Deliverable 2.2

Conclusion of data register, feature report and accreditation

UK participants in Horizon Europe Project PHOEBE are supported by UKRI grant numbers 10038897 (The International Road Assessment Programme – iRAP) and 10056912 (The Floow)

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101076963

Document Control Page

Deliverable title	*Conclusion of data register, feature report and accreditation*
Deliverable number	*2.2*
Deliverable version	*1 (V1.0)*
Work Package number	*2*
Work Package Title	*Data specification and collection (AI & machine learning)*
Due date of delivery	*30/04/2025*
Actual date of delivery	*30/04/2025*
Dissemination level	*Sensitive*
Type	*Report*
Editor(s)	*Sam Chapman, FLOOW* *Mark Burke, FLOOW*
Contributor(s)	*Sam Chapman, FLOOW* *Mark Burke, FLOOW*
Reviewer(s)	*Haris Sideris, O7* *Marcel Sala, AIM*
Project name	*Predictive Approaches for Safer Urban Environments*
Project Acronym	*PHOEBE*
Project starting date	*01/11/2022*
Project duration	*45 months*
Rights	*PHOEBE consortium*

PM efforts per beneficiary that contributed to the deliverable

#	Partner	PM effort in the Deliverable
1	*iRAP*	*0.01*
2	*AIM*	*0.028*
3	*FLOOW*	*0.21*
4	*O7*	*0.02*
5	*NTUA*	*0.02*
6	*TUM*	*0.01*
7	*UPV*	*0.01*
Total	*PHOEBE Consortium*	*0.308*

Document History

Version	Date	Beneficiary	Description
0.1	*6/03/2025*	*Sam Chapman, FLOOW* *Mark Burke, FLOOW*	*Outline of the Deliverable structure and draft*
0.2	*15/04/2025*	*Sam Chapman, FLOOW* *Mark Burke, FLOOW* *Marcel Sala, AIMSUN* *Apostolos Ziakopoulos, NTUA* *Shanna Lucchesi, iRAP* *Arun Putatunda, TUM*	*Refined prefinal version with review of AI section parts from project partners and internal review of all before other partner review submission.*
0.3	*16/04/2025*	*Sam Chapman, FLOOW*	*Prefinal version with fixed citations and references for external in project review*
1.0	*28/04/2025*	*Sam Chapman, FLOOW* *Haris Sideris*, O7 *Marcel Sala*, AIM	*Final version with corrections from internal reviewers*

Project summary

The EU-funded 'Predictive Approaches for Safer Urban Environment' (PHOEBE) project aims to develop an integrated, dynamic human-centred predictive safety assessment framework in urban areas. This will be achieved by bringing together the interdisciplinary power of traffic simulation, road safety assessment, human behaviour, mode shift and induced demand modelling and new and emerging mobility data.

Focused on vulnerable road users' safety, the 3.5-year-long PHOEBE project will draw inspiration from real-world scenarios in the three pilot cities of Athens (GR), Valencia (ES) and West Midlands (UK). Testing activities will be performed across the use cases to simulate and forecast the impact of changes on safety in different scenarios of disruptions or transitions across urban transport networks.

Predicting and visualising the safety and socioeconomic outcomes of new forms of transport, new technologies, or regulatory and behavioural changes from the individual (micro) level up to the network wide (macro) level will also be a significant game-changer for urban stakeholders. The results of PHOEBE can be used as a blueprint by other European cities to develop their knowledge products, such as socioeconomic analysis model, urban road safety assessment, human behaviour and choice modelling.

PHOEBE pilot cities

List of participating cities:

- Athens (Greece)
- Valencia (Spain)
- West Midlands (United Kingdom)

Social Links:

https://twitter.com/Project_PHOEBE

https://www.linkedin.com/company/phoebe-project/

https://www.youtube.com/@phoebeproject

For further information please visit **WWW.PHOEBE-PROJECT.EU**

Project Partners

Organisation	Country	Abbreviation
EVROPSKI INSTITUT ZA OCENJEVANJE CEST - EURORAP	SI	**EIRA**
ETHNICON METSOVION POLYTECHNION	EL	**NTUA**
TECHNISCHE UNIVERSITEIT DELFT	NL	**TUD**
TECHNISCHE UNIVERSITAET MUENCHEN	DE	**TUM**
AIMSUN SLU	ES	**AIM**
POLIS AISBL	BE	**POLIS**
FACTUAL CONSULTING SL	ES	**FC**
UNIVERSITAT POLITECNICA DE VALENCIA	ES	**UPV**
OSEVEN SINGLE MEMBER PRIVATE COMPANY	EL	**OSEVEN**
THE FLOOW LIMITED	UK	**FLOOW**
INTERNATIONAL ROAD ASSESSMENT PROGRAMME	UK	**iRAP**

List of definitions, glossary and abbreviations.

Term	Meaning
AADF / AADT	Average Annual Daily Flow / Traffic- average number of vehicles passing a specific point on a road network each day over a full year.
ADT	Average Daily Traffic – average number of vehicles passing a fixed point per day.
AI	Artificial Intelligence
ATC	Automated Traffic Count surveys.
CONSUMERS	A stakeholder with a need for data or information to support work or framework usage in the PHOEBE project
EDPB	European Data Protection Board – body that advises on common interpretation of GDPR.
FAIR	Findable, Accessible, Interoperable, Reusable (FAIR data principals)
FCD	Floating Car Data - telematics data from vehicles
FEATURES	The feature inputs required for the PHOEBE framework parts to operate. Each feature may be informed by one or more sources of data.
GOLD STANDARD	Gold standard is a representative exemplar for data quality or accuracy in an area to which other data or accuracy may be compared.
IP / IPR	Intellectual Property / Intellectual Property Rights
KSI / FSI	Killed and Seriously Injured / Fatal and Seriously Injured
META DATA	A set of data that describes and gives information about other data, such as that presented in the D2.2 registers.
ODD	Operational Design Domain - the specific operating conditions and environmental factors under which an automated driving system or feature is designed to function safely
OD	Origin Destination – matrix data that tracks the frequency of movement from one location (the "origin") to another (the "destination").
PHOEBE	Predictive approacHes fOr safEr urBan Environments
PII	Personally Identifiable Information
PRODUCERS	A stakeholder with potential to supply data or information to support work or framework usage in the PHOEBE project
RAP	Road Assessment Program and the models to which support these
RTC	Road Traffic Collision - relates to data utilised to monitor road traffic incidents
SATC	Static Automated Traffic Counter - fixed point vehicle counting sensors and related data about locations
SSM	Surrogate Safety Measures - proxy indicators of risk that can be used in road safety analyses to quantify unsafe traffic events and event frequency
VRU	Vulnerable Road User
WP	Work Package

1 Table of Contents

List of tables

List of figures

Deliverable executive summary

The PHOEBE Framework is a methodological approach designed for cities to improve understanding of the safety implications of future changes in the transport systems, such as behavioural changes, redesign or new infrastructure, or evolving modes of travel. To do so, this requires evidential data and proven approaches. This document details the use of data in the PHOEBE framework. This is supported by:

1. An account of the purpose of this document.
2. A review of the project data including:
    a. A review of prior project data registers and the need for public facing registers.
    b. A documented summary of the *D2.2* live document public registers related to data. This includes:
        i. Data resource and FAIR register, detailing the aim and structure of this resource.
        ii. Features framework register, detailing the aim and structure of this resource.
3. A review of AI usage in the project, including:
    a. Summary of management controls and processes for AI management in the project.
    b. Extensive details on the usage of AI and its management for compliance.
4. The use of accreditation approaches to ensure new used data provides ensured accuracy and impact value.

