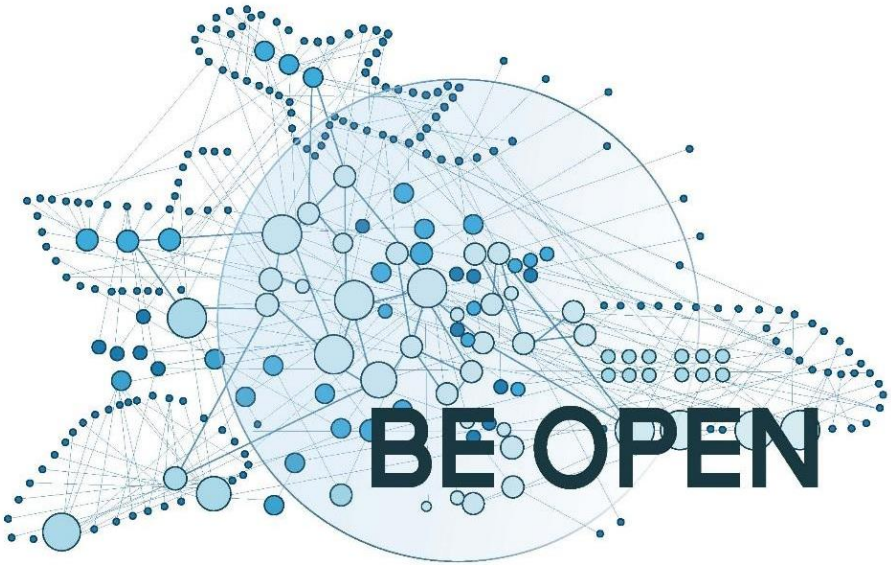
Pending EC/INEA approval

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824323



European forum and oBsErvatory for OPEN science in transport

Project Acronym:	**BE OPEN**
Project Title:	**European forum and oBsErvatory for OPEN science in transport**
Project Number:	**824323**
Topic:	**MG-4-2-2018 – Building Open Science platforms in transport research**
Type of Action:	**Coordination and support action (CSA)**

D4.2 Transport Open Data: Properties and Specifications for Open Science

Final Version

Deliverable Title:	Transport Open Data: Properties and Specifications for Open Science
Work Package:	WP4
Due Date:	31/10/2020
Submission Date:	19/01/2021
Start Date of Project:	01/01/2019
Duration of Project:	30 months
Organisation Responsible of Deliverable:	ATHENA RC (ARC)
Version:	Version 2.0
Status:	Final
Author name(s):	Alessia Bardi (OpenAIRE), Harry Dimitropoulos (ARC), Yannis Foufoulas (ARC), Afroditi Anagnostopoulou (CERTH)
Reviewer(s):	Katerina Folla (NTUA), Jakob Michelmann (VDI/VDE)
Nature:	☒ R – Report ☐ P – Prototype ☐ D – Demonstrator ☐ O - Other
Dissemination level:	☒PU - Public ☐ CO - Confidential, only for members of the consortium (including the Commission) ☐ RE - Restricted to a group specified by the consortium (including the Commission Services)

Document history			
Version	**Date**	**Modified by (author/partner)**	**Comments**
0.1	09/09/2020	Harry Dimitropoulos (ARC)	First draft version
0.2	30/10/2020	Harry Dimitropoulos (ARC), Alessia Bardi (OpenAIRE)	Chapter 1, Chapter/Section 2.1-2.3 & 2.5, Chapter 3
0.3	25/11/2020	Harry Dimitropoulos (ARC), Alessia Bardi (OpenAIRE)	Chapter/Section 2.4, Chapter 3
0.4	09/12/2020	Harry Dimitropoulos (Arc), Yannis Foufoulas (ARC)	General updates and edits, adding Section 3.2.2 on Anonymisation
0.5	15/12/2020	Alessia Bardi (OpenAIRE)	Adding section 3.2.1 for Argos; chapters re-numbered; revision of Introduction; draft for conclusion and discussion
0.6	19/12/2020	Harry Dimitropoulos (ARC)	Overall edits and conclusions
1.0	21/12/2020	Harry Dimitropoulos (ARC)	For internal review
1.1	23/12/2020	Harry Dimitropoulos (ARC)	Ordered and fixed References
1.2	14/01/2021	Harry Dimitropoulos (ARC), Alessia Bardi (OpenAIRE)	After internal review version
2.0	19/01/2021	Harry Dimitropoulos (ARC), Alessia Bardi (OpenAIRE)	Final version

Contents

List of Figures

List of Tables

Abbreviations and Terminology

Abbreviations	Terminology
API	Application Programming Interface
BDE	Big Data Europe (project)
BDTL	Big Data in Transport Library
C4R	CAPACITY4RAIL (project)
CC	Creative Commons
CSV	Comma-Separated Values
DMP	Data Management Plan
EC	European Commission
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, Reusable
FOT	Field Operational Test
FPS	Frames Per Second
HDFS	Hadoop Distributed File System
HLEG	High-Level Expert Group
IPR	Intellectual Property Rights
JSON	JavaScript Object Notation
LD	Linked Data
LOD	Linked Open Data
NDS	Naturalistic Driving Studies
OBD	Open Big Data
OD	Origin-Destination
OF	Open Format
OL	Open License
RDA	Research Data Alliance
RE	machine REadable
ReST	Representational State Transfer (API)
Rec	Recommendation
RDM	Research Data Management
RPC	Remote Procedure Call
SQL	Structured Query Language
TOPOS	Transport Observatory/fOrum for Promoting Open Science
TfL	Transportation for London
TT	Transforming Transport (project)
TSV	Tab-Separated Values
URI	Uniform Resource Identifier
WMS	Web Map Service
XML	eXtensible Markup Language

Executive summary

Deliverable D4.2 is the second deliverable of WP4 "Code of Conduct on Open Science in Transport". Following D4.1, which presents the main legal issues and fundamental principles of utilizing Open Science in transport research, this deliverable's (D4.2) objectives are to: i) identify available open big data in the transport research sector, and ii) provide recommendations for making them FAIR, in order to feed the TOPOS Observatory and Forum development in WP3. TOPOS aims to promote communication and networking among stakeholders in the transport research area.

Open Big Data in the transport research sector are not easy to find. This report confirms the remarks of the HLEG report and the NOESIS project regarding the low uptake of FAIR and Open principles by the transport research community, especially when big datasets are involved. Transport research is inherently cross-disciplinary and so provides an ideal context in which to apply principles such as FAIR but unfortunately, most of the data that transport researchers collect is used once and then stored away in locations that are inaccessible to other researchers.

The OpenAIRE Gateway on Transport Research[1] (which is part of the BE OPEN TOPOS) identifies, as of December 2020, only about 800 research data correctly deposited and made available via a trusted repository in the OpenAIRE network[2]. None of those could be identified as "big" according to the famous 4 Vs. In fact, for the identification of the open big datasets analysed in Chapter 2, the most effective methodologies have been going through the deliverables of projects in the domain and through word-of-mouth. This way a number of big data sets were found but unfortunately most of the transport research data, although big, are rarely open. The open big datasets that we discovered were less than twenty and are presented in Chapter 2, followed by an analysis of their general characteristics and properties, including an overview of big data file formats and their licenses.

The majority of the identified open big datasets belong to the categories of operational and governmental data, rather than original research data obtained by observations or data used in published transport research appearing in scientific venues. Only two open big datasets from Chapter 2 can be considered research data: *pNEUMA* and *Cityscapes 3D*. The two datasets went through a FAIR assessment process that revealed a low level of FAIRness in both cases. The assessment was conducted considering the FAIR principles and the set of indicators defined by the RDA FAIR Data Maturity Model Working Group. It is worth highlighting that very simple actions like depositing bibliographic metadata about the dataset in an open institutional or thematic repository and assigning a persistent identifier could already, with minor effort, have a massive impact on the level of FAIRness, especially for its findability and discoverability thanks to its inclusion in the open scholarly communication infrastructure. The FAIRness assessment is detailed in Chapter 3, preceded by a set of 14 recommendations for the research community in transport research, obtained as a synthesis of the RDA FAIR Data Maturity Model indicators and the recommendations of the EC HLEG report. The 14 recommendations suggest actions not only to single researchers but also to the research community as a whole. In fact, achieving FAIRness is a process that one person cannot conduct alone: the research community, as a whole, should agree on best practices at the local, national, and international level. In this context, the role of the TOPOS that BE OPEN is setting up is

[1] https://beopen.openaire.eu

[2] Mostly from generic research data repositories: Figshare and Zenodo.

fundamental because it could be the hub where different stakeholders of the domain gather, communicate with each other, and have access to guidelines and best practices that have been previously discussed and agreed in the TOPOS as well. Some of those guidelines and best practices, however, need not be defined from scratch, since several aspects of data FAIRness and data management are generic and can be addressed by building on top of the results obtained by other more mature research communities.

Chapter 4 focuses on two major challenges of open big data management in transport research: planning data management and anonymisation of sensitive data. In fact, the two challenges are generic: they also apply to other domains, as well as to "small" research data, and can therefore be tackled with already existing tools, that can be possibly customised for given peculiarities, if needed. As examples, we briefly described two OpenAIRE products that address the two challenges: *Argos* for the definition of machine-actionable Data Management Plans and *Amnesia* for the anonymisation.

1. Introduction

The objectives of the BE OPEN project are a) to create a common understanding on the practical impact of Open Science and b) to identify and put in place the mechanisms to make it a reality in transport research. As such, BE OPEN follows a two-fold action plan:

1. to engage key transport and open science-related communities in a participatory approach fostering a dialogue on Open Science (what exists, what should be done, how it should be done) among relevant stakeholders in Europe and around the world, and
2. develop a detailed roadmap for the implementation of sustainable open science modules which include key practices, infrastructures, policies and business models, all taking into account the specificities of the transport research domain, and the use and integration of existing-infrastructures and the emerging EOSC initiative.

Deliverable D4.2 is the second deliverable of WP4 “Code of Conduct on Open Science in Transport”. Following D4.1, which presents the main legal issues and fundamental principles of utilizing Open Science in transport research, this deliverable’s (D4.2) objectives are to i) identify available open big data in the transport research sector, and ii) provide recommendations for making them FAIR, in order to feed the TOPOS Observatory and Forum development in WP3. TOPOS aims to promote communication and networking among stakeholders in the transport research area.

The deliverable is organised as follows. Chapter 2 identifies and analyses open big datasets in the transport sector. Chapter 3 presents the FAIR principles and a set of recommendations to the transport research community for improving the FAIRness of their research data. FAIRness principles and indicators are evaluated on two selected open big research data among those identified in chapter 2: pNEUMA and the Cityscapes 3D dataset. Chapter 4 presents two useful tools for Research Data Management practices: Argos for the definition of data management plans and Amnesia for the anonymisation of sensitive information in research data. Chapter 5 draws conclusions.

2. Open Big Data in Transport Research

In this chapter, we describe our work on attempting to identify available Open Big Data (OBD) in the transport research sector. We first start with defining the Open Science terminology used in this report, followed by outlining some of the main challenges in using ODB in transport. We then describe the various OBD sources we found, including relevant research projects, organisations and institutions that collect real-time transport data. Some of the OBD that we discovered are described, followed by an analysis of their characteristics and properties, including big data file formats and their licenses.

2.1. Terminology

As deliverable D1.2 [1] pointed out, a global definition of Open Science terminology is key to ensure a common understanding between all stakeholders. Borrowing from section 3.1 of D1.2 [2], in order to share quality data and to create a common understanding in terminology within the "data sharing of transport research data", the following definitions are proposed:

Data: "any piece of information whose value might be used during analysis and impact its result." This means that participants, characteristics, weather, traffic and driving conditions could be considered as data and part of a dataset.

Open Data: are online, free of cost, accessible data that can be used, reused and distributed provided that the data source is attributed.

Big Data: Big Data refers to data sets that are too large, or too complex, or too fast-moving or too weakly structured to be evaluated by manual and conventional methods of data processing. In 2001, Doug Laney [3] introduced the famous three dimensions of Big Data, the "3 Vs": Volume, Variety, and Velocity. As the years went by and the definition of Big Data evolved, new "Vs" were gradually introduced: 4, 5[3], 7[4], 10 Vs[5] or more. A nice visual summary of the "4 Vs" (even if now dated) can be downloaded in the form of an infographic from IBM's Big Data and Analytics Hub[6] (shown in Figure 1).

Open Big Data (OBD): Big Data that are also Open Data.

Metadata: any piece of information necessary to use or properly interpret data. It could be divided into the following categories [2]:

- *Descriptive Metadata*: describes precisely each component of the dataset, including information about its origin and quality.
- *Structural Metadata:* describes how the data is organized.
- *Administrative Metadata*: sets how the data can be accessed and implemented.

Dataset: one definition for a dataset is "any piece of information necessary to use or properly interpret data".

