQL-compliant and open-source objective-RDBMS.

4.1.1.2 The Model-less (NoSQL) Approach

The NoSQL approach of structuring the data consists without formal constraints and strict relations, which essentially liberates the means of keeping, querying, and using information. NoSQL databases offer many different types of ways to keep and work with the data for specific use cases efficiently based on e.g. document or key-value stores. Unlike traditional relational databases, it is possible to group

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

collections of data together with certain NoSQL databases, such as the MongoDB. In any case NoSQL databases do not have a common way to query the data (which is the case relational databases that use the SQL standard) and each solution provides its own query system.

Some popular document based NoSQL database management systems are (note that based on the current understanding of the HARMONY data, document based databases are the most relevant for storing the available semi-structured data (e.g. CSV data) and logs generated by the messaging middleware):

- MongoDB, a cross-platform document-oriented database
- Apache CouchDB, an open-source document-oriented NoSQL database, implemented in Erlang

4.1.1.3 Characteristics of SQL and No-SQL Database Management Systems

The main differences between SQL and No-SQL databases are:

Structure and type of data being kept: SQL/Relational databases require a structure with defined attributes to hold the data, NoSQL databases commonly allow free-flow operations.

Querying: Relational databases implement the SQL standard to a certain degree and can be queried using the Structured Query Language (SQL). NoSQL databases use proprietary query protocols for accessing and managing the data they store.

Scaling: Both solutions are easy to scale vertically (i.e. by increasing system resources). However, NoSQL solutions usually offer much easier means to scale horizontally (i.e. by creating a cluster of multiple machines).

Reliability: SQL databases provide higher data reliability and safe guarantee of performed transactions (through atomicity, consistency, isolation, durability of transactions (ACID)).

Support: Relational database management systems have existed for a longer period and are extremely popular resulting to the fact that it easier to find both free and paid support. If an issue arises, it is therefore much easier to solve than recently-popular NoSQL databases – especially if said solution is complex in nature (e.g. MongoDB).

Complex data keeping and querying needs: Relational databases are the go-to solution for complex querying and data keeping needs. They are much more efficient and excel in this domain.

4.1.2 The HARMONY Approach

Taking into account the abovementioned requirements a hybrid approach combining SQL and noSQL based database systems is foreseen for the HARMONY TSDW (see Figure 4). An SQL database for relational data storage and access as well as geospatial indexing and querying capabilities will be used (this will mainly consist the “project area”) along with a noSQL database for logging and initial data importing for transformation (mainly the staging area).

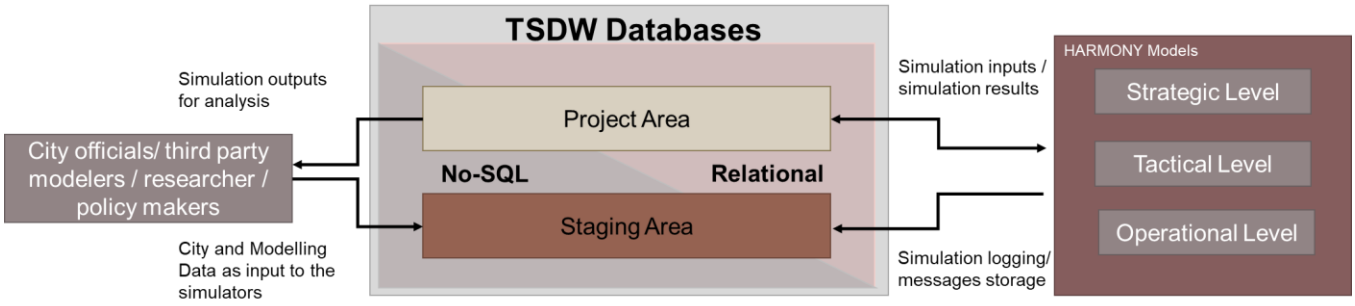

Figure 4: The TSDW Project and Staging areas.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

Based on our analysis, experience and the above requirements the following database systems will be used:

- PostgreSQL[8] with PostGIS[9] extension
- MongoDB[10]

The use of PostreSQL and PostGIS is in line with state of art approaches as it is used in implementations of transport simulators, for example, the Simmobility[11] suite of simulators which to some extent follows a paradigm related to HARMONY's model suite, relies on a relational database to manage data and specifically makes use of PostreSQL. The use of MongoDB will allow to utilize more flexible noSQL data schemas where possible.