Ultimately this report summarises from a data management perspective the use of data, AI and accreditation in the PHOEBE project.

2 Purpose of the deliverable

2.1 Attainment of the objectives

This document aims to provide commentary to accompany the data registers, which collectively form *D2.2*. This commentary accompanies registers pertaining to existing data sources, new data sources, and AI and novel data-capture approaches. This document also describes the iRAP data accreditation approaches when used to support the PHOEBE framework.

This deliverable meets the objectives as laid out in the original work plan for the PHOEBE project, providing data registers supporting the project needs.

These objectives include:

1. Providing a public conclusion and report concerning the project data needs, requirements and registers. This report aims to inform a wider audience about the projects data needs.
2. An overview of features data as used within the PHOEBE framework and detailed in a dedicated features register.
3. Providing a discussion and conclusion of how project data needs are supported by new data sources, new AI approaches, and the accreditation process.

Finally, this document discusses a dedicated public data register, which consists of two parts. These are:

1. Data resources and FAIR combined register
2. Features framework register

Both collated registers remain ‘live’ documents and are subject to change should new data be identified in the remainder of the project work.

2.2 Intended audience

The intended audience of this deliverable are the PHOEBE project partners, project officers, selected use case stakeholders, and for interested parties in the wider research community. This publication includes the registers themselves and this accompanying deliverable report.

As this deliverable is a public-facing document it should be noted that it may restrict any sensitive information related to named data owners or protected aspects about either data or features. Such restrictions help to support data privacy and protections in line with the PHOEBE data management plan (*Deliverable D7.3*) and the rights of project participants required exploitation and proprietary restrictions.

2.3 Links with other work packages and deliverables

Data underpins the scientific investigation and model refinement throughout the PHOEBE project. Links to all work packages are summarised in Table 1.

WP#	WP Name	Connection with D2.2
1	PHOEBE framework - Methodological and technical approach	Requirements and state of the art regarding data usage and classifications, supporting the methodological and technical approaches in: **Tasks 1.1, 1.2,** and **1.4** supporting **D1.1** and **D1.2**.
2	Data specification and collection (AI & machine learning)	**D2.2** forms a key deliverable of **WP2** activity however it remains closely tied to wider **WP2** tasks related to data and registers related to these. In particular, **D2.2** extends from deliverable **D2.1**. although is written to not be dependent upon it.
3	Model development and enhancement for VRU and urban safety	Supporting model development and testing with appropriate data: **Tasks 3.1, 3.2, 3.3,** and **3.4** supporting **D3.1** and ultimately **D3.2**.
4	Safety use case implementation	Supporting data discovery and needs in relation to use case activities, stakeholders and scenarios: **Task 4.1, 4.2, 4.3, 4.4** and **4.5** supporting **D4.1** and ultimately **D4.2**.
5	Systems integration and transferability	Standardisation and preparation of data and its requirements will be needed to support uniform model processing and evaluation in the use cases and can support: **Tasks 5.1, 5.3,** and **5.4** supporting **D5.1** and **D5.2**.
6	Communication, dissemination and exploitation	Within external communication data or outputs will be required to support messaging. Later in the project data management for releases will be required. **WP6** tasks support **D6.1** and ultimately **D6.2**.
7	Project Management	Within project management three key areas connect to data and the scope of **D2.2**. These are: **Task 7.3** Risk management - data privacy and data handling risks are detailed in the project risk tables. **Task 7.4** Data management - where data registers capture information on FAIR principles, IP protections and restrictions that can apply to project data. This gathered information is held and tracked in the LIVE **WP2** data registers and supports **D7.2**. **Task 7.5** Ethical management - where data gathering and potential usage requires ethical review to confirm ethical standards are met for gathering and usage of data. These reviews and confirmations are held in the ethical management registers supporting **D7.3**.

Table 1 - The connection from activity documented in D2.2 to other work packages, tasks and deliverables.

3 PHOEBE project data

The EU-funded 'Predictive Approaches for Safer Urban Environment' (PHOEBE) project aims to develop an integrated, human-centred predictive safety assessment framework suitable for changing urban areas. This is achieved by bringing new data, approaches and models together in investigative development and a new framework. The PHOEBE framework brings together traffic simulation, road safety assessment, human behaviour, mode shift, and induced demand modelling approaches to holistically understand the impact of new infrastructure and policy. This complex process requires robust data management to ensure best reuse of existing data, consistent data usage, and replicability. Data management is facilitated by maintaining a portfolio of data registers, aiming to manage and control data needs across the project to help best support, accuracy, portability and impact from the resulting framework.

3.1 Introduction to project data registers

Data registers are often component parts of data and investigative management in data and analytic projects. Registers are used to help collate complex data requirements and constraints whilst helping to support data related tasks that can manage and minimise data complexity.

3.1.1 The need for PHOEBE project data registers

Data registers have been structured to help provide a framework for data needs and usage following three data management areas (we term this the "three Cs):

- **Connect** – to help discover and make data available that may be needed using a coherent approach
- **Comprehend** – to understand the type, content and value of data as may be used within the framework
- **Control** – to help support data usage encouraging accuracy, consistency, replicability of the framework system and reuse of its data

These areas of data management are further broken down in how each area supports more specific data related tasks supporting the PHOEBE framework. These are defined in more detail in the following Table 2.

Area of data management (the 3 C's)	Data tasks that registers help to support	How this task area helps to support the PHOEBE project and its aims.
Connect – make available data needed	Discovery	Supporting the data discovery process for available data that may support urban risk estimation.
	Collection	Enabling the collection and targeted gathering of data that may support urban risk estimation.
	Generation	Supporting the generation of specific data that helps to address data gaps supporting urban risk estimation.
Comprehend – to understand the potential and use of data	Cataloguing	Creating common taxonomy and catalogues of data to group common features and sources of data to allow common understanding.
	Mapping	Supporting potential mapping between 'data providers' and data consumers'.
	Purpose	Supporting understanding of the goals of using data to ensure data is fit for purpose.
	Accuracy	Supporting understanding of data accuracy and its integrity to ensure the best resource is universally used where able.
	Minimisation	Supporting consolidation of data sources to select best available sources when multiple may exist.
	Coverage	To understand the temporal and geospatial coverage of data to ensure fit for purpose.
	Storage	To understand the location of data to understand its availability for a given need.
Control - to support data usage for the framework system	Accountability	To manage ownership, provenance, and who uses data in the framework system.
	Protection	To manage data privacy processing controls and compliance aspects of data management.
	Access	To manage the operational access of data for use in the framework.
	Reuse	To manage potential data reuse beyond original purpose to support onward reuse of resources.
	Coordination	To manage coordination of data resources used between systems parts, partners and tools.

Table 2 2 – Areas of data management (the 3 Cs) and the data tasks that they help to support.

Overall, the use of registers aid the PHOEBE framework data management areas and the underlying data tasks to help handle and minimise complexity.

3.1.2 A review of PHOEBE data requirement and availability registers

Throughout the PHOEBE project research and development activities, a range of initial registers have helped to support the connection, comprehension and control of data of relevance to the project. These prior registers are detailed in Table 3 below.