[3] https://www.bbva.com/en/five-vs-big-data/
[4] https://impact.com/marketing-intelligence/7-vs-big-data/
[5] https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx
[6] https://www.ibmbigdatahub.com/infographic/four-vs-big-data

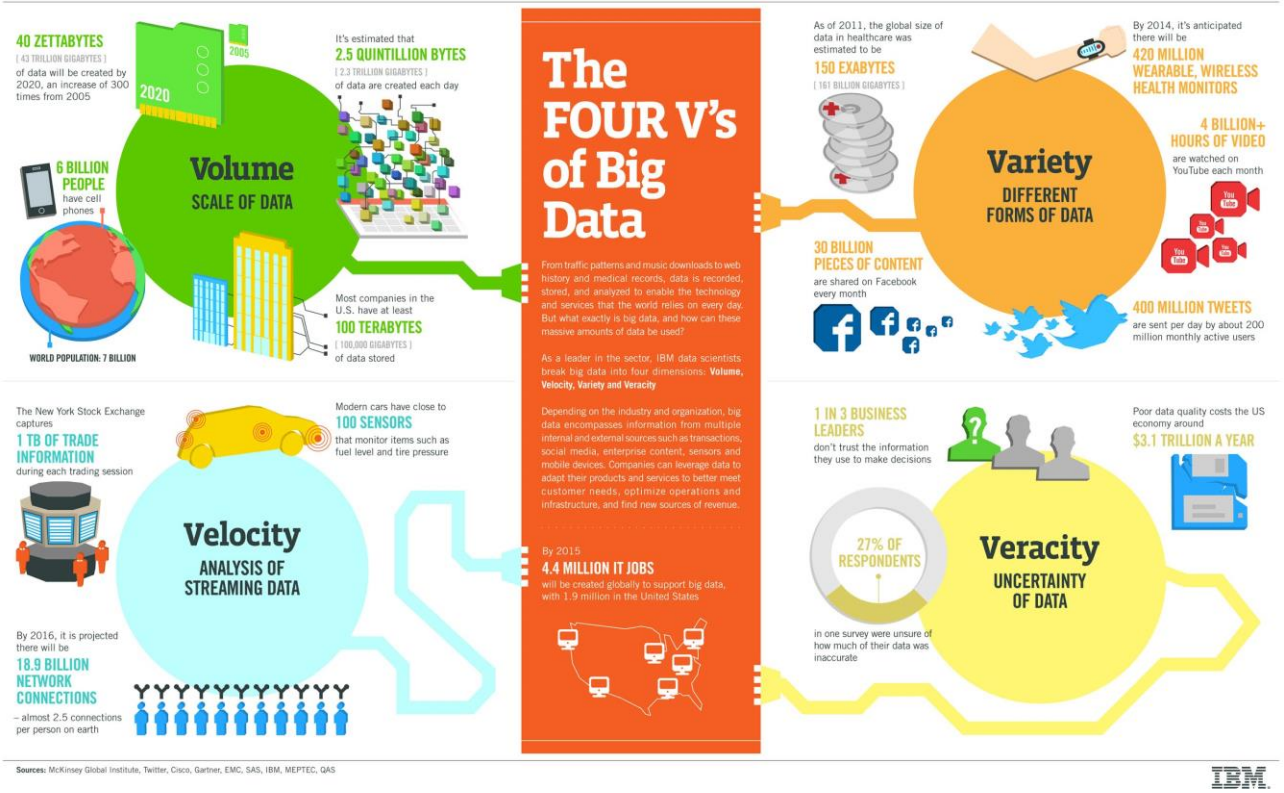


Figure 1: The 4Vs of Big Data (Available for download from https://www.ibmbigdatahub.com/infographic/four-vs-big-data)

2.2. Main challenges

As the "Analysis of the State of the Art, Barriers, Needs and Opportunities for Setting up a Transport Research Cloud" HLEG report produced by the EC concluded, although transport researchers collect and generate large amounts of data whether from monitoring actual freight/person movements, recording sensor data from vehicles and infrastructures, or capturing video of various transport related phenomena, they come from different background and apply a wide range of methods: transport research is inherently cross-disciplinary and so provides an ideal context in which to apply principles such as FAIR. "Unfortunately, most of the data that these researchers collect is used once and then stored away in locations that are inaccessible to other researchers." [4] Our experience in attempting to identify sources of available open big data confirms this.

The "Policy Briefs" document [5], produced by the NOESIS projects (mentioned in more detail in 2.3) and which pulled together the outputs and conclusions from all its work packages, presented the following main obstacles with regards to the implementation of Big Data solutions in the transportation sector:

- Lack of skilled technical personnel to work with data for Transport Authorities
- Lack of data collaboration and sharing among institutions
- Lack of understanding between transportation experts and data scientists
- Lack of access to different data sources and data variability
- Lack of standards in databases
- Lack of data quality
- Lack of infrastructure for storage
- Cost of purchasing the data

- Potential of open data not fully realized
- Uncertainty about data ownership regulation
- Uncertainty on data privacy and protection law
- Uncertainty in the benefits provided by the investment in Big Data
- Lack of financing or business models

The obstacle "potential of open data not fully realized", is very relevant here, as it seems that public or private transport organisations have not yet recognized the full potential of opening up their data [6].

As noted also by deliverable D2.2 [7] of the NOESIS project (described in the following section): "In transportation, data collection and dissemination are often fragmented and inconsistent. Though there are quite a few open data portals available at national and EU level, many transport authorities or organisations that create transport related data do not publish data in these portals because a) data is not open (there are restrictions for releasing data as open), or b) they have their own portals and they prefer to publish it there, or c) they do not wish to share data at all. In the latter case, they just keep data internally and use it only for their purpose."

Deliverable D4.1 [8] of the LeMo project (described also in the next section), which uncovers different barriers to the utilisation of big data in the transport sector, also includes a number of legal barriers and limitations specifically regarding Open Data. For example, one of the identified barriers is that Openness of (types of) data may differ between Member States, limiting in certain circumstances the scope of big data applications, including from a territorial point of view. Such issues are also covered in more depth, in their deliverable D2.2 on Legal Issues [9], which identifies and examines various legal issues that are relevant to the production of, access to, linking of and re-use of big data in the transport sector. There is a dedicated section to Open Data that analyses the opportunities and challenges in relation to open data.

In BE OPEN deliverable D4.1, "Open Science in transport research: legal issues and fundamental principles" [10], issues of data protection, IPR, security aspects, ethical concerns, and privacy and legal issues, are thoroughly presented.

2.3. Sources

Previously (in Task 2.1 [11]), the following projects related to Big Data were identified:

- **LeMO**[7] – Leveraging Big Data for managing Transport Operations: targets at developing a strategy that defines the necessary research efforts for the realisation of the big data economy in the transport sector. This is a Horizon2020 project that ended on 31 October 2020.
- **AutoMat**[8] - project with the objective to innovate an open ecosystem for Vehicle Big Data. The need to make mined and anonymous vehicle data, while building upon current trends in Big Data, is highlighted. This Horizon2020 project ended on 31 March 2018.

[7] https://lemo-h2020.eu
[8] https://automat-project.eu

In deliverable D2.2 [12], under Chapter 3.1 on "Open and FAIR data", a list of sources for accessing datasets in transport is provided. However, despite being open, none of the datasets in that list satisfy the criteria of being Big Data, so we had to expand our search.

The project **NOESIS**[9] – NOvel Decision Support tool for Evaluating Strategic Big Data investments in Transport and Intelligent Mobility Services, a 24 months Horizon2020 project, that ended on 31 October 2019.

NOESIS created a library of 73 Big Data Transport Use Cases[10] where big data has been used. Unfortunately, very few of the data are open or accessible, and many use cases have incomplete or no metadata. More use cases seem to have been added later and are now organised in a Big Data in Transport Library (BDTL) toolbox[11], containing a list of 106 Big Data in transport use cases and their generated socioeconomic value, where filters can be used to group them per transport sector (freight/passenger/both) and mode (Air/Rail/Road/Maritime/Inland waterways).

The key results of the project[12] include the NOESIS Decision Support Tool[13], an interesting online tool using machine learning techniques to predict the potential value of Big Data investments in Transport. The prediction algorithms are based on the library of use cases that the NOESIS consortium analysed. The Tool can be used for pre-screening when considering Big Data in Transport investments and solutions.

In the project's "Handbook on Key Lesson learnt and Transferable Practices" [13] the key lessons learnt related to big data use cases derived from the NOESIS Big Data in Transport Library are presented. Analysing the type of data, location and data from monitoring are used in ¾ of all use cases. It is hard to make a conclusion on data size, as more than half use cases (50.82%) did not provide sample size information; however, of those that did indicate sample size, 27.87% were of one million samples or more, and 11.48% of 100K-1M samples.

In addition, in deliverable D3.1 [14] of the project, where an earlier analysis of the BDTL attributes is carried out in order to select appropriate features for the NOESIS Decision Support Tool, we see that many of the "Data Related Information" attributes cannot be used due to incomplete or missing data. For example, apart from "data size" information already mentioned above, spatial and temporal resolution attributes (>70% empty), data velocity (40% missing), data veracity (54% missing), relational data management system and data storage information (>82% missing), and data processing technique/tool - such as Hadoop, Spark, Samza, Storm, Flink, excel/simulation model, PostgreSQL, Multiple, Mathematica (Wolfram Alpha) - (54% missing), are all data attributes that NOESIS could not use for training their decision support tool. There was enough information to mainly utilise the "data variety" attribute (only 30% missing) specifying if the data were Structured, Semi-structured, or Unstructured, and the "data analysis method" (Classification, Regression, Clustering, Bayesian networks, Cleansing and forecasting, Data mining, Cross-referencing, SPCs, Descriptive statistics, Diagram mapping, Geospatial Analysis).

[9] http://noesis-project.eu
[10] https://noesis-project.eu/survey/use_cases_overview.php
[11] https://noesis-project.eu/toolbox
[12] https://noesis-project.eu/results/
[13] https://noesis-project.eu/toolbox/decision.php

Similarly, of the "Privacy, Security, and Governance Information" attributes, the following could not be used either due to a large number of missing entries: personal privacy concern specification, current personal privacy concern level, business confidentiality problem specification, problem level, legal concern specification, and more importantly for this report the current data openness level information.

No detailed analysis of the openness of the data is provided in [13] or [14]. However, our own analysis of the NOESIS Extended Survey reports of the 106 use cases found in the BDTL showed that 78 (73.58%) of the use cases had "Null" values in the field "Current data openness level", 7 (6.60%) use cases the string "Not available", 1 (0.94%) use case also the string "Cannot disclose", and 8 (7.55%) use cases were marked as "Low", 9 (8.49%) as "Medium", and just 3 (2.83%) as "High" openness level. Thus, over 81% of the use cases have provided no openness information. Interestingly, one of the 3 "High" openness level use cases had also the remark "It should not be open", under the "Privacy Concerns" section of the Extender Survey, in the "Personal privacy concern specification" entry field. This situation once more confirms that most of the transport research data, although big, are rarely open.

The project **Track & Know**[14] – Big Data for Mobility Tracking Knowledge Extraction in Urban Areas, is a Horizon2020 project, with a focus on Big Data. The project's objective is to research, develop and exploit a new software framework that aims at increasing the efficiency of Big Data. This will be applied in the transport, mobility, motor insurance and health sectors. It is a 3-year project that will end on 31 December 2020.

Track & Know provides an Online observatory[15] which contains relevant literature and research papers, Track & Know datasets, other Big Data resources and Track & Know software. Looking at their datasets, we currently find two sets, the London Graph dataset [15] and the Tuscany City Features dataset [16].

The London Graph dataset consists of aggregate origin-destination (OD) flows of private cars in London augmented with feature data describing city locations and dyadic relations between them. The geographical location of each cell in the OD graph is not provided, for privacy protection, since the extension of each area is relatively small. The dataset consists of a file containing the nodes (grid cells) of the graph with associated features, and another one for the edges (flows between two cells). Dataset size: 6,791 Nodes, each described by 36 features, and 2,3062,231 edges, each described by 13 features (+ references to start and end cells) [17].

The Tuscany City Features dataset consists of mobility- and road network-related aggregates computed for each municipality in Tuscany (Italy), based on private cars GPS traces and the OpenStreetNetwork graph. The application that originated this dataset is the study of geographical transferability of mobility models, in order to understand under which conditions a model (e.g., to predict traffic) built in one area can be used in another one, or what kind of adjustment is needed. The dataset consists of 5 CSV files, each providing a set of features of the same family for each municipality of Tuscany [18].