4.2 Interfaces and Communication Protocols

The approach for developing the HARMONY TSDW is to rely on open standards to the largest possible extent and this applies to the data import and access interfaces as well as communication protocols. Data interoperability will be ensured to the largest possible extent through standards such as GML[12], GeoJSON[13] and Shapefiles[14] for geographical information, GTFS[15] for public transportation schedules and DATEX II[16] for traffic information. The related approaches vary widely and depend on the implementation and available client-side implementations and related client software packages for implementing the connections to the databases. In any case, it is important to limit technology approaches to avoid complicated implementations which could cause operational, maintenance and long-term problems, while ensuring reusability of the project outcomes. In the following a description of the main approaches that the TSDW will instantiate are described.

4.2.1 Open database connectivity

Open database connectivity (ODBC) will be used for direct access to the underlying database systems. Such connectivity will be carefully provided as it is essentially a scheme that allows an application to have direct access to the schema and tables of the database. A set of rules need to be implemented in this case granting filtered and / or read only access to the data.

4.2.2 RESTful API connectivity

This type of connectivity refers to the implementation of web services over the HTTP protocol, that perform a set of operations on data resources (create update - delete (CRUD)). Each data resource is identified by a URL and is available by calling the specific URI with a request that contains the operation to be performed. In general, a RESTful request contains the URI, request parameters and type of operation which can generally be one of the following GET, POST, PUT, DELETE, and defines the actions that need to performed by service on the server. In more details:

- A GET request asks the server to return information regarding the resource identified by the request parameters. Commonly the parameters of a GET are part of the request URI in the form of a query string with a collection of key/value pairs.
- A POST request asks the server to create a new data resource. The information regarding the resource to be created is commonly located in request body, for example in the form of a json document.

[8] http://www.postgresql.org/
[9] http://www.postgis.net/
[10] https://www.mongodb.com/
[11] https://its.mit.edu/software/simmobility
[12] https://en.wikipedia.org/wiki/Geography_Markup_Language
[13] https://en.wikipedia.org/wiki/GeoJSON
[14] https://en.wikipedia.org/wiki/Shapefile
[15] https://en.wikipedia.org/wiki/General_Transit_Feed_Specification
[16] https://www.datex2.eu/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- A DELETE request asks the server to delete a data resource as described in the query parameters. Similarly to the case of GET, the parameters of a DELETE are part of the request URI in the form of a query string with a collection of key/value pairs.
- A PUT request asks the service to update a data resource. The information to be updated is located in the request body like for example a JSON document.

The RESTful approach is completely stateless which means that the contents of variables between requests are not stored. A predefined specification describing the URIs and parameters contained in the requests between the client and server provides the common understanding of the content and context being exchanged. The service producer and service consumer have a mutual understanding of the context and content being passed along. Since there is no formal way to describe the web services. RESTFul implementations are open in the sense that there is great flexibility in the definition of the schemas that describe the data being exchanges and related processing methods. Furthermore, such an approach is aligned with the event-based architecture and messaging middleware that is proposed in the HARMONY technical architecture. With the implementation of software wrappers the data requests the different simulators can be handled with the RESTful API of the TSDW.

4.3 Approach for deployment

The HARMONY TSDW will be containerized in order to support different deployment needs and approaches. This means that there will options to provide the TSDW as a virtual machine or a docker[17] image(s). This approach allows for great adaptability to the needs for the users of the HARMONY model suite. In cases where the model suite deployment is performed within the premises of the end users because for example the data to be imported cannot be shared or uploaded to public servers, a local installation of the virtual machine or docker image will be possible. Similarly, it will be possible to deploy the TSDW in public clouds such as Amazon or Microsoft azure.

4.4 Non-functional requirements considerations

4.4.1 Configuration (compute – memory – storage requirements)

The TSDW is being designed such that it can scale to HARMONY's transport simulation needs. At this point we identify a set of configuration guidelines which will be revisited and further refined as the development of the HARMONY simulators as well as the integration work progresses. In terms of storage we foresee a need for at least one TB which should rely on SSD hard drives for optimal performance (for the PostgreSQL and MongoDB instances). With respect to network connectivity, the server hosting the TSDW should be connected by at least a gigabit connection with less than one millisecond of latency. The processing should be performed in servers with enough capacity. Based on prior experience machines with at least 8 cores and 32 GB of memory should be used.