PHOEBE supporting register (D2.1)	Register data task and aims within PHOEBE	Overview of the prior data register (D2.1)
Consumer data needs register	Discovery Purpose	This deprecated register aimed to capture the initial needs of 'data consumers' to discover urban risk model data needs and requirements.
Producer data needs register	Discovery Accountability Daat Coverage Data Storage	This deprecated register aimed to discover potential data available across potential data owners that could be related to the project or its use cases.
Types and classifications register	Cataloguing Data Minimisation	This classification register aimed to group similar types of data to a taxonomy and common description across data stakeholders.
Consolidated availability requirements register	Discovery Cataloguing Mapping Data Minimisation	This set of deprecated registers combined data supplier and data consumer needs across all possible suppliers and consumers to data classifications and taxonomies.
Initial gap analysis register	Discovery Cataloguing Mapping Data Minimisation Data Gathering	This deprecated register merged data producer and data consumer data to determine gaps in available data. This supported data minimisation, mapping and early stages of gathering of data for project purposes.
Use case experimental design registers	Purpose Cataloguing Data Coverage Accuracy Accountability Data Protection Data Reuse	This deprecated register collated data to ensure it was fit for the project different use case need. It captured data related to coverage, accuracy, protection and accountability to support reuse in and beyond the project aims.
Consolidation registers	Gathering Cataloguing Mapping	This deprecated register mapped data required for use cases and the project development to source data using a revised taxonomy related to data sources.
Secondary gap analysis registers	Data Gathering Cataloguing Mapping	This secondary gap analysis created a series of deprecated registers detailing the areas without data supply where data gathering would be required.
Served data registers	Cataloguing Data Collection Data Coverage Data Access Accountability	This now deprecated series of registers (one per use case) catalogued available data for each use case supporting data access for use case needs of known available data.
Unserved data registers	Cataloguing Gathering	This now deprecated register (one per framework model sub-component) collated data requirement gaps to support minimised targeted data gathering.
Outputs and intermediate data register	Cataloguing Accountability Data Protection Data Access Data Reuse Coordination	This now deprecated register collated additional information about data produced within the project or by the process of parts of the PHOEBE framework. It focuses upon data protection, attributions, accountability, access and reuse to support FAIR principles for data access.

PHOEBE supporting register (D2.1)	Register data task and aims within PHOEBE	Overview of the prior data register (D2.1)
Publication of FAIR data register	Cataloguing Accountability Data Protection Data Access Data Reuse Coordination	This live register is an internal project register related to publication and scientific work related to the PHOEBE project. This register is principally related to WP6 to manage and support dissemination requirements of the wider project. This register however also provides details to better encourage data reuse including data related to project scientific works and publications. More details on publications are included in wider *D6.1 Communication, dissemination and exploitation plan*.

Table 3 - table of prior D2.1 supporting data registers used in the PHOEBE project and the data tasks that each supported in overall data management

3.1.3 The need for public data registers

To ensure beneficial research outputs the PHOEBE project follows a range of good practice approaches to ensure dissemination and knowledge sharing. These approaches encourage open access publications, dissemination activities and conform to FAIR principles throughout the project activity. These efforts are reported wider in *WP6* (*D6.1 Communication, dissemination and exploitation plan*) and regarding data management in the deliverable *D7.3 Data management plan*. Beyond these efforts the contents of data registers themselves need consideration to ensure following FAIR principles.

Public data registers provide the following benefits:

- Enabling data discoverability providing suitable metadata.
- Improving accessibility of existing and new data sources.
- Promoting data interoperability and where possible standardisation.
- Support research beyond the funded project.
- Encourages collaboration and citation leading to scientific impact from project work.

4 D2.2 Data Registers

This document is coupled with a dedicated public data register comprised of two main component parts. These are:

1. Data resources and FAIR combined register
2. Features framework register

These are each now detailed in sections 4.1 and 4.2.

4.1 Data resources and FAIR combined register

4.1.1 Aim

This register supports open dissemination of project related data following the four FAIR principles (Wilkinson, M. D. 2016). These guiding principles are documented in deliverable *D7.3 Data Management Plan.* In summary, the FAIR (Findable, Accessible, Interoperable, Reusable) principles are:

1. **Findable** – Data should be easy to find by having consistent, standardised, and machine-readable metadata.
2. **Accessible** – The data should be accessible by anyone with a computer and an internet connection. Where data is sensitive in nature then any access conditions should be clearly detailed.
3. **Interoperable** – Data should use standard formats and use consistent metadata to help enable integration with other datasets and existing systems and software.
4. **Reusable** – Data should be well-documented, with clear licensing and provenance information, so others can use it effectively.

In addition to these FAIR principles this register also aims to link data and ethical management reviews in relation to each data source as well as document its linkage and usage in the PHOEBE project.

4.1.2 Structure

The structure of the *D2.2* register for data resources is organised to detail key metadata about data utilised within the project helping to support FAIR principles. The structure aims to be minimal to best support external understanding, enabling onward usage, exploration and filtering. To support filtering a range of constrained metadata fields has been included to filter resources and help exploration for aiding onward discovery and potential reuse of project utilised resources.

It should be recalled that a much longer list of potential data was considered in *D2.1* earlier in the project however this longer list contained data ultimately discarded. The following D2.2 data register should therefore be intended to identify only the validated data sources used in project activity. This approach can help direct data consumers and researchers more easily away from ultimately less useful or problematic resources.

The structure of the data resources and fair register are detailed in the following Table 4.

Fields in DATA resources and FAIR	Decsription of the field and its contents
Data ID	Unique identifier for data used in the project.
Data title	Textual Description of the data.
Appears in register	Prior linkage to D2.1 internal registers (project usage only).
Data type	Type of Data in relation to the project.
Data Owner	Data ownership in relation to the project.
Data provider to the project	Project partner responsible for the data resource used within the project.
Use Cases covered	Use cases where data is applicable to allow filtering.
Known data VS expected data	If data is 'known' and available or 'expected' data that may not yet be available (e.g. final project KPIs).
Part of project dataset (T2.1.2)	Link to managed data areas as declared in D7.3 Data Management Plan and its related records.
Relation to workpackage(s)	Relationship to specific workpackages in the PHOEBE project to allow filtering by parts of the project.
Ethical review identifier	Ethical review identifiers that pertain to the data usage within the project to ensure meeting the requirements of ISO 26000 for ethical handling. (Each identifier relates to an ethical check documented in D7.4 Ethical Management Plan and its related registers).
DPIA ID	Data privacy Impact Assessment (DPIA) records related to this data and its usage in the project data processing. Each record relates to records held in the project register 'GDPR and AI processing register' related to D7.3. Thjis register incorporates all data privacy impact assessments for project data and processing.
Brief description of data	A textual description of the data resource to enable understanding of its content and potential usage.
Location of Data	Location or means of access to data used in the project to enable reuse of prior and new data resources used in project work. This may be a URL or can be a contact point for requested access if given a need to request access to the resource for onward usage. PLEASE NOTE not all data is available for onward reuse to protect propriatry data that the project may of utilised OR for partners to protect potential exploitation of data involved in the project work.
FAIR Rating	An indicator reviewing the FAIR staus of the data resource and its overall potential for sharing and reuse.
FAIR Reasoning	This details any reasons behind and OPEN or restricted designation for FAIR sharing of data.
Data Size Estimation	This is an upper estimate of data size bands to help understand the size of the resource should a user wish to obtain the data for direct or indirect reuse.
Temporal extent	Information regarding the temporal extent of the data resource to understand the range of real world measurement, representation or capture that the data may contain.
Geographical extent	The geographical coverage extent detailing the coverage of the available resource.
Current Accessibility	Accessibility filter by means of accessibility of the data resource.
Current Interoperability	Interoperability filter to select by means of interoperability readyness from the data sources.
License for reuse	License that applies for potential reuse of the data beyond the project.
Keywords for reusability - to aid findability	Keywords about the data resource to allow finding and fast understanding of its content.