[14] https://trackandknowproject.eu

[15] https://trackandknowproject.eu/file-repository/

Apart from the above Horizon2020 projects, deliverable D2.2 of project NOESIS [7] and D1.1 of project LeMO [19] mention a few additional EU-funded initiatives that are/were working in the Big Data in transport sector, including **Transforming Transport**[16] (TT), **BigDataEurope**[17] (BDE), **OPTIMUM**[18], **INOMANS²HIP**[19], **DATA SIM**[20], **CAPACITY4RAIL**[21] (C4R), and **In2Rail**[22]. Of those, Transforming Transport is the most recent and relevant project to this report, especially since NOESIS consortium was collaborating with TT and a number of its use cases were recorded to the NOESIS BDTL. TT was a Horizon2020 project (2017-2019) representing a consortium of 47 transport, logistics, and information stakeholders in Europe that demonstrated the effects of big data on mobility and logistics.

The oneTRANSPORT project[23] (Collaborative R&D, Innovate UK), with an aim to make transport more accessible and usable, provides an open and scalable platform for multi-modal and multi-system transport integration across local authorities. oneTRANSPORT focuses on integrating transport modes and systems from 4 contiguous counties. Their data catalogue can be browsed here: https://onetransport.io/data-catalogue. However, in order to use this platform, a subscription fee is required.

Looking beyond projects, there are some organisations that collect real-time data (e.g., mobility) and have their data open. Examples of such organisations are:

- Transport for London (https://tfl.gov.uk/info-for/open-data-users/our-open-data): opening up TfL which was valued at 17-68 million Euros per year, resulted in over 200 travel applications developed by private companies [20], thus generating large economic benefits by reducing travel time, traffic congestion, pollution and greenhouse gas emissions.
- DB, the German national railway company (https://data.deutschebahn.com)
- EMT in Madrid (https://www.emtmadrid.es/EMTBUS/Mibus)
- Rennes Métropole (https://data.rennesmetropole.fr/explore/?sort=modified)
- Austin Texas' Open Data Portal (https://data.austintexas.gov)

Data in transport can come from different sources, such as:

- Route-based (sensors along a particular route such as a highway or rail)
- Vehicle-based (generated by sensors in a vehicle and GPS)
- Traveller-based (collected from travellers and mobile devices)
- Wide-area data (collected by sensor networks monitoring traffic flows, as well as by drones and space-based satellites)

and offer a variety of data, including:

[16] https://transformingtransport.eu
[17] https://www.big-data-europe.eu
[18] http://www.optimumproject.eu
[19] https://cordis.europa.eu/project/id/266082/reporting
[20] https://cordis.europa.eu/project/id/270833
[21] http://www.capacity4rail.eu
[22] http://www.in2rail.eu
[23] https://onetransport.io

- location data from mobile devices and objects that show where, when, and how objects or people more, such as GPS feeds from vehicles, or from mobile phones and wearable devices, data collected from Wi-Fi of Bluetooth devices, RFID tags, and so on,
- hot sports, such as bus roots, car parking locations, speed limit datasets, locations of bus stops, ports, rails, airports, etc.,
- monitoring devices, e.g., surveillance, traffic or speed cameras, aerial photos, cameras that count people or vehicles, wearable devices, various sensors including parking and traffic sensors, or sensors that monitor the functioning condition of different parts of cars, trucks, buses, trains, etc.,
- physical events, such as roadwork locations, real-time traffic incident reports, planned road closures, etc.,
- environmental data from sensors that create real-time data for air quality, weather, rainfall, wind, etc.
- transaction data that reveal consumption patterns, including credit card spending data, petrol prices, fares, loyalty card data, bookings and reservations, shipments, rates, etc. (but mostly closed data)
- social media data from Twitter feeds, Facebook, etc.
- digital maps that can be very useful for OBD applications in transport, such as those provided by OpenSteetMap

Below we present a few examples of transport related datasets that we identified either directly or were made aware of them via the above-mentioned projects and resources. These are datasets that are both open and big data (OBD).

Example of Identified OBD datasets are provided below, the first of which was discovered by searching the Transport Research Community Gateway in OpenAIRE[24], which forms part of the TOPOS Observatory[25] for Organizations (whereas Scipedia hosts the TOPOS Observatory for Individuals).

- πNEUMA, hosted at EPFL https://open-traffic.epfl.ch/, is a large-scale dataset of naturalistic trajectories of half a million vehicles that have been collected by a one-of-a-kind experiment by a swarm of drones in the congested downtown area of Athens, Greece. A unique observatory of traffic congestion, a scale an-order-of-magnitude higher than what was not available until now, that researchers from different disciplines around the globe can use to develop and test their own models. The dataset is ideal for multimodal research (the vehicles available are: Car, Taxi, Bus, Medium Vehicle, Heavy Vehicle, Motorcycle) but can be used only for academic, non-commercial purposes. The data are provided in CSV format, split in different files depending on location, data and time. The data frequency is 25 FPS. See also the relevant publication available in the TOPOS Observatory. https://beopen.openaire.eu/search/publication?articleId=dedup_wf_001::a2b3bea9d0324f308b75041c38150600 (https://doi.org/10.1016/j.trc.2019.11.023)

- Live daily data on traffic for the city of Thessaloniki, Greece, provided by CERTH. It also provides Historical Data exported in CSV format. https://trafficthess.imet.gr

[24] https://beopen.openaire.eu
[25] https://www.topos-observatory.eu

- Almost real-time Taxi data from the city of Thessaloniki, also by CERTH. The data are provided under a CC BY-NC 2.0 license in a variety of formats: JSON, XML, CSV, KML, MAP. http://opendata.imet.gr/dataset/fcd-gps

- The Cityscapes 3D Dataset, released this August (2020), is an extension of the original Cityscapes with 3D bounding box annotations for all types of vehicles as well as a benchmark for the 3D detection task, as described in a paper, presented at the CVPR 2020 Workshop on Scalability in Autonomous Driving[26]. It includes 3D bounding box annotations of all vehicle types, i.e., car, truck, bus, on rails, motorcycle, bicycle, caravan, and trailer. The box annotations feature a full 3D orientation including yaw, pitch, and roll labels. The annotations are available for free download on their download page; however, registration is required first. They also provide a toolbox available on GitHub[27]. https://www.cityscapes-dataset.com/cityscapes-3d-dataset-released/

- Real time bus and river bus locations across all Transport for London (TfL) bus stops. http://countdown.api.tfl.gov.uk/interfaces/ura/instant_V1

- UK Oyster card data, as well as multimodal datasets for Journey Planning (current and future), Status (current and future), Disruptions (current) and Planned works (future), Arrival/departure predictions (instant and web-sockets), Timetables, Embarkation points and facilities, Routes and lines (topology and geographical), and Fares. https://tfl.gov.uk/corporate/privacy-and-cookies/oyster-card

- Timetable data for each line on the London Underground, provided in CVS files, based on version 2.0 of the Open Government Licence. http://timetables.data.tfl.gov.uk

- Real time traffic incident reports from Austin-Travis County, updated every 5 minutes, available in a number of formats including CSV, TSV, RDF, XML, etc. https://data.austintexas.gov/Transportation-and-Mobility/Real-Time-Traffic-Incident-Reports/dx9v-zd7x

- Dataset containing information about traffic cameras in Austin, TX. Traffic cameras are owned and operated by the City of Austin Transportation Department and are used to monitor traffic conditions across the city. Available in a number of formats including CSV, TSV, RDF, XML, etc. https://data.austintexas.gov/Transportation-and-Mobility/Traffic-Cameras/b4k4-adkb

- Real time traffic cameras in Washington with open API. The complete reference documentation for TrafficView API is found here: https://www.trafficview.org/developers https://www.trafficview.org/traffic_cameras/

- Traffic information data of the German national railway company DB, including live schedule data (departure and arrival times) and delays. The data is available for free use and further

[26] https://arxiv.org/abs/2006.07864
[27] https://github.com/mcordts/cityscapesScripts

processing, in machine-readable and openly licensed form, in various formats (including a ReST API, JSON files, etc.), permanently and free of charge. https://data.deutschebahn.com

- iRail supports digital creativity concerning mobility in Belgium. This is an attempt to make the railway time schedules in Belgium easily available for anyone. The iRail open data API allows anyone to query trains, stations, liveboards and connections.
  https://hello.irail.be/api/1-0/

- Flightradar24 is a global flight tracking service that provides real-time information about thousands of aircraft around the world. Their service is currently available online and for iOS or Android devices. It is a subscription-based service but includes a Basic free plan for private use.
  https://www.flightradar24.com

- Data contains a list of all Metro and Ulsterbus Bus Stops across Northern Ireland, under an Open Government License, and in CVS and GeoJSON formats.
  https://data.gov.uk/dataset/495c6964-e8d2-4bf1-9942-8d950b3a0ceb/translink-bus-stop-list

- Payment parking places in the City of London, provided in WMS format.
  https://data.gov.uk/dataset/937bdb4b-4e0c-4ea9-9b43-ec6eaffb9aa4/payment-parking-places

There are also ODB sources of relevant environmental data, such as:

- Live access to air quality information and data for London in a structured format, under an Open Government License.
  http://www.londonair.org.uk/LondonAir/API/

- API delivering 2 Billion weather forecasts per day. A subscription-based service but with a free basic account option also.
  https://openweathermap.org/api

- 1,000 real time rain gauges which are connected by telemetry, provided by the UK Environment Agency's API. https://environment.data.gov.uk/flood-monitoring/doc/rainfall

2.4. Characteristics and Properties

Lack of standards in databases is one of the challenges mentioned in section 2.2. Similarly, Kumar and Prakash [21] pointed out that there is a lack of standards regarding data formats and structures for many types of data relevant to transport. In many open data portals, raw data feeds are presented as CSV, XLS, and XML files, as well as PDF documents and a variety of other formats [7]. Our analysis of the OBD datasets presented in the previous section, confirm that the majority of datasets are available mostly as CSV, XLS, JSON, and/or XML files, or machine-readable data feeds.

The inventor of the Web and Linked Data initiator, Tim Berners-Lee, suggested a 5-star deployment scheme for Open Data [22], as shown in Figure 2, which can be used to measure the technical usability of data.

1. The 1-star badge is given to data that is available on the web in whatever format (e.g., a PDF document) under an Open License (OL).
2. The 2-star badge is given to data available as structured data (e.g., XLS/Excel file, instead of an image scan of a table) and so is also machine REadable (RE).
3. The 3-star badge is given to data available in a non-proprietary (e.g., CSV file) Open Format (OF).
4. The 4-star badge is given to Uniform Resource Identifiers (URIs) to denote things, so that people can point at it (e.g., RDF data).
5. The 5-star badge is given to Linked Data (LD), so that the data are linked to other data to provide context (e.g., LOD cloud).

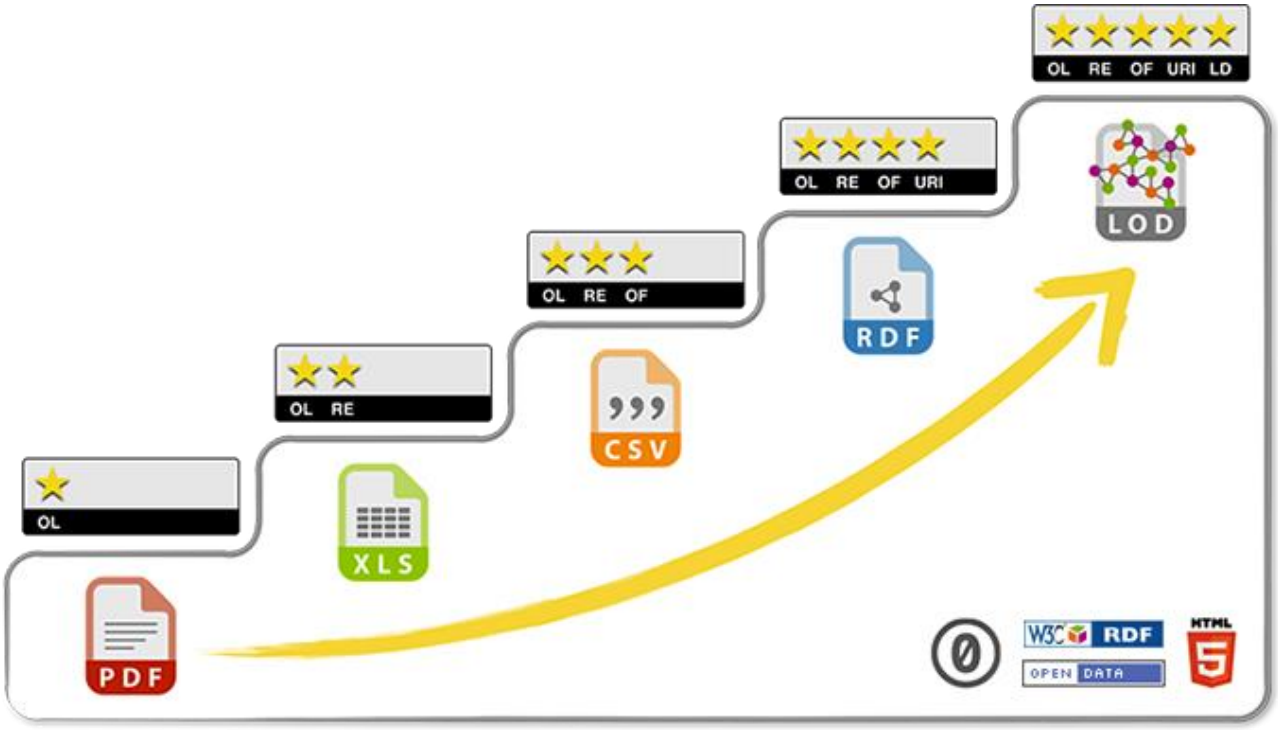


Figure 2: The 5-star deployment scheme for Open Data

Although this 5-star scheme was created with Web and Linked Open Data in mind, what applies also to OBD are the goals of having data provided under an Open Licence (OL), data that are machine REadable (RE), and data available in a non-proprietary Open Format (OF); the 3-star badge.