4.4.2 Data security considerations

To ensure a unified and manageable data security and authorisation capability across all the components of HARMONY, the TSDW will make use of the OAuth protocol for RESTful interfaces and will make use of secure data connections in cases of direct access. OAuth is a widely used and efficient simple method to interact with protected resources. It provides access to requested resources without credentials sharing and relies on access tokens that have limited lifetime. A role-based access control approach will be used for assigning different access rights to different actors. Data access levels will be defined to provide proper content to different users and components of the platform. At this point we foresee the following access levels:

- Admin: has access to all resources available in the TSDW. This role will be assigned to the TSDW administrator and will be used for development and implementation purposes.
- Simulation component: provides access to all data required by a HARMONY component.

[17] https://www.docker.com/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

- Transport modeller: has access to simulation input data relevant for her/his modelling purposes as well as the corresponding simulation outcomes. This user analyses the related data and infers conclusions.

Note that we foresee to define the types of operations (read/write/edit/update) permitted per user in a specific access level and per dataset. This will allow to implement concepts of data ownership so that representatives of an area can upload datasets relevant for that area and provide access only to other members / simulation components that require these datasets and corresponding simulation results. Moreover, extensions to these access levels will be made as new requirements for data access arise during the implementation of the HARMONY components.

4.4.3 Data privacy and anonymization

The TSDW will not hold any personally identifiable data which require e.g. anonymization. HARMONY collects personal data from travel surveys as part of task T3.1. The data collection is performed using the MobyX app which adheres to all privacy and anonymization related regulations. The TSDW will be populated with aggregate data only and knowledge which will be the outcome of analysis of data coming from travel surveys. In this respect, no special considerations for privacy and anonymization are required for the TSDW.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

5 Conclusions and Next Steps

This deliverable provided the technical design of the HARMONY TSDW addressing requirements of the components of the HARMONY model suite. In this respect, the following main design elements of the TSDW have been considered in this deliverable:

- The specification of the required data schemas for storing transport and spatial data. Starting from the conceptual architecture of the HARMONY Model Suite (MS) (deliverable D1.3) a thorough record of data produced, consumed and exchanged between the simulators has been compiled. This set of data has been analysed in order to derive a harmonized data model where the main data entities of the model suite have been defined together with their properties.

- A data processing and management pipeline including the selection of data storage and processing technologies has been identified. In terms of technologies we aim to rely on PostgreSQL and MongoDB.

- Open Database Connectivity and RESTful APIs have been identified as the main interfaces and communication protocols interacting with external data providers and consumers.

- Processing and storage capacity. Based on the identified data structures and the expected volume and data processing functions, preliminary figures of the required processing and storage capacity have been identified in Section 4.4.1.

The technical design is already feeding with information the other tasks of WP3 which concern the implementation and population of the TSDW. In the next period the databases of the TSDW will start being populated with data, and integrations with the different simulators will be performed.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269

6 References

Azevedo, Ana Isabel Rojão Lourenço, and Manuel Filipe Santos. "KDD, SEMMA and CRISP-DM: a parallel overview." IADS-DM (2008).

Caserta, J., & Cordo, E. (2015). Big ETL: The Next 'Big'Thing. Retrieved from DataInformed: http://data-informed. com/big-etl-next-big-thing.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. Library Review.

Gartner Research, 2013, Hype Cycle for Big Data, available at: https://www.gartner.com/en/documents/2574616/hype-cycle-for-big-data-2013

IBM (2020), The Four V's of Big Data, available at http://www.ibmbigdatahub.com/infographic/four-vs-big-data

OECD/ITF 2015, Big Data and Transport: Understanding and Assessing Options, available at: https://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf

Piatetsky-Shapiro, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. KDnuggetsTM Data Mining, Analytics, Big Data and Data Science.

Piatetsky-Shapiro, G., Brachman, R. J., Khabaza, T., Kloesgen, W., & Simoudis, E. (1996). An overview of issues in developing industrial data mining and knowledge discovery applications. In KDD (Vol. 96, pp. 89-95).

SAS Institute, SEMMA model, available at: https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en

Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1), 217-222.

Shearer, C. (2000) The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5, 13-22.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815269