Table 4 - Fields in the D2.2 Data resources and FAIR register.

To facilitate reuse and following FAIR principles constrained fields help to allow filtering of the project data. Firstly, key terms classify data into existing stakeholder and regional data sources or data that is newly

gathered, annotated, generated via AI or model outputs or as a part of scientific investigative works. This is detailed below in Figure 1.

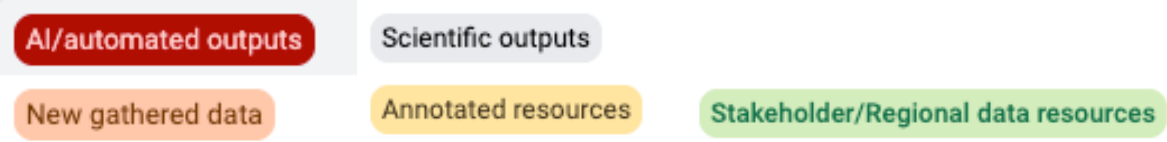


Figure 1 – Data Type classifications key terms allowing filtering to data resource types in relation to the PHOEBE project

Like data types aiding findability a range of themed keywords also help to organise data and allow quick discovery of potentially interesting data resources. Keywords used to focus upon PHOEBE related data are detailed in Figure 2 and with distributions in Figure 3.

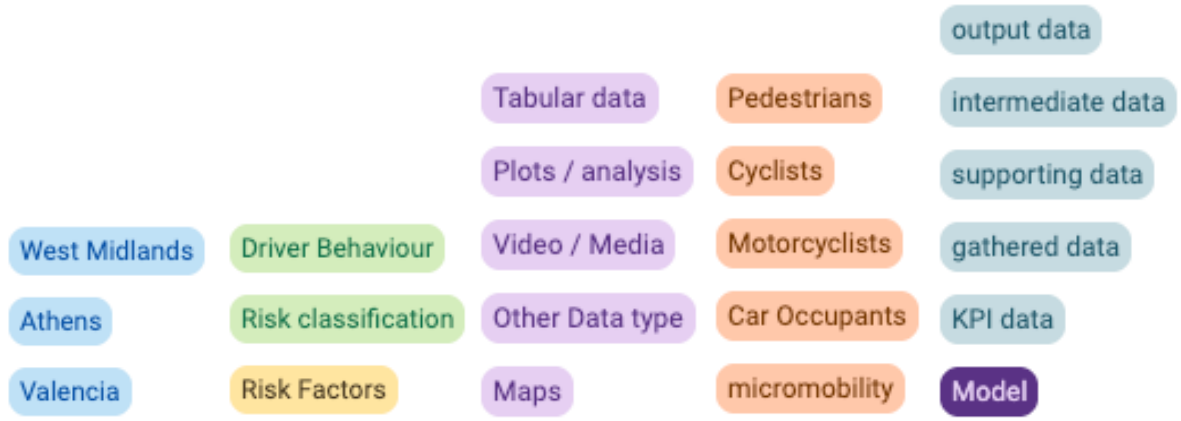


Figure 2 - Keyword filters (multi-selection) allowing filtering to data in relation to the PHOEBE project to aid FAIR findability. The coloured areas indicate: Geographic, Types of data content, Data formats, Mode of travel, Period of data covered, Model

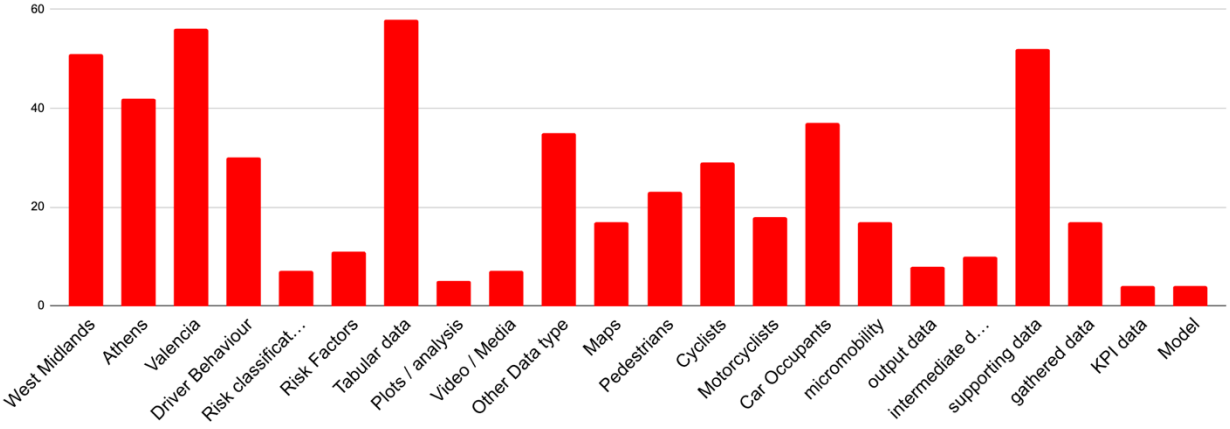


Figure 3 - Keyword distributions frequencies across all meta data keywords. Keywords support findability [accurate as of 1 APR 2025]

To allow filtering upon owners and providers key terms are used to help filtering data by project partners or when appropriate external stakeholders. These selections are detailed in Figure 4.

Figure 4 - Identifiers to allow filtering upon data in relation to project partners and providers as well as managed data ownership for each area.

To support onward usage the project identifies data by the terms related to its permitted reuse. This refers to its license when declared. Data will be under an open license where possible to support. However in cases of third-party data reuse, proprietary data or formats or if data is not distributed to support exploitation, this will be indicated. This record of permitted reuse supports a central FAIR principal to disclose licensing meta data to facilitate potential reuse, the key terms supporting this are detailed in Figure 5 and Figure 6.

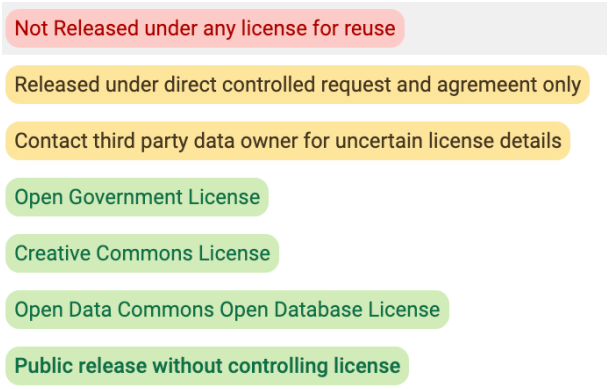


Figure 5 - License for reuse – to allow controlled release of data in relation to the project aims

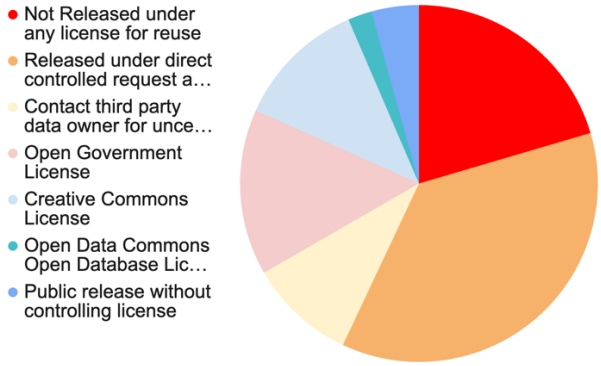


Figure 6 - Percentage distributions of assigned licenses in project utilised data. [accurate as of 3rd APR 2025].