As OBD are open by definition, the need for OL is self-explanatory, and the requirement for OF is for alignment with FAIR principles (see next chapter). The data must also be RE because of the sheer volume of big data: data need to be in a format that can be easily processed by a computer without human intervention, as humans are unable to process such huge volumes of data. That means that data must be structured data, but it does not mean that they cannot also be human-readable or understandable by humans.

For tabular data, which is the most common structure for data, the simplest format is CSV and its varieties like Tab-Separated Values (TSV). Note that CSV files may have issues with the separator

which can lead to data quality issues. Although the most usable format for data is likely to be one in which the dataset was originally created, in many cases this format may be a proprietary software programme like Microsoft Excel. Excel, and other similar spreadsheet software, come with richer content, such as styling on tables and graphs of data which can help to give context to a human who is trying to understand the data, but this does not satisfy the OF criterion. Thankfully, most such programmes have the ability to export data in open formats[28]. CSV does not support column types (no difference between text and numeric columns), there is no standard way to present binary data, and has poor support for special characters. Despite these limitations, CSV files are the most common choice for data exchange of tabular data, as CSV is human-readable, easy to read and edit, provides a simple scheme, can be processed by virtually all existing applications, is easy to implement and parse, and is compact (column headers are written only once, unlike XML tags that are repeated for each column in each row).

However, not all data can be suitably expressed in a tabular form. Other key structures to be aware of are hierarchical and network data. Hierarchical data show the relationships between data points, such as a family tree. If the dataset depends on the relationship between data points and follows a structure in which data points are linked in vertical 'trees', a hierarchical data structure in a format such as JSON is ideal. JSON is often compared to XML because it can store data in a hierarchical format. Both formats are user-readable, but JSON documents are typically much smaller than XML. They are therefore more commonly used in network communication, especially with the rise of ReST-based web services [23]. Most computer languages provide simplified JSON serialization libraries or built-in support for JSON serialization/deserialization, and there is built-in support in most modern tools. JSON supports lists of objects, helping to avoid chaotic list conversion to a relational data model, and is a widely used file format for NoSQL databases, such as MongoDB. JSON is the standard for communication on the Internet (APIs and websites communicate with JSON due to its convenient features of clearly defined schemes and the ability to store complex hierarchical data). However, JSON seems to be the worst format to use for performance [23].

MessagePack[29] is like JSON but faster and smaller. It offers an efficient binary serialization format that can exchange data among multiple languages like JSON. Small integers are encoded into a single byte, and typical short strings require only one extra byte in addition to the strings themselves. Similarly, BSON[30] is another binary-encoded version of JSON (its name stands for Binary JSON) and was created by MongoDB. It is lightweight, traversable, and efficient. BSON also con-tains ex-ten-sions that al-low rep-res-ent-a-tion of data types that are not part of the JSON spec. For ex-ample, BSON has a Date type and a BinData type.

Network structured data allows relationships to exist between any combination of elements in any direction, such as in a social network. The Web is another example of a network data structure, where web pages link to any number of other pages in any direction.

A particular set of open data are geospatial data which are oftentimes more complex than simple tabular datasets and can exist as a hierarchical dataset (e.g., detailing countries and counties/states), or as a network dataset (e.g., detailing roads) [24]. Geospatial data are usually

[28] https://www.europeandataportal.eu/elearning/en/module9/#/id/co-01
[29] https://msgpack.org
[30] http://bsonspec.org

published in data formats like GeoJSON (based upon JavaScript Object Notation, JSON)[31] or KML (based upon Extensible Markup Language, XML)[32]. These formats are specifically designed with usability in mind and can easily be imported and exported from specialist mapping tools like Open Street Map[33] and CartoDB[34].

Tabular data is the best suited for download. As we've seen in the previous section (2.3), most government data portals provide tabular data. However, as the data is frequently too large, we see that it is frequently split and organised into smaller datasets. For services which provide data that are updated very frequently, and are live or near-live, an API is provided instead. A machine interface has the advantage that it can directly integrate these services into other Web applications. We have already seen (section 2.3) many examples of services providing an open data API, such as the Transport for London (TfL), the German national railway company DB, the UK Environment Agency, and the Belgian railway company iRail, among others.

Other big data file formats include:

Avro[35], a data serialization system, which is very good for storing row data, very efficient. Avro provides rich data structures, a compact, fast, binary data format, a container file, to store persistent data, Remote Procedure Call (RPC), and simple integration with dynamic languages. Code generation is not required to read or write data files nor to use or implement RPC protocols. Code generation is provided as an optional optimization, only worth implementing for statically typed languages. Avro has a schema and supports evolution, providing excellent integration with Kafka, and supports file splitting. This allows old software to read new data, and new software to read old data; a critical feature for evolving data. All versions of the schema are stored in a human-readable JSON header, making it easy to understand all available fields. Avro is a relatively compact option for both permanent storage and wire transfer. Since Avro is a row-based format, it is the preferred format for handling large amounts of records as it is easy to add new rows.

Protocol Buffers[36] are Google's language-neutral, platform-neutral, extensible mechanism for serializing structured data; similar to XML, but smaller, faster, and simpler. You define how you want your data to be structured once, then you can use special generated source code to easily write and read your structured data to and from a variety of data streams and using a variety of languages. Protocol Buffers are great for APIs, especially for gRPC[37] (RPC stands for Remote Procedure Call), or machine learning. It supports schema and it is very fast.

BSON, mentioned earlier, can be compared to binary interchange formats, like Protocol Buffers. BSON is more "schema-less" than Protocol Buffers, which can give it an advantage in flexibility but also has a slight disadvantage in space efficiency (BSON has overhead for field names within the serialized data)[38].

[31] https://geojson.org
[32] https://developers.google.com/kml/documentation/kml_tut
[33] https://www.openstreetmap.org/
[34] https://carto.com
[35] https://avro.apache.org
[36] https://developers.google.com/protocol-buffers/
[37] https://grpc.io
[38] http://bsonspec.org

Apache *Parquet*[39] is a columnar storage format available to any project in the Hadoop ecosystem, regardless of the choice of data processing framework, data model or programming language. It has schema support and works very well with Hive[40], Spark[41], and Dask[42] as a way to store columnar data in deep storage that is queried using SQL, but also NoSQL (for example, Dask first converts Parquet to Pandas[43]). It is designed for HDFS but can also be stored on other file systems. Because it stores data in columns, query engines will only read files that have the selected columns and not the entire data set as opposed to Avro (the concept is called *projection pushdown*). Useful as a reporting layer. As a columnar storage format, data can also be highly compressed (up to 75% with snappy[44]). Parquet files are binary files that contain metadata about their contents. It is the fastest format for read-heavy processes. It also provides *predicate pushdown,* reducing the further cost of transferring data from storage to the processing engine for filtering. One of its disadvantages is that it does not support data modification (files are immutable) or schema evolution.

Apache *ORC*[45]: Similar to Parquet, offers better compression. It also provides better schema evolution support as well, but it is less popular. ORC is a self-describing type-aware columnar file format designed for Hadoop workloads. It is optimized for large streaming reads, but with integrated support for finding required rows quickly. Storing data in a columnar format lets the reader read, decompress, and process only the values that are required for the current query. Because ORC files are type-aware, the writer chooses the most appropriate encoding for the type and builds an internal index as the file is written.

In conclusion, CSV is typically the fastest format to write, JSON the easiest to understand for humans, Avro the fastest to read all columns at once, while Parquet the fastest to read a subset of columns. Columnar formats are typically used where several columns need to be requested rather than all columns as their column-oriented storage design is well suited for this purpose. On the other hand, row-based formats are used where all fields in a row need to be accessed. This is why Avro is usually used to store the raw data because all fields are usually required during ingestion. And Parquet is used after pre-processing for further analytics because usually all fields are no longer required there [23]. For working with text formats, JSON and CSV when one can control the data types (no text with special characters) are good choices. Unless the data source imposes it, it is best to avoid selecting XML. It is practically not used in data processing because of its high complexity, often requiring a custom parser. In terms of streaming processing, Avro and Protocol Buffers are privileged formats. Avro is a flexible and powerful format, while also providing great support for schema evolution. In addition, Protocol Buffers is the foundation of gRPC on which the Kubernetes[46] ecosystem relies on. However, while Protocol Buffers are efficient in streaming workloads, this format is not splitable and not compressible which makes it less attractive for storing data at rest. Finally, column storage offered by ORC and Parquet provide a significant performance boost in data and business Intelligence analytics [25]. Table 1 provides a comparison of the different big data file formats.

[39] https://parquet.apache.org
[40] https://hive.apache.org
[41] https://spark.apache.org
[42] https://dask.org
[43] https://pandas.pydata.org
[44] https://github.com/google/snappy
[45] https://orc.apache.org
[46] https://kubernetes.io

Table 1: Overview of Big Data File Formats (adapted from [25])

Types	CSV	JSON	XML	AVRO	Protocol Buffers	Parquet	ORC
Text versus binary	text	text	text	metadata in JSON, data in binary	text	binary	binary
Data type	no	yes	no	yes	yes	yes	yes
Schema enforcement	no (minimal with header)	external for validation	external for validation	yes	yes	yes	yes
Schema evolution	non	yes	yes	yes	non	yes	non
Storage type	row	row	row	row	row	column	column
OLAP/OLTP	OLTP	OLTP	OLTP	OLTP	OLTP	OLAP	OLAP
Splitable	yes, in its simplest form	yes, with JSON lines	no	yes	no	yes	yes
Compressed	no	no	no	no	no	yes	yes
Stream	yes	yes	non	yes	yes	non	non
Typed data	non	non	non	non	yes	non	non
Ecosystems	popular everywhere for its simplicity	API and web	enterprise	Big Data and Streaming	RPC and Kubernetes	Big Data and BI	Big Data and BI

2.5. Licenses

The various licenses for open data are described in BE OPEN Deliverable D4.3 [26] and apply also for OBD. These are normally *Creative Commons (CC)*[47] copyright licenses that come in a number of flavours, as depicted in Figure 3.

All Creative Commons licenses have many important features in common. Every license helps creators (licensors) retain copyright while allowing others to copy, distribute, and make some uses of their work, at least non-commercially. Every Creative Commons license also ensures licensors get the credit for their work they deserve. Every Creative Commons license works around the world and lasts as long as applicable copyright lasts (because they are built on copyright). These common features serve as the baseline, on top of which licensors can choose to grant additional permissions when deciding how they want their work to be used.