To aid potential interoperability metadata is also included to detail aspects of its reusability. Ideally all shared data should be presented in common reusable formats whenever possible. However, in some cases particularly with complex or proprietary data formats this may not be the case. This filter allows understanding of interoperability as can be seen in Figure 7 and Figure 8.

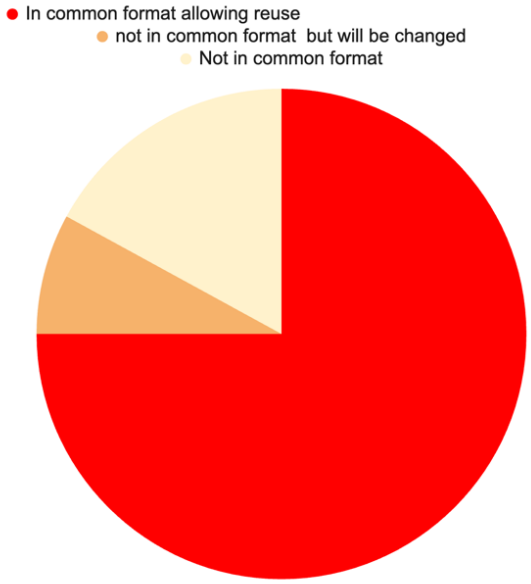


Figure 7 - Percentage plots of data by the interoperable meta data assigned on them showing approximately 75% having common reusable formats [data accurate as of 1APR2025].

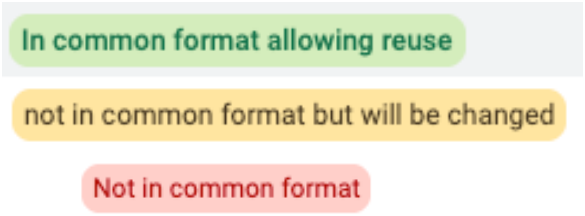


Figure 8 - Interoperability key terms allowing identification of resources in interoperable formats for possible reuse

To facilitate reuse of key term meta data, filters exist to help address accessibility of resources to understand the means to gain access to them. These key terms are detailed in Figure 9 and distributions of these are shown in Figure 10.

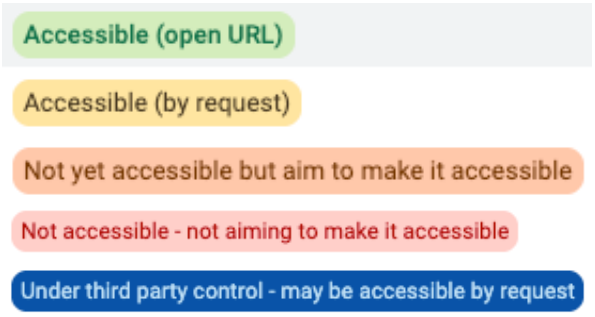


Figure 9 - Accessibility key terms allowing filtering of available resources for reuse addressing aims of FAIR data governance.

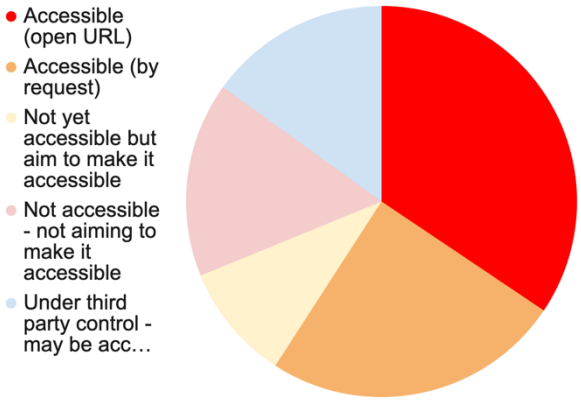


Figure 10 - Percentage plots of data accessibility from meta data with over 50% of data accessible [data accurate as of 1APR2025].

To facilitate the understanding of data resources banded size estimates are used to indicate resource sizes of available data to aid reusability. These are detailed in Figure 11.

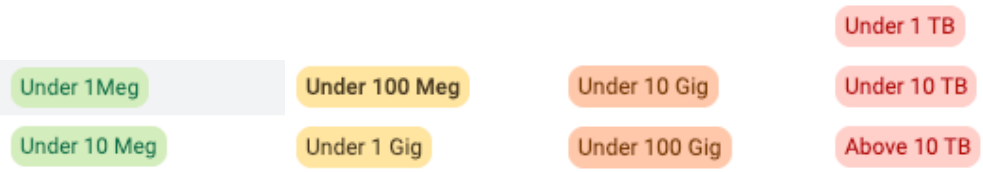


Figure 11 - Data Size estimate key terms allowing filtering of resources by the estimated banded size of each data resource to aid reusability.

Finally, to allow a geographical focus data of coverage or relevance to each use case is also indicated, where data can apply to more than one region. This metadata allows filtering particularly to understand potential regional reuse potential in line with the explored use cases. These are detailed in Figure 12.

Figure 12 - Use case key terms covered to allow filtering of data resources used in differing use cases.

Ultimately, across all features more than 90 distinct utilised data sources are detailed in the register. It should be noted that the register (unlike this document) is a live document that is subject to change as new data may be added over time. This is still likely until the project end as project progress may add new research outputs into available data[1]. A full example of a singular record is presented as an example of a singular meta data record within the register in Table 5. The entire register for data resources and FAIR consists of several thousand metadata.

[1] This will include new updates from work in WP5 in particular which extends beyond WP2 where this register will be maintained.

Fields in DATA resources and FAIR	Example singular data record
Data ID	D43
Data title	Valencia - speed limits
Appears in register	PRODUCER Data records Information-Data-Producers
Data type	Stakeholder/Regional data resources
Data Owner	EXTERNAL STAKEHOLDERS
Data provider to the project	UPV
Use Cases covered	Valencia
Known data VS expected data	Known already
Part of project dataset (T2.1.2)	DMPDS4 - Collated Data from the use case regions and stakeholders
Relation to workpackage(s)	WP2
Ethical review identifier	E8 E9
DPIA ID	DPIA 5 DPIA 6 DPIA 7
Brief description of data	Georeferenced database of speed limit for each street
Location of Data	https://valencia.opendatasoft.com/explore/dataset/velocitat-carrers-velocidad-calles/export/ AND https://valencia.opendatasoft.com/explore/dataset/velocitat-carrers-velocidad-calles/information/?location=16,39.50133,-0.37493&basemap=e4bf90
FAIR Rating	Already FAIR
FAIR Reasoning	Used in project model configuration and enhancements
Data Size Estimation	Under 100 Meg
Temporal extent	current only
Geographical extent	Valencia
Current Accessibility	Accessible (open URL)
Current Interoperability	In common format allowing reuse
License for reuse	Creative Commons License
Keywords for reusability - to aid findability	Valencia Tabular data Driver Behaviour Car Occupants supporting data

Table 5 - An example singular record in the Data Resources and FAIR register.

4.2 Features framework register

4.2.1 Aim

The features framework register details the individual features needed within the various parts of the PHOEBE framework and the data origins for each related to the use cases. This register supports FAIR principles (as per section 4.1.1). In this register, each feature is mapped to the data used for the use cases.

4.2.2 Structure

The structure of the *D2.2* register for features used in the PHOEBE framework details key terms and some metadata about the features and the origin data for them. The live document aims to be minimal making it easy to follow, update and review. Within the register a range of key terms are present allowing exploration and filtering for internal and onward usage. The structure of the features framework register is detailed in the following Table 6.

Fields in features FRAMEWORK register	Decsription of the field and its contents
FeatureID	Unique identifier for features used in the project.
Features	Textual Description of the feature.
Athens Source Data ID(s)	A list of origin data sources used to support the need within the PHOEBE framework within the ATHENS use case
West midlands Source Data ID(s)	A list of origin data sources used to support the need within the PHOEBE framework within the WEST MIDLANDS use case
Valencia Source Data ID(s)	A list of origin data sources used to support the need within the PHOEBE framework within the VALENCIA use case
Use-Case	Indicator of within which use case(s) the data is used (from WEST MIDLANDS, ATHENS and VALENICA)
Supporting models	Indicator of the specific models parts in the PHOEBE framework that need this feature
Use in Setup stage	Indicator of the SETUP phase usage in WP3 when the feature is utilised.
Use in Simulation stage	Indicator of the SIMULATION phase usage in WP3 when the feature is utilised.
Use in Refinement stage	Indicator of the REFINEMENT phase usage in WP3 when the feature is utilised.
NOTES	Specific notes (1) about the feature and its usage in the PHOEBE framework. This may point to external documentation of the feature or notes of restrictions that may apply in the PHOEBE framework usage.
NOTES 2	Specific notes (2) about the feature and its usage in the PHOEBE framework. This may point to external documentation of the feature or notes of restrictions that may apply in the PHOEBE framework usage.

Table 6 - Fields in the D2.2 Features FRAMEWORK register.

This register maps data IDs from the Data resources and FAIR register to data used in the PHOEBE framework. In each of these areas these links to data utilised are detailed by the data IDs shown in Figure 13. Some of the utilised data is used to support multiple features. For instance, '*D2 - iRAP geolocated*

survey images' are utilised across a wide range of features where survey imaging is used to support encoding of a large range of observed features.

etc.

Figure 13 - Feature usage of differing data linked by a set of Data ID's (see section 4.1).

Many features are required across each use case region however some are targeted specifically to support specific setup or scenarios in just some of the use cases. Use case key terms support faster exploration and filter by the use cases that features are applied to Figure 14.

Figure 14 - Use cases regions covered to allow filtering of features used in differing use cases.

Figure 15 shows that similar numbers of features are applicable in each use case (Athens has 109, Valencia 127 and the West Midlands uses 141).

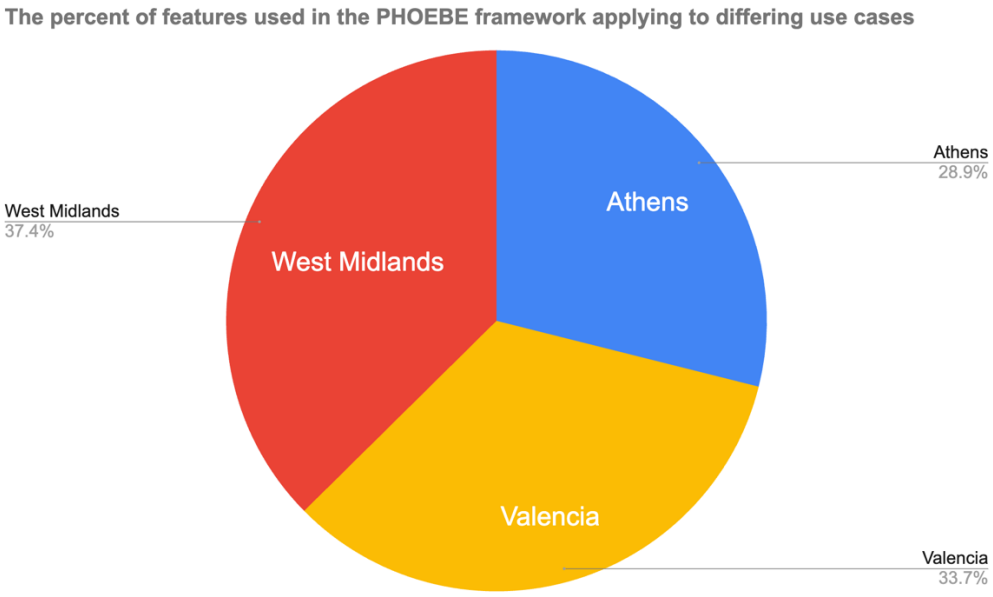


Figure 15 - the distribution of features used in each use case in the PHOEBE framework showing a similar distribution of feature numbers for each use case [data accurate as of 15 APR 2025].

Using both the Data IDs and the use case regions the distribution of data usage across the use case features can be visualised in Figure 16. This visualisation highlights the strong and central role of certain highly utilised data sources, and the supportive role often aligned to individual use cases and scenarios of others. Some data sources are not used in the PHOEBE framework for its setup and usage, these show as zero usage the operating PHOEBE framework. These data sources instead have supported investigatory, academic and developmental work even if not utilised in the operating PHOEBE framework.