[47] https://creativecommons.org/licenses/

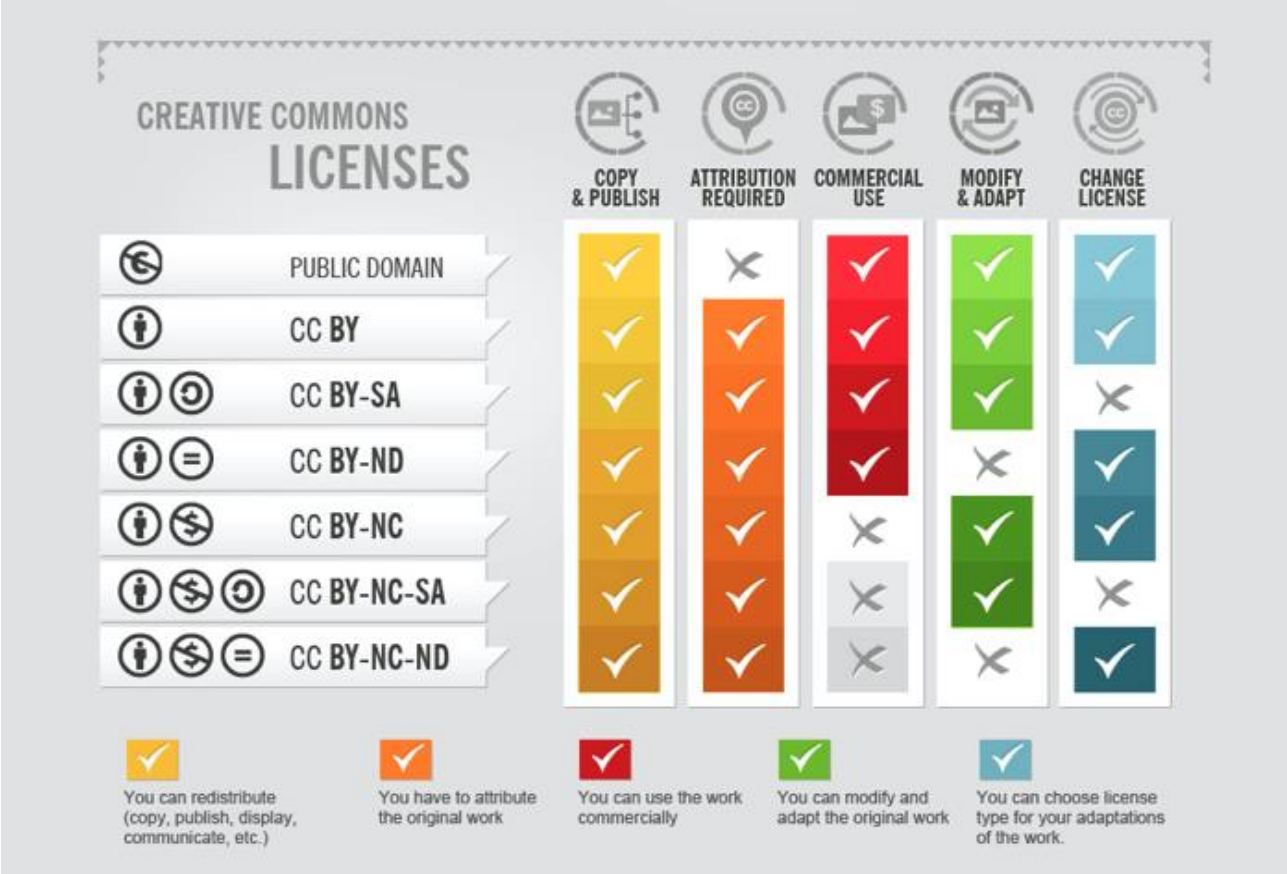


Figure 3: Creative Commons Licenses[48]

The reader of section 2.3 may have noticed the mention of an *Open Government License*. This is a copyright license from Crown Copyright works published by the UK Government. UK public sector bodies may apply it to their publications. Developed and maintained by The National Archives[49], it is compatible with the Creative Commons Attribution (CC-BY) licence. A similar open license has been created by the French government[50].

In addition, there are bespoke or custom-made licences. These are created by various data publishers and introduce specific conditions with which the user must comply. Bespoke or custom-made licences can be written by the publisher or adapted from a standard licence through the addition of new conditions and/or the modification of existing ones. However, these licences can increase complexity for users of open data, as they can introduce specific conditions that limit usage, restrict data integration and, in some cases, are difficult for users to comply with.

[48] https://foter.com/blog/how-to-attribute-creative-commons-photos/
[49] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/
[50] https://www.etalab.gouv.fr/licence-ouverte-open-licence

3. FAIR Recommendations

The increasing availability of online resources means that data need to be created with longevity in mind. Providing other researchers with access to data facilitates knowledge discovery and improves research transparency. In this context, during the Lorentz Workshop "Jointly Designing a Data FAIRport" (2014), participants formulated the FAIR data vision to optimise data sharing and reuse by humans and machines, which resulted in the publication of “The FAIR Guiding Principles for scientific data management and stewardship” [1] .

The FAIR principles define 15 concepts (or high-level recommendations) to ensure that research data is Findable, Accessible, Interoperable, and Reusable by humans and machines. The OpenAIRE guide “How to make your data FAIR” [3] summarises the 15 principles, described in detail in [1,2] by highlighting four basics for achieving FAIRness:

- Adoption of a standard identification system (PID, persistent identifiers, like DOI or Handle)
- Openness of metadata
- Adoption of formats that can be understood by humans and machines
- Use the least restricted license you can use

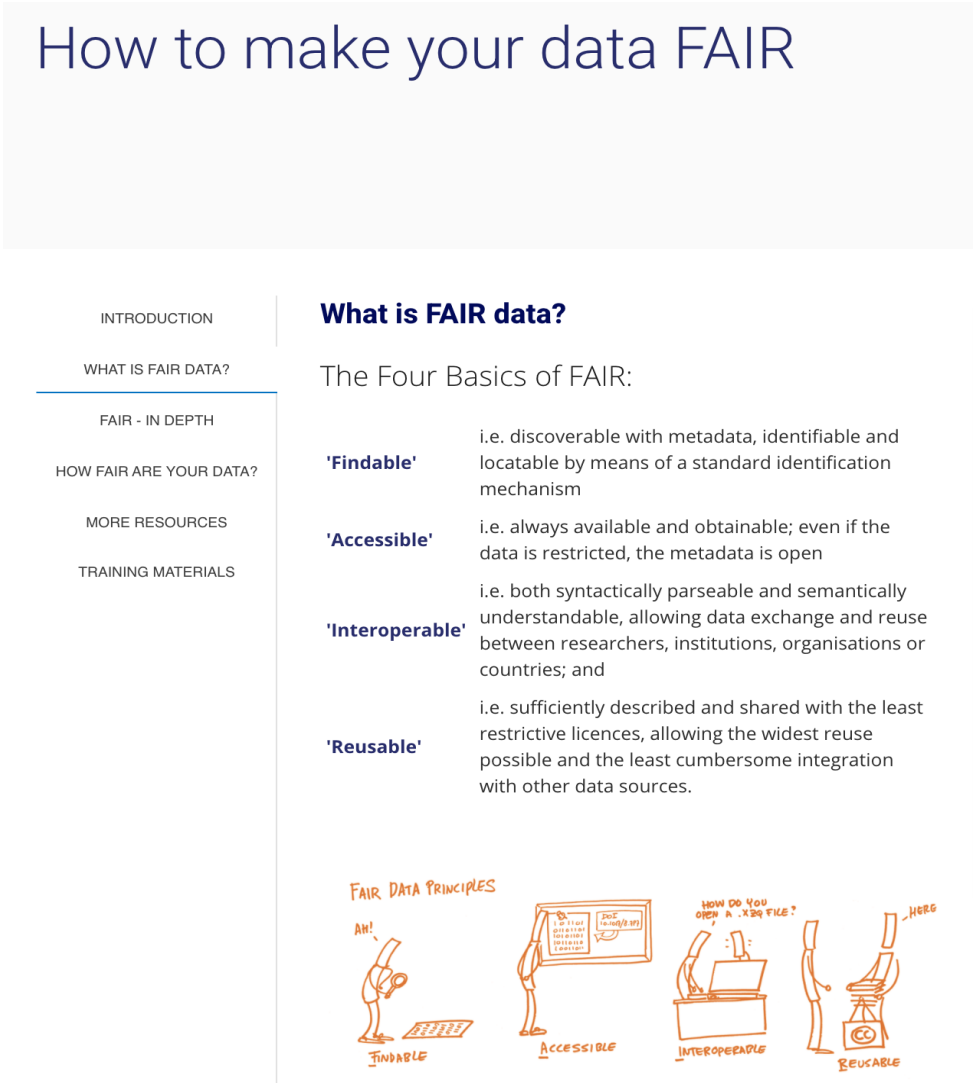


Figure 4: OpenAIRE guide “How to make your data FAIR” [3]

The final report of the expert group for the European Commission "Analysis of the State of the Art, Barriers, Needs and Opportunities for Setting up a Transport Research Cloud" [4] highlights the low uptake of FAIR principles in the Transport research community and calls for actions to raise awareness and empower the community with tools supporting the implementation of Open Science and FAIR principles to transport research data.

Based on the categorization of transport research data presented in [4], the hereafter recommendations apply to *original research data obtained by observations* (e.g., data from Field Operational Tests (FOTs), Naturalistic Driving Studies (NDS)) and *data used in published transport research appearing in scientific venues* (e.g., research results, models from papers published in journals or conference proceedings). Unless otherwise specified, the term "transport research data" in the remainder of the deliverable denotes those two types of data (i.e., excluding, for example, operational and governmental data).

Recommendations have been synthesized from the report of the expert group on Transport Research Cloud [4] and the RDA endorsed recommendation on FAIR data maturity model [7]. While [4] addresses the specific discipline of Transport research, [7] provides insights on domain-independent FAIRness indicators. The two reports together form a valuable source of information and knowledge for the formulation of pragmatic recommendations for FAIR data in transport research.

It is also worth to highlight that FAIRness should not be intended as a yes/no feature, but rather a continuum: when analysing the FAIRness of datasets, we should not ask if datasets are FAIR or not, but rather how FAIR they are and how the FAIRness could be improved [2,7].

Table 2 summarises the recommendations for FAIR Open transport research data. Some of the recommendations target researchers, while others target the research community as a whole and could be seen as activities suggested to the BE OPEN consortium to support, foster, promote, and facilitate researchers of the domain to make their data comply to as much as FAIR principles as possible.

Table 2: Recommendations for FAIR Open transport research data

FAIR principle	Indicator	Priority	Recommendation	Target
Findable				
F1. (Meta)data are assigned a globally unique and persistent identifier	Metadata is identified by a persistent identifier	Essential	**Rec. 1**: Adoption of persistent unique identifiers (PIDs) issued by organisations with a clear persistence policy (e.g., DOI, Handle). A list of possible PID registration agencies is available on the RDA-endorsed registry	Researchers & Research community
	Data is identified by a persistent identifier	Essential		
	Metadata is identified by a	Essential		

	globally unique identifier		FAIRsharing[51] [7] **Rec. 2**: Ensure a PID resolves to the landing page of the data that contains both the metadata and the access URL to the actual data content. Otherwise, metadata and data must have one PID each and the metadata should include the PID of the data. Note that the Transport research community may want to align on the approach and modify the recommendation to include one of the two options, as suggested in [7].	
	Data is identified by a globally unique identifier	Essential		
F2. Data are described with rich metadata	Rich metadata is provided to allow discovery	Essential	**Rec. 3**: adopt a metadata format based on the principles of DCAT-AP that also include: links to publications and other related research products, links to funding and projects, temporal and geographic information, access information (e.g., access rights, URLs, formats, licenses), quality information (e.g., update frequency, collection methodology, original purpose of the collection). Evaluate the opportunities of	Researchers & Research community

[51] https://fairsharing.org/standards/?q=&selected_facets=type_exact:identifier%20schema

			adoption (possibly after modification) of ISO/IEC 11179, ISO 14817, and ASTM E2468-05 [4].	
F3. Metadata clearly and explicitly include the identifier of the data they describe	Metadata includes the identifier for the data	Essential	See **Rec. 2**	Researchers & Research community
F4. (Meta)data are registered or indexed in a searchable resource	Metadata is offered in such a way that it can be harvested and indexed	Essential	**Rec. 4**: deposit data and metadata on a trusted Open Access repository. A list of possible repositories can be found on re3data[52] **Rec. 5**: the list of thematic trusted repositories should be made available by the BE OPEN TOPOS **Rec. 6**: in case of lack of thematic repository for transport research data, the consortium may create and manage a Zenodo community that researchers can use to deposit the data and metadata	Researchers & Research community
Accessible				
A1. (Meta)data are retrievable by their identifier using a standardised communications protocol	Metadata contains information to enable the user to get access to the data	Important	See **Rec. 3**	Researchers & Research community
	Metadata can be	Essential	See **Rec. 4**	Researchers (this

[52] https://www.re3data.org/

	accessed manually (i.e., with human intervention)			indicator can be addressed by choosing a proper repository for deposition)
	Data can be accessed manually (i.e., with human intervention)	Essential	See **Rec. 4**	Researchers (this indicator can be addressed by choosing a proper repository for deposition)
	Metadata identifier resolves to a metadata record	Essential	See **Rec. 2**	Researchers (this indicator can be addressed by following the community policy on PIDs)
	Data identifier resolves to a digital object	Essential	See **Rec. 2**	Researchers (this indicator can be addressed by following the community policy on PIDs)
	Metadata is accessed through standardised protocol	Essential	See **Rec. 4**	Researchers (this indicator can be addressed by choosing a proper repository for deposition)
	Data is accessible through standardised protocol	Essential	See **Rec. 4**	Researchers (this indicator can be addressed by choosing a proper repository for deposition)
	Data can be accessed automatically (i.e., by a computer program)	Important	See **Rec. 2** and **Rec. 4**	Researchers (this indicator can be addressed by choosing a proper repository for deposition)
A1.1 The	Metadata is	Essential	See **Rec. 4**	Researchers (this

protocol is open, free, and universally implementable	accessible through a free access protocol			indicator can be addressed by choosing a proper repository for deposition)
	Data is accessible through a free access protocol	Important	See **Rec. 4**	Researchers (this indicator can be addressed by choosing a proper repository for deposition)
A1.2 The protocol allows for an authentication and authorisation procedure, where necessary	Data is accessible through an access protocol that supports authentication and authorisation	Useful	See **Rec. 4**	Researchers (this indicator can be addressed by choosing a proper repository for deposition)
A2. Metadata are accessible, even when the data are no longer available	Metadata is guaranteed to remain available after data is no longer available	Essential	See **Rec. 4**	Researchers (this indicator can be addressed by choosing a proper repository for deposition)
Interoperable				
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	Metadata uses knowledge representation expressed in standardised format	Important	**Rec. 7**: use FAIR controlled vocabularies for relevant properties of the metadata format (see also **Rec. 3**)	Researchers
	Data uses knowledge representation expressed in standardised format	Important	**Rec. 8**: Define the data model using proper standards (see also **Rec. 3**) **Rec. 9**: Use standard non-proprietary formats for the data.	Researchers & Research community