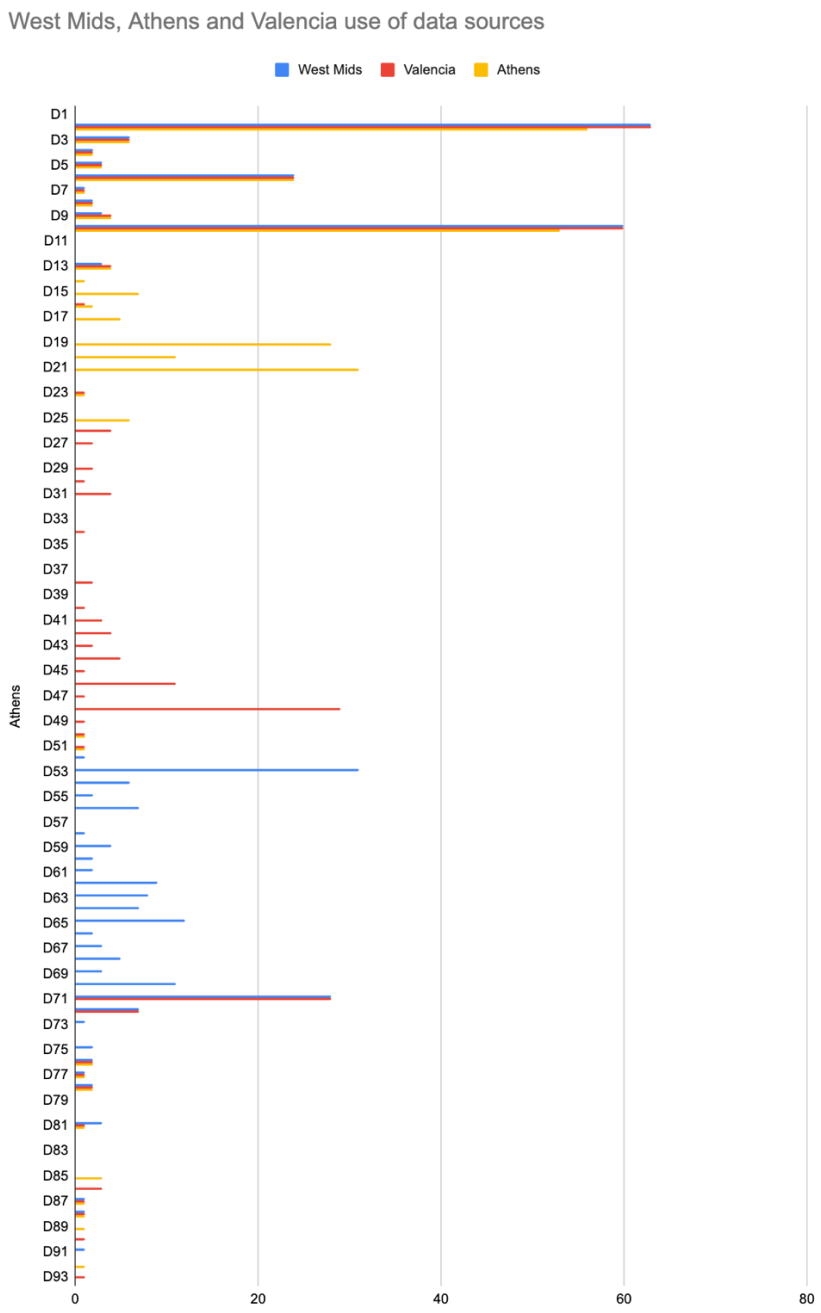


Figure 16 - The frequency of data source usage to support needed model features per use case (counted by the features that they support). [data accurate as of 15 APR 2025]

Features are also encoded across the component parts of the PHOEBE framework to help support filtering and onward usage of separated parts outside of the overall framework if used in isolation. These key terms aid discovery for differing parts of the framework to see where features are utilised. The key terms for model parts of the framework are detailed in Figure 17.

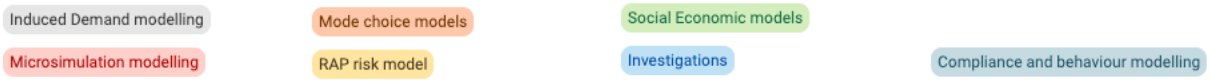


Figure 17 - Key terms for the specific models parts in the PHOEBE framework that utilise this feature.

The distribution of features applied to differing model parts is shown in Figure 18. This highlights the large number of features in RAP models to help encode risk features of the network.

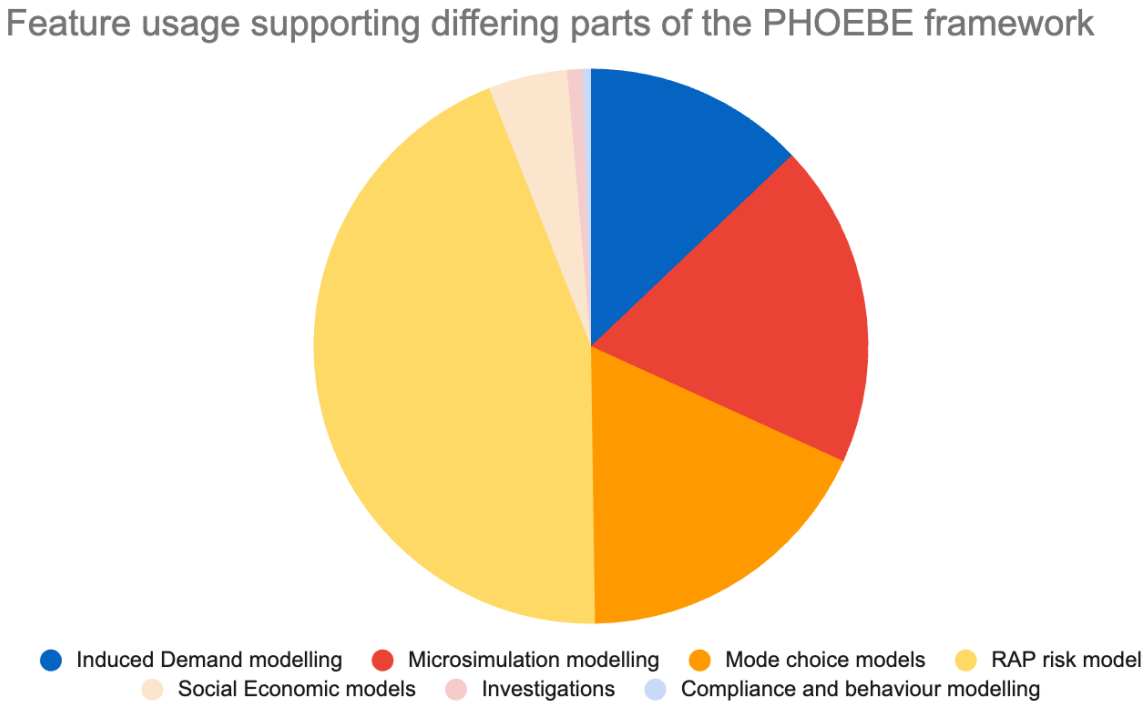


Figure 18 - Feature usage levels by proportion across the differing parts of the PHOEBE framework. [data accurate as of 15 APR 2025]

Within the PHOEBE framework three stages of processing exist:

- **SETUP** – the stage where initial configuration is supported with data enabling model configuration for the application region.
- **SIMULATION** – the stage where features support the operation of simulation and modelled outputs.
- **REFINEMENT** – the stage where features help improve outputs to ultimately produce KPIs for the overall PHOEBE framework.

Within each of these stages features: may not be required, may be required for the first time, updated, or simply reused without alteration. These stages utilised key terms detailed in Figure 19.

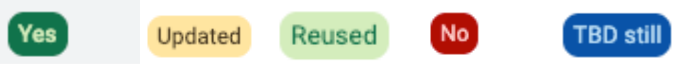


Figure 19 - Filter indicators of feature usage at the SETUP, SIMULATION and REFINEMENT stages of models in the PHOEBE framework.

Overall the collective metadata and key terms enable to see the full range of data utilised in the PHOEBE framework. A singular example record for a feature detailing potential fixed obstacles at network segments is shown in Table 7.

Fields in features FRAMEWORK register	Example singular data record
FeatureID	F40
Features	Fixed obstacles
Athens Source Data ID(s)	D2 D10 D19 D21 D6
West midlands Source Data ID(s)	D2 D10 D71 D53 D6
Valencia Source Data ID(s)	D2 D10 D71 D48 D6
Use-Case	Athens Valencia West Midlands
Supporting models	RAP risk model Microsimulation modelling
Use in Setup stage	Yes
Use in Simulation stage	Reused
Use in Refinement stage	Reused
NOTES	Encoded data detailing this data field - encoded types are detailed in https://resources.irap.org/Specifications/iRAP_Coding_Manual_Drive_on_Left.pdf
NOTES 2	N/A

Table 7 - An example singular record in the Features FRAMEWORK register. In this example the encoded feature comes from a variety of potential data sources, these differ by use case. These values support two models but only in the setup phase.