			Rec. 10: Analyse the formats suggested by the European ITS Directive (2010/40/EU) [8] (DATEX II, NaTEX, SIRI, TN-ITS, Inspire) to verify they fit the FAIR requirements [7]	
	Metadata uses machine-understandable knowledge representation	Important	See **Rec. 3** and **Rec. 4**	Researchers & Research community
	Data uses machine-understandable knowledge representation	Important	See **Rec. 9** and **Rec. 10**	Researchers and Research community
I2. (Meta)data use vocabularies that follow FAIR principles	Metadata uses FAIR-compliant vocabularies	Important	See **Rec. 7**	Researchers
	Data uses FAIR-compliant vocabularies	Useful	See **Rec. 8** and **Rec. 9**	Researchers & Research community
I3. (Meta)data include qualified references to other (meta)data	Metadata includes references to other metadata	Important	See **Rec. 3**	Researchers & Research community
	Data includes references to other data	Useful	See **Rec. 8** and **Rec. 10**	Researchers & Research community
	Metadata includes references to other data	Useful	See **Rec. 3**	Researchers & Research community
	Data includes qualified references to other data	Useful	See **Rec. 8**	Researchers & Research community

	Metadata includes qualified references to other metadata	Important	See **Rec. 3**	Researchers & Research community
	Metadata include qualified references to other data	Useful	See **Rec. 3**	Researchers & Research community
Reusable				
R1. (Meta)data are richly described with a plurality of accurate and relevant attributes	Plurality of accurate and relevant attributes are provided to allow reuse	Essential	See **Rec. 3**	Researchers & Research community
R1.1. (Meta)data are released with a clear and accessible data usage license	Metadata includes information about the licence under which the data can be reused	Essential	**Rec. 11**: Adopt a license that allows access to researchers from different countries [4]. See also **Rec. 3**	Researchers
	Metadata refers to a standard reuse licence	Important	**Rec. 12**: Adopt a standard license (like Creative Commons[53] or Open Data Commons[54]) [7]	Researchers
	Metadata refers to a machine-understandable reuse licence	Important	**Rec. 13**: The adopted license should be available in machine-readable format See also **Rec. 12**: standard licenses usually are available in a machine-readable format	Researchers
R1.2. (Meta)data	Metadata	Important	See **Rec. 3**	Researchers &

[53] https://creativecommons.org/about/cclicenses/
[54] https://opendatacommons.org/licenses/

are associated with detailed provenance	includes provenance information according to community-specific standards			Research community
	Metadata includes provenance information according to a cross-community language	Useful	**Rec. 14**: use a cross-domain standard to describe provenance (e.g., PROV-O[55]) or a community-specific standard that can be mapped to a cross-domain standard [7] See also **Rec. 3**	Researchers & Research community
R1.3. (Meta)data meet domain-relevant community standards	Metadata complies with a community standard	Essential	See also **Rec. 3** and **Rec. 7**	Researchers & Research community
	Data complies with a community standard	Essential	See also **Rec 8 Rec. 9** and **Rec. 10**	Researchers & Research community
	Metadata is expressed in compliance with a machine-understandable community standard	Essential	**Rec. 15:** Describe the metadata model in a machine-readable format See also **Rec. 3** and **Rec. 7**	Researchers & Research community
	Data is expressed in compliance with a machine-understandable community standard	Important	**Rec. 16:** Describe the data model in a machine-readable format See also **Rec 8 Rec. 9** and **Rec. 10**	Researchers & Research community

[55] https://www.w3.org/TR/prov-o/

3.1. Evaluating FAIRness of Open Big Datasets in Transport

Section 2 of this deliverable identified a number of Open Big Datasets in transport and analysed their characteristics and properties. Among those datasets, two can be considered transport research data: pNEUMA and the Cityscapes 3D Dataset. The other datasets belong to the categories of operational and governmental data, which do not fit in the definition of transport research data provided at the beginning of this section[56]. The evaluations are performed based on the indicators available in Table 2.

i. Checking the FAIRness of pNEUMA

pNEUMA, hosted at EPFL https://open-traffic.epfl.ch/, is a large-scale dataset of naturalistic trajectories of half a million vehicles that have been collected by a one-of-a-kind experiment by a swarm of drones in the congested downtown area of Athens, Greece. It is a unique observatory of traffic congestion, with a scale an order of magnitude higher than what was available until now. Researchers from different disciplines around the globe can use it to develop and test their own models. The dataset is ideal for multimodal research (the vehicles available are: car, taxi, bus, medium vehicle, heavy vehicle, motorcycle) but can be used only for academic, non-commercial purposes. Data is provided in CSV format, split in different files depending on location, data and time. The dataset was identified by the BE OPEN consortium via its journal publication, with a search on the TOPOS Observatory[57].

The level of FAIRness of pNEUMA has been assessed based on the information available in the accompanying publication [10], on the web site (https://open-traffic.epfl.ch/) and part of the data as downloaded from the pNEUMA web site.

As detailed in the paragraphs below, the pNEUMA dataset does not have a high level of compliance to the FAIR principles. In fact, dataset creators focused on humans, providing everything needed for a researcher to access, interpret, and re-use the dataset on a dedicated web site. On the other hand, FAIR principles focus on machine-actionability, which has not been addressed when publishing the dataset. Nevertheless, even a basic FAIRification process (e.g., creation and deposition of metadata in a certified repository, assignment of a PID) would have a major effect on the FAIRness of pNEUMA and would foster wider awareness of the dataset among the research community.

Findability

F1. (Meta)data are assigned a globally unique and persistent identifier
The pNEUMA dataset does not have a globally unique and persistent identifier, although it can be identified by the URL of the web site dedicated to it: https://open-traffic.epfl.ch/ . Citations to the dataset are made via the publication that describes the experiment thanks to which the dataset was generated and how the data can be used for multi-modal transport research activities [10]. The

[56] It does not mean that operational and governmental data are not relevant for transport research. They are indeed relevant and useful to conduct research studies and analysis, however they fall beyond the type of data we consider "transport research data" in the context of this deliverable. As highlighted in the HLEG report [4] there is no consensus yet on a definition of transport research data and one of its recommendations is to bring together all stakeholders to agree on a definition to be used in the context of the Transport Research Cloud.

[57] Link to the pNEUMA publication in the TOPOS Observatory: https://beopen.openaire.eu/search/publication?articleId=dedup_wf_001::a2b3bea9d0324f308b75041c38150600. (DOI: https://doi.org/10.1016/j.trc.2019.11.023) Last Accessed 11 December 2020

dataset must also be acknowledged in the acknowledgement statement of publications that used the dataset, as described in the Download FAQs[58].

F2. Data are described with rich metadata
No accompanying metadata could be found in the web site[59] from which it is possible to download the data. The dataset is described in detail in human readable form in the accompanying journal article. This is very useful for human consumption of the dataset, but not very helpful to machine consumption, which is also the main focus of the FAIR principles.

F3. Metadata clearly and explicitly include the identifier of the data they describe
No metadata is available.

F4. (meta)data are registered or indexed in a searchable resource
The dataset has been found via the OpenAIRE Community Gateway for Transport Research (part of the BE OPEN TOPOS), which included the journal article describing the experiments that produced the research data. The dataset (or a description of it) could not be found in any of the following scholarly communication sources: OpenAIRE, Datacite, BASE, Google Dataset Search, Mendeley Data, Academic Torrents. Based on the information available on the pNEUMA dataset and its accompanying publications, no information about the storage and the adopted policies to ensure the persistency of the dataset

Accessibility

A1. (meta)data are retrievable by their identifier using a standardized communications protocol
No retrievable metadata is available: the dataset is described by its accompanying publication. The data itself is instead retrievable via the web site https://open-traffic.epfl.ch/index.php/downloads/ via https. The download option is available for humans but not machines.

A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
The data can be downloaded by humans submitting a form. The data is downloaded as a CSV file via HTTPS.

A2. metadata are accessible, even when the data are no longer available.
If the data will no longer be available, the web site will probably be shut down. However, the human readable metadata contained in the publication describing the data will still be accessible.

Interoperability

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
No retrievable metadata is available: the dataset is described by its accompanying publication. Data can be downloaded as a CSV file where columns and string values are written in English. No unit of measures are specified in the CSV.

[58] pNEUMA Download FAQ: https://open-traffic.epfl.ch/index.php/downloads/. Last accessed 12 November 2020
[59] pNEUMA dataset web site: https://open-traffic.epfl.ch/. Last accessed 30 October 2020

I2. data use vocabularies that follow FAIR principles.
Data contains only one column whose values could be terms of a vocabulary: column "type", informing about the type of vehicles. The Glossary for Transport Statistics [9] could have been used, but it seems this was not the case.

I3. data include qualified references to other (meta)data.
No.

Reuse

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
The dataset is accompanied by the web site and the publication, which provide very detailed information about the dataset. Data provenance is described in detail on the web site and in the publication in human readable format. The dataset is open for academic, non-commercial purposes. The licensing information is available in the FAQ of the web site. No standard license is referred.

ii. Checking the FAIRness of Cityscapes 3D

The Cityscapes 3D Dataset is a large-scale dataset that contains a diverse set of stereo video sequences recorded in street scenes from 50 different cities, mainly in Germany, with high quality pixel-level annotations of 5 000 frames in addition to a larger set of 20 000 weakly annotated frames. The dataset is presented in a scientific publication [11], and it is accompanied by a toolbox, which includes a set of scripts, on GitHub that support its re-use[60] and by a web site (https://www.cityscapes-dataset.com) where one can find all information for accessing, interpreting, and re-using the dataset.

The level of FAIRness of Cityscapes 3D Dataset has been assessed based on the information available in the accompanying publication, on the web site, and on the GitHub project of the toolbox. As detailed in the paragraphs below, the Cityscapes 3D dataset does not have a high level of compliance to the FAIR principles. In fact, dataset creators focused on humans, providing everything needed for a researcher to access, interpret, and re-use the dataset on a dedicated web site. On the other hand, FAIR principles focus on machine-actionability, which has not been addressed when publishing the dataset. Nevertheless, even a basic FAIRification process (e.g., creation and deposition of metadata in a certified repository, assignment of a PID) would have a major effect on the FAIRness of the dataset and would foster wider awareness of the dataset among the research community.

Findability

F1. (Meta)data are assigned a globally unique and persistent identifier
The Cityscapes 3D dataset does not have a globally unique and persistent identifier, although it can be identified by the URL of the web site dedicated to it: https://www.cityscapes-dataset.com . Citations to the dataset are made via the publication presenting the dataset, as suggested by dataset creator and maintainers on the web site.

F2. Data are described with rich metadata
Data comes with rich metadata, such as GPS coordinates and outside temperature from vehicle

[60] https://github.com/mcordts/cityscapesScripts

sensors, that is intended to enrich the data and support additional use cases. Available metadata is not, therefore, devised to support the findability of the dataset itself. The dataset is described in detail in human readable form in the accompanying scientific publication and the web site. This is very useful for human consumption of the dataset, but not very helpful to machine consumption, which is the main focus of the FAIR principles.

F3. Metadata clearly and explicitly include the identifier of the data they describe
No metadata is available.

F4. (meta)data are registered or indexed in a searchable resource
The dataset could not be found in a searchable scholarly communication resource for datasets. In particular, we searched for it in the following online resources: OpenAIRE (it includes the scientific publication that describes the dataset[61]), Datacite, BASE, Mendeley Data. A search on Google Dataset Search, instead, gave us two relevant results: one entry from Kaggle, where we can find a processed subsample of the dataset[62] and one entry on Academic Torrent that is apparently a 2016 copy of the dataset[63], in possible infringement to the official license agreement[64]. We could not find any information about the adopted policies to ensure dataset persistence on the web site and in the related publications.

Accessibility

A1. (meta)data are retrievable by their identifier using a standardized communications protocol
No retrievable metadata is available: the dataset is described by its accompanying publication. The data itself is instead retrievable via the web site https://www.cityscapes-dataset.com upon user registration.

A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
Data cannot be downloaded automatically but requested by registering to the web site.

A2. metadata are accessible, even when the data are no longer available.
If the data will no longer be available, the web site will probably be shut down. However, the human readable information contained in the publication describing the data will still be accessible.

Interoperability

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

[61] OpenAIRE EXPLORE page for the scientific publication: https://explore.openaire.eu/search/publication?articleId=dedup_wf_001::715e4e3409e65fae4e0306637ca2087d. Last accessed 11 December 2020.
[62] Cityscapes Image Pairs on Kaggle https://www.kaggle.com/dansbecker/cityscapes-image-pairs (2018). Last accessed 11 December 2020.
[63] The Cityscapes Dataset for Semantic Urban Scene Understanding on Academic Torrents https://academictorrents.com/details/4f76b97fbb851fac002dcc55dcc55883e9728db7 (2016). Last accessed 11 December 2020.
[64] Cityscapes 3D Dataset license agreement: https://github.com/mcordts/cityscapesScripts/blob/master/license.txt. Last accessed 11 December 2020

No retrievable metadata is available: the dataset is described by its accompanying publication. The format and language of the data is not explicitly described on the website.

I2. data use vocabularies that follow FAIR principles.
Authors of the datasets defined their own vocabularies for classes and labels as explained in the supplementary material of the publication[65] and on the web site. Classes and labels are not available in machine readable format nor described by dedicated metadata records.

I3. data include qualified references to other (meta)data.
No.

Reuse

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
The dataset is accompanied by the web site and the publication, which provide detailed information about the dataset. Data provenance is described in detail in the publication in human readable format. The dataset is open for research, non-commercial purposes. The licensing information is available in the GitHub page of the toolbox on GitHub[66]. No standard license is referred.

[65] The Cityscapes Dataset for Semantic Urban Scene Understanding – SUPPLEMENTAL MATERIAL –. https://www.cityscapes-dataset.com/wordpress/wp-content/papercite-data/pdf/cordts2016cityscapes-supplemental.pdf Last accessed 11 December 2020
[66] Cityscapes 3D Dataset license https://github.com/mcordts/cityscapesScripts/blob/master/license.txt Last accessed 11 December 2020

4. Open Big Data Management

Research data is the basis on top of which scientific knowledge grows. As such, it is important that research data, open, restricted, big, or "small", is properly managed so that it is not lost and remains accessible and usable on the long term by the research community. More and more funders and research organisations are mandating clear Research Data Management (RDM) practices and proposing RDM guidelines. For example, the European Commission considers responsible data management a fundamental part for an effective implementation of the Open Science paradigm [1, 2].

As reported in the previous chapters, open big research data in the transport sector are not easy to find for different reasons. The lack of best practices and guidelines for the transport research community and the low level of awareness of RDM among researchers are two of those reasons. However, on the BE OPEN TOPOS platform provided by OpenAIRE[67], one can see that when transport data are generated in the context of EC H2020 projects, researchers do deposit the data according to the guidelines of the Commission.

The promotion of the Open Science paradigm and supporting its implementation are key objectives of the OpenAIRE initiative. As a member of the BE OPEN consortium, OpenAIRE can contribute to the implementation of Open Science practices across the community in transport research by raising awareness in the research community on policies, practices, and providing technical services to ease RDM.

The adoption of responsible RDM practices can be supported by tools that facilitate the definition of Data Management Plans (DMPs) and the management of sensitive and personal data. As of January 2021, the EOSC marketplace includes 87 services under the broad "Data Management" category[68]. Of those, three services are for the definition of DMPs: DMPOnline by the Digital Curation Centre[69], the Data Stewardship Wizard[70], and Argos by OpenAIRE[71], which we describe in section 4.1. For the management of sensitive and personal data, other two services could be found: Services for Sensitive Data (TDS) by University of Oslo[72], and AMNESIA by OpenAIRE[73], which we describe in section 4.2

As guidance on RDM practices (in and out of the scope of the European Commission), OpenAIRE provides an RDM handbook [4] and guides [5] on different aspects of responsible RDM. In addition, OpenAIRE offers two services for RDM practices: Argos for the definition of Data Management Plans (DMPs) (section 4.1), and Amnesia for the anonymisation of sensitive and personal data (section 4.2).

[67] https://beopen.openaire.eu
[68] https://marketplace.eosc-portal.eu/services/c/data-management
[69] https://marketplace.eosc-portal.eu/services/dmponline?fromc=data-management
[70] https://marketplace.eosc-portal.eu/services/data-stewardship-wizard
[71] https://marketplace.eosc-portal.eu/services/argos?fromc=data-management
[72] https://marketplace.eosc-portal.eu/services/tds?fromc=data-management
[73] https://marketplace.eosc-portal.eu/services/amnesia?fromc=data-management

4.1. DMP tools (Argos)

DMPs hold information about novel datasets (including raw and auxiliary data) that are produced, but also about data that are re-used for the purpose of a scientific mission/ project. DMPs offer valuable documentation regarding how data have been handled, processed, curated, published and preserved throughout a data management lifecycle and therefore serve as the bridge to reproducibility of research and reusability of data by other researchers and interested parties, such as SMEs, easing their further and long-term exploitation [3].

Several DMP tools are being developed and embedded in the RDM lifecycle by funders and organisations. Those tools differ at conceptual (e.g., target audience, functionalities, cost) and/ or technical (e.g., data model, technology) levels. Despite aforementioned differences and deviations, most DMP tools are gradually moving towards applying the Research Data Alliance (RDA) standard [6] for interoperability and machine-actionability with successful examples being, among other Argos[74] by OpenAIRE, Data Stewardship Wizard[75] and the DMPonline[76].

OpenAIRE's Argos is a running instance of the OpenDMP open-source software[77], a result of a collaboration between OpenAIRE and EUDAT CDI[78] to deliver an open platform for Data Management Planning. Details on the functions and architecture of Argos can be found in [3].

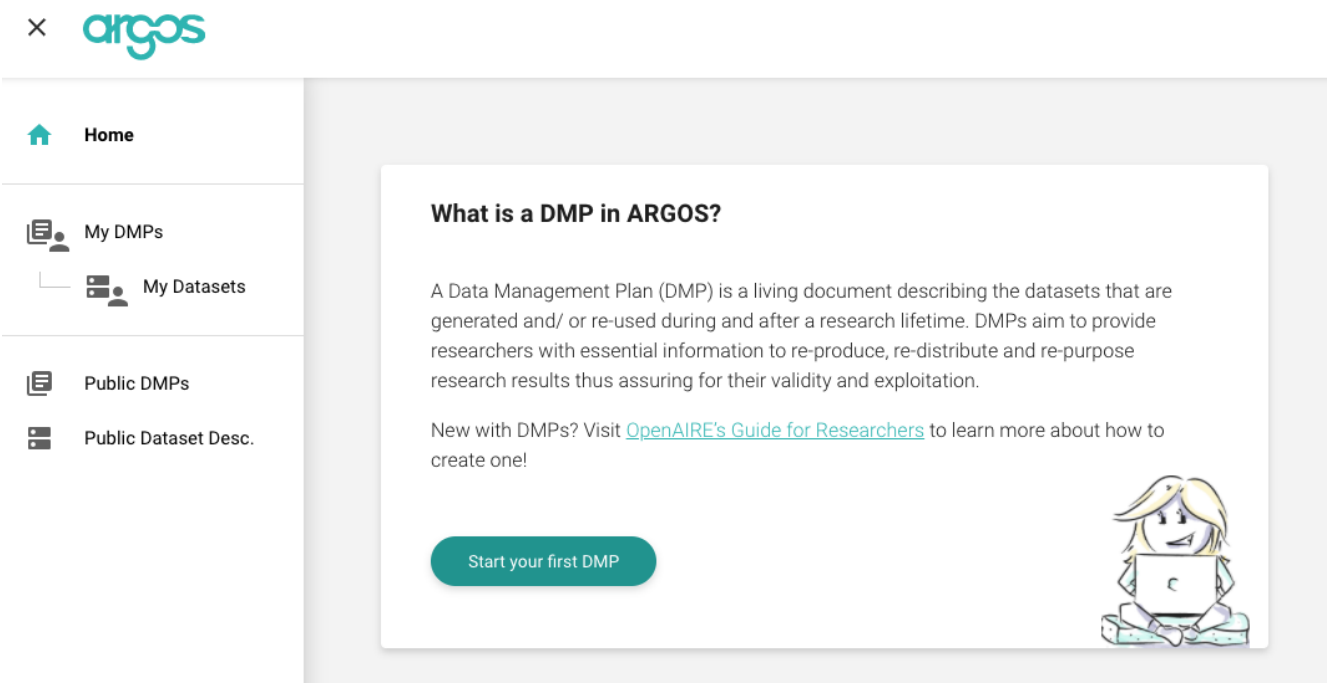


Figure 5: Home page of Argos

[74] https://argos.openaire.eu/
[75] https://ds-wizard.org/
[76] https://dmponline.dcc.ac.uk/
[77] Wiki page https://gitlab.eudat.eu/dmp/OpenAIRE-EUDAT-DMP-service-pilot/-/wikis/home
[78] https://eudat.eu/eudat-cdi

4.2. Anonymization: dealing with sensitive data (Amnesia tool)

Research data may contain sensitive information, and this is a limiting factor for sharing. Co-developed with Athena Research & Innovation Center as part of the OpenAIRE suite of services, Amnesia[79] is a tool to help researchers publish data which contain sensitive information, by anonymizing the data and ensuring GDPR compliance. There are a couple of open software anonymization tools available, but with its high accuracy and state-of-the-art algorithms, which are continuously updated, Amnesia offers one easy solution.

Amnesia is an application written in Java and JavaScript and should be used locally for anonymizing a dataset. The basic idea behind anonymization is that a file containing personal data is loaded to Amnesia (the original dataset) and Amnesia transforms it to an anonymous dataset, which can then be stored locally. The transformation is guided by user choices and provides an anonymization guarantee for the result. An easy-to-use graphical interface allows users to tailor the anonymization to their needs, by guiding the algorithm and deciding trade-offs with simple visual choices. In addition, developers can incorporate Amnesia anonymization engine to their project through a ReST API.

The guarantees currently supported by Amnesia are *k-anonymity* and *km-anonymity*. That means that the original data are transformed by generalizing (i.e., replacing one value with a more abstract one) or suppressing values to achieve the statistical properties required by the anonymisation guarantees. Amnesia allows minimal reduction of information quality and guarantees no links to the original data. The transformed datasets are treated as statistics by GDPR. Hence, anonymous data can be used without the need for consent or other GDPR restrictions, greatly reducing the effort needed to extract value from them. The documentation page[80] guides users on how to follow the simple 5-step anonymization process for anonymizing a dataset.

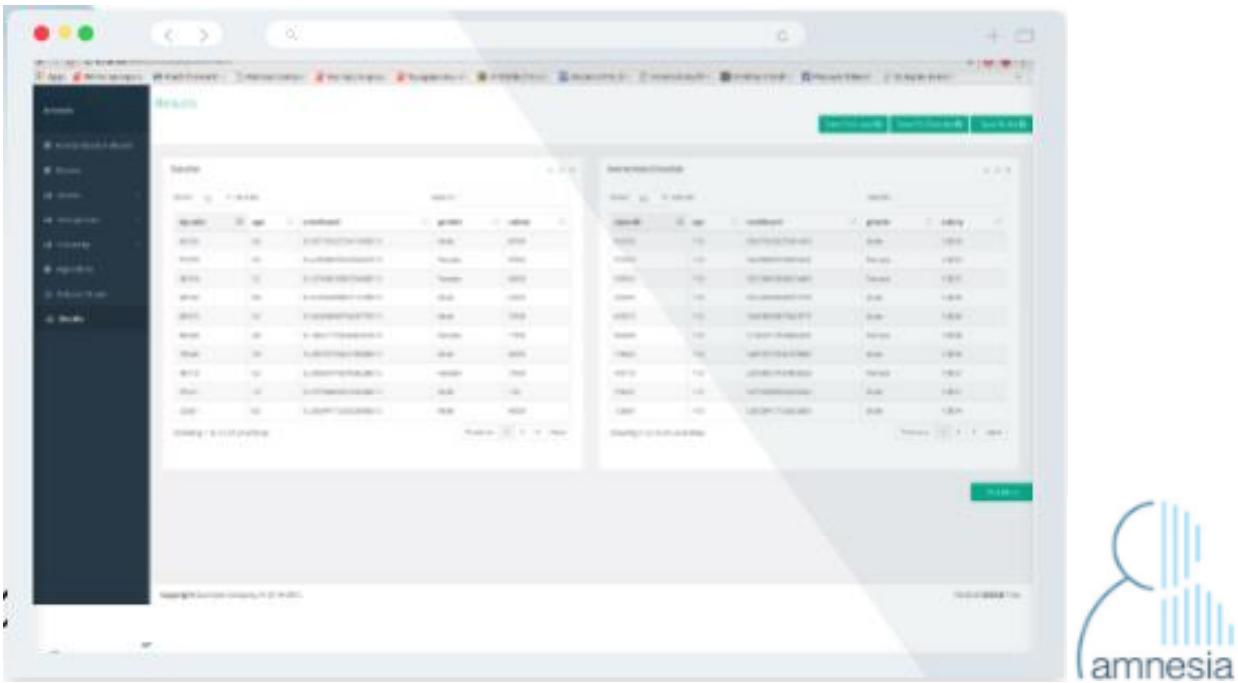


Figure 6: Amnesia anonymization tool

[79] https://amnesia.openaire.eu
[80] https://amnesia.openaire.eu/about-documentation.html

5. Conclusions and Discussion

Open Big Data in the transport research sector are not easy to find. This report confirms the remarks of the HLEG report and the NOESIS project regarding the low uptake of FAIR and Open principles by the transport research community, especially when big datasets are involved. Transport research is inherently cross-disciplinary and so provides an ideal context in which to apply principles such as FAIR but unfortunately, "most of the data that these researchers collect is used once and then stored away in locations that are inaccessible to other researchers." [1]

The OpenAIRE Gateway on Transport Research[81] (which is part of the BE OPEN TOPOS) identifies, as of December 2020, only about 800 research data correctly deposited and made available via a trusted repository in the OpenAIRE network[82]. None of those could be identified as "big" according to the famous 4 Vs. In fact, for the identification of the open big datasets analysed in Chapter 2, the most effective methodologies have been going through the deliverables of projects in the domain and through word-of-mouth. This way a number of big data sets were found but the challenge then was to find which of those were also openly available. For example, our own analysis of the NOESIS Extended Survey reports of the 106 transport use cases found in the Big Data in Transport Library toolbox[83] showed that over 81% of the use cases provided no openness information, confirming that most of the transport research data, although big, are rarely open.

Eventually, the open big datasets that we discovered were less than twenty and are presented in Chapter 2, followed by an analysis of their general characteristics and properties, including an overview of big data file formats and their licenses. The majority of datasets identified are provided mostly as CSV, XLS, JSON, and/or XML files, or machine-readable data feeds, and are made available under one of the Creative Commons (CC) copyright licenses for open data, or very similar licenses such as the UK *Open Government License*.

Nonetheless, most of the identified open big datasets belong to the categories of operational and governmental data, rather than original research data obtained by observations or data used in published transport research appearing in scientific venues. Only two open big datasets from Chapter 2 can be considered research data: *pNEUMA* and *Cityscapes 3D*. The two datasets went through a FAIR assessment process that revealed a low level of FAIRness in both cases. The assessment was conducted considering the FAIR principles and the set of indicators defined by the RDA FAIR Data Maturity Model Working Group. It is worth highlighting that very simple actions like depositing bibliographic metadata about the dataset in an open institutional or thematic repository and assigning a persistent identifier could already, with minor effort, have a massive impact on the level of FAIRness, especially for its findability and discoverability thanks to its inclusion in the open scholarly communication infrastructure. The FAIRness assessment is detailed in Chapter 3, preceded by a set of 14 recommendations for the research community in transport research, obtained as a synthesis of the RDA FAIR Data Maturity Model indicators and the recommendations of the EC HLEG report. The 14 recommendations suggest actions not only to single researchers but also to the research community as a whole. In fact, achieving FAIRness is a process that one person cannot conduct alone: the research community, as a whole, should agree on best practices at the local,

[81] https://beopen.openaire.eu
[82] Mostly from generic research data repositories: Figshare and Zenodo.
[83] https://noesis-project.eu/toolbox

national, and international level. In this context, the role of the TOPOS that BE OPEN is setting up is fundamental because it could be the hub where different stakeholders of the domain gather, communicate with each other, and have access to guidelines and best practices that have been previously discussed and agreed in the TOPOS as well. Some of those guidelines and best practices, however, need not be defined from scratch, since several aspects of data FAIRness and data management are generic and can be addressed by building on top of the results obtained by other more mature research communities.

Chapter 4 focuses on two major challenges of open big data management in transport research: planning data management and anonymisation of sensitive data. In fact, the two challenges are generic: they also apply to other domains, as well as to “small” research data, and can therefore be tackled with already existing tools, that can be possibly customised for given peculiarities, if needed. As examples, we briefly described two OpenAIRE products that address the two challenges: *Argos* for the definition of machine-actionable Data Management Plans and *Amnesia* for the anonymisation.

REFERENCES

Chapter 2 – Open Big Data in Transport Research

[1] D1.2: Open Science framework, terminology and instruments, 2019. BE OPEN project. https://beopen-project.eu/storage/files/beopen-d12-open-science-framework-terminology-and-instruments.pdf

[2] Helena Gellerman, Erik Svanberg, Yvonne Barnard. Data sharing of transport research data. s.l.: Science Direct, 2016.

[3] Laney, D., 3D Data Management: Controlling Data Volume, Velocity, and Variety, Technical report, META Group, 2001.

[4] Analysis of the State of the Art, Barriers, Needs and Opportunities for Setting up a Transport Research Cloud, European Commission, Directorate-General for Research and Innovation, DOI: 10.2777/77906, Luxembourg: Publications Office of the European Union, 2018.

[5] Policy Briefs, D6.5 of NOESIS project, 2020. Available at: https://drive.noesis-project.eu/index.php/s/cNup8ZG1pUL7BCG#pdfviewer

[6] Soriano, F.R., Samper, J.J., Martinez, J.J., Cirilo, J.V., & Carrillo, E. (2016) Smart cities technologies applied to sustainable transport. Open data management. In Telematics and Information Systems (EATIS), 2016 8th Euro American Conference on (p. 1-5). IEEE.

[7] D2.2: Big Data implementation context in transport, 2018. NOESIS project. Available at: https://drive.noesis-project.eu/index.php/s/MmtDNXgQ2th3lVq

[8] D4.1: Report on the characterization of the barriers and limitations, 2019. LeMo project. https://lemo-h2020.eu/newsroom/2019/10/1/new-report-published-identification-and-characterisation-of-barriers-and-limitations

[9] D2.2: Report on Legal Issues, 2018. LeMo project. Available at: https://lemo-h2020.eu/newsroom/2018/11/1/deliverable-d22-report-on-legal-issues

[10] D4.1: Open Science in transport research: legal issues and fundamental principles, 2020. BE OPEN project. Available at: https://beopen-project.eu/storage/files/beopen-d41-open-science-in-transport-research-legal-issues-and-fundamental-principles.pdf

[11] D2.1: Open access publications and the performance of the European transport research. 2019. BE OPEN project. Available at: https://beopen-project.eu/storage/files/beopen-d21-open-access-publications-and-the-performance-of-the-european-transport-research.pdf

[12] D2.2: Open/FAIR data, software and infrastructure in European transport research, 2019. BE OPEN project. Available at: https://beopen-project.eu/storage/files/beopen-d22-open-fair-data-software-and-infrastructure-ineuropean-transport-research.pdf

[13] D4.3: Handbook on Key Lesson learnt and Transferable Practices, 2019. NOESIS project. Available at: https://drive.noesis-project.eu/index.php/s/Uigt0ggLZtQ6WS2#pdfviewer

[14] D3.1: Learning Framework Methodology and Architecture, 2018. NOESIS project. Available at: https://drive.noesis-project.eu/index.php/s/10Iz4nw1GHGbMXh

[15] London Graph dataset (PDF) description. Available at: https://trackandknowproject.eu/wp-content/uploads/simple-file-list/Track-and-Know-Datasets/London_Graph_Dataset.pdf

[16] Tuscany City Features dataset (PDF) description. Available at: https://trackandknowproject.eu/wp-content/uploads/simple-file-list/Track-and-Know-Datasets/Tuscany_City_Features-dataset.pdf

[17] Gevorg Yeghikyan, Felix L. Opolka, Bruno Lepri, Mirco Nanni, Pietro Lio. Learning Mobility Flows from Urban Features with Spatial Interaction Models and Neural Networks. 2020 IEEE International Conference on Smart Computing (SMARTCOMP), to appear. Available at: https://arxiv.org/abs/2004.11924

[18] Leonardo Longhi, Mirco Nanni. Car telematics big data analytics for insurance and innovative mobility services. In Journal of Ambient Intelligence and Humanized Computing (JAHC), p. 1-11, Springer, 2017. DOI: 10.1007/s12652-019-01632-4
[19] D1.1: Understanding and Mapping Big Data in Transport Sector, 2018. LeMO project. https://static1.squarespace.com/static/59f9cdc2692ebebde4c43010/t/5b49c213352f534ffb42e3d8/1531560480749/20180711_D1.1_Understanding+and+mapping+big+data+in+transport+sector_LeMO.pdf
[20] Sweeney, R. Opening data fully to improve London's transport network, Int. Assoc. Public Transp., 2018, UK
[21] Kumar, A., & Prakash, A. (2014) The role of big data and analytics in smart cities. International Journal of Science and Research IJSR, 6(14), 12-23.
[22] Five-star Open Data webpage. https://5stardata.info/
[23] Big Data file formats, webpage: https://luminousmen.com/post/big-data-file-formats
[24] Choosing the right format for open data, European Data Portal webpage: https://www.europeandataportal.eu/elearning/en/module9/#/id/co-01
[25] Comparison of different file formats in Big Data, webpage: https://www.adaltas.com/en/2020/07/23/benchmark-study-of-different-file-format/
[26] D4.3: New business models to implement Open Access in transport research, Oct 2020. BE OPEN project (to be available here: https://beopen-project.eu/resources/deliverables)
[27] Data formats for preservation, OpenAIRE webpage: https://www.openaire.eu/data-formats-for-preservation/
[28] File formats recommended and accepted by the UK Data Service for data sharing, reuse and preservation: https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats
[29] Big Data file formats explained, webpage: https://towardsdatascience.com/big-data-file-formats-explained-dfaabe9e8b33
[30] EU Best Practice: Open Up Public Transport Data, 2016. Available at: https://www.w3.org/2013/share-psi/bp/ptd/

Chapter 3 – FAIR Recommendations

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[2] Website of [1] above: https://www.go-fair.org/fair-principles/
[3] OpenAIRE Guide "How to make your data FAIR" https://www.openaire.eu/how-to-make-your-data-fair. Last accessed: September 2020.
[4] Analysis of the State of the Art, Barriers, Needs and Opportunities for Setting up a Transport Research Cloud, European Commission, Directorate-General for Research and Innovation, DOI: 10.2777/77906, Luxembourg: Publications Office of the European Union, 2018
[5] D2.2: Draft Governance Framework for the European Open Science Cloud, 2017. EOSC pilot project. Available at: https://eoscpilot.eu/content/d22-draft-governance-framework-european-open-science-cloud
[6] OpenAIRE webpage: https://www.openaire.eu/mission-and-vision
[7] RDA FAIR Data Maturity Model Working Group (2020). FAIR Data Maturity Model: specification and guidelines. Research Data Alliance. DOI: 10.15497/RDA00050
[8] European Union Directive 2010/40/EU of the European Parliament and of the Council of 7 July 2010, "on the framework for the deployment of Intelligent Transport Systems in the field of road transport and for interfaces with other modes of transport," Official Journal of the European Union, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32010L0040

[9] European Commission. Statistical Office of the European Union., United Nations., & International Transport Forum (ITF). (2019). Glossary for transport statistics, 2019. Publications Office. Available at: https://doi.org/10.2785/675927

[10] Barmpounakis, Emmanouil, and Nikolas Geroliminis. "On the new era of urban traffic monitoring with massive drone data: The pNEUMA large-scale field experiment." Transportation research part C: emerging technologies 111 (2020): 50-71. https://doi.org/10.1016/j.trc.2019.11.023

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. https://arxiv.org/abs/1604.01685

Chapter 4 – Open Big Data Management

[1] Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information C/2018/2375 http://data.europa.eu/eli/reco/2018/790/oj

[2] European Commission. (2016). European Cloud Initiative - Building a competitive data and knowledge economy in Europe. Luxembourg: Office for Official Publications of the European Communities. https://ec.europa.eu/digital-single-market/en/news/communication-european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe Last accessed 15 December 2020

[3] Papadopoulou, E., Bardi, A., Kakaletris, G., Tziotzios, D., Manghi, P., & Manola, N. (2020). Data Management Plans and Linked Open Data: exploiting machine actionable data management plans through Open Science Graphs. 1st Workshop on Research Data Management for Linked Open Science (DaMaLOS@ISWC). https://doi.org/10.4126/FRL01-006423283

[4] OpenAIRE's handbook on Research Data Management: https://www.openaire.eu/rdm-handbook Last accessed 15 December 2020

[5] OpenAIRE's guides on Open Science: https://www.openaire.eu/guides . Last accessed 15 December 2020

[6] Research Data Alliance DMP Common Standard https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard . Last accessed 15 December 2020

Chapter 5 – Conclusions and Discussion

[1] Analysis of the State of the Art, Barriers, Needs and Opportunities for Setting up a Transport Research Cloud, European Commission, Directorate-General for Research and Innovation, DOI: 10.2777/77906, Luxembourg: Publications Office of the European Union, 2018.