5 AI usage in PHOEBE

The PHOEBE project utilises Artificial Intelligence (AI) tools and technologies in some areas. This section of the report details the management and use of AI in the activities carried out by project partners and use cases. For the purpose of this document AI is defined as:

Artificial intelligence (AI) is a broad set of technologies that enable computers to perform a variety of advanced functions on variable inputs, including the ability to see, understand, predict, analyse data, estimate outcomes or are used to automatically inform models or analysis used in decision making.

5.1 Management controls for AI

AI tools and methods have many potential societal benefits, particularly in terms of efficiency and the automation of tasks. However, there are undoubtedly risks from utilising the technology. Users should consider the potential for producing inaccurate results, violating privacy, introducing statistical biases, and the potential impact from AI decision making on humans. To manage these potential risks, controls for AI processing have been introduced in the project to ensure AI is used responsibly. These controls extend on good practice principles aligned to regulation, such as the General Data Protection Regulation Act (EU 2016/679) and the EU AI Act (EU 2024/1689). AI processing management include augmenting the GDPR article 30 processing registers with a dedicated list of all AI usage and risk reviews. For each AI task, the types of data processed are recorded, potential impacts of processing are considered, and we detail the extent to which risks are mitigated. The AI Register is a live document that may be updated throughout the project life, this can be seen in Figure 20.

PHOEBE AI Register

OWNER FLOOW LAST UPDATE 18-Feb-25

DETAILS

This register describes AI or AI like systems that may be used in the project work to ensure record keeping and review of AI systems. This includes approaches that provide prediction, estimation, models, automated analysis or aother areas that may or not include AI algorithms. This register aims to capture risk levels and any needed mitigations for any such system, program or tool usage to access where it may have any impact upon individuals, the environment or wider

AI REGISTER CAVEAT

Despite this register each partner is seperately has its own responsibilities under the EU AI ACT and EU DATA ACT that may pertain to AI management. This register creates a record indicating ownership of AI system and project based risk reviews and mitigations. The AI system owner and responsible party may have additional details about AI systems beyond this register in its own compiance activities. Each partner is seperately accountable for AI system usage and any risks thsi may bring.

AI SYSTEM OWNER AND RESPONSIBLE PARTY(S)	AI SYSTEM AND USAGE	TYPES OF PERSONAL DATA *	POTENTIAL IMPACTS OF AI PROCESSING IMPACTING INDIVIDUALS / ENVIRONMENT / SOCIETY	RISK LEVEL FROM AI IMPACTS	SPECIFIC MITIGATION MEASURES FOR AI PROCESSING OR ITS OUTPUTS
IRAP	AI meeting notes (Microsoft Teams Pro) - used in meetings to securely capture records of some project meetings for use strictly within the project coordination, running and record keeping.	PII - names and connected informations about individuals in connection to the PHOEBE project.	Automated approaches may capture in part or full additional details about meeting participants mentioned within meetings that can be captured in meeting records. All meeting records are used within the confines of the collaboration agremeent and the project contract including its terms for protection and confidentiality.	LOW	Outputs are reviewed by humans and used only within the project activity subject to confidentiality.

Figure 20 - The PHOEBE AI Register which captures and supports management of project AI activities.

5.2 AI usage in the PHOEBE project

For each instance of AI usage, the use of personal information is considered, impacts from processing are reviewed, and risk assessments are detailed. The use of AI processing is detailed below.

5.2.1 Assistive tools

Standard AI tools have been utilised to aid productivity by making day-to-day efficiencies. One example is the use of AI meeting assistants and minute taking. Such tools are not used in all meetings but have helped in capturing records of some project meetings. These records are held securely in the project for internal use only. While these tools capture details about the participants they are regarded as low risk in terms of the subsequent processing performed and are only used with the consent of all attendees. To allow the use

of such tools the details of processing and data retention have been checked to examine risk and ensure no third-party data reuse.

5.2.2 International Road Assessment Program Risk model

A key part of the PHOEBE project is the use and extension of road risk models developed by the international road assessment programme (iRAP). Such iRAP models have been developed considering many studies carried out by both academics and safety practitioners[2]. These models assess road safety by analysing road features and assigning them risk values, which are then used to estimate the potential numbers of fatal and serious injuries (FSI's) at a given location and to recommend interventions to reduce those risks. These models require a comprehensive parameterisation of a given road and the surrounding infrastructure into more than 50 key features, the majority of which is obtained by performing a manual survey along a route.

Within the PHOEBE project to better fit the needs of changing urban environments these models have been extended and improved. Improvements include model integration and incorporation of non-compliance behaviours, induced demand, and mode shift components. The altered risk assessments are directly aligned with the project's goal of reducing fatalities and severe injuries on urban roads. The approach also supports the standardisation of road safety evaluation utilising a fixed methodology and approved data (see section 6 on accreditation). The enhancements to the methodology provided within PHOEBE facilitate the incorporation of new data sources, such as telematics data for vehicle behaviours, section 5.2.6. These data sources improve efficiency by minimising manual data collection and enhances the reliability and detail of the results. The risk assessment relies on various data sources to estimate the risk and occurrence of fatal and severe injuries per road segment. The required data includes infrastructure-related information about parts of the road network. In PHOEBE, data sources encompass survey assessment videos, telematics data, traffic counts, and historical incident records. Traditional iRAP assessments rely on periodic site surveys, however, new enhancements allow connection with traffic simulations tools (also detailed in this section) to support continuous monitoring and dynamic risk analysis. These new data sources and AI approaches have been used to refine iRAP's Star Rating methodology, ensuring that road safety insights are more accurate and data driven. The enhancements to the risk assessment component of the PHOEBE project will be made available to all iRAP stakeholders worldwide.

Throughout the use of RAP models all data utilised is fully anonymised without any personal identifiable information. Outputs are reviewed by humans before any changes are recommended that could positively impact fundamental human safety or the built environment. Model outputs can influence change programs on road networks; however, this is performed without the use of personalised data and encourages positive safety changes to the public realm improving the fundamental right to safety. This use of AI use is regarded as a low risk and is in the public benefit.

5.2.3 Microsimulation models

The PHOEBE framework combines models to predict traffic behaviours in changing urban environments, which in turn influence risk. One key component part of the overall PHOEBE framework that supports this modelling are traffic microsimulation models. The PHOEBE project utilises AIMSUN microsimulations, which include several related AI capabilities. AIMSUN can model realistic mobility behaviours over a road network under different conditions. Specific models have been developed for each use case region in the PHOEBE project. Within the microsimulation models several component AI models are used, these include:

[2] https://irap.org/methodology/ For details of the methodology and reference materials. [link - last accessed 7th APR 2025]

- **A vehicle following model** – controls the acceleration and breaking of vehicle agents in the simulation model to emulate realistic behaviours of vehicles[3]. This model helps understand risk behaviours of modelled vehicles to help understand changing locational risk. To support PHOEBE needs this model has been configured to each of the use case scenarios using data of road-user behaviours.
- **A lane changing model** - simulates when vehicles may realistically change lanes[4]. To support PHOEBE needs the configuration has been adapted to emulate traffic weaving and lane-change behaviours within urban multi-lane environments.
- **A path finding algorithm** – helps to identify best paths that vehicles may take between desired origin and destination regions[5]. Beyond configuration for each of the urban settings and mode-based passage restrictions the project also added specific new cost functions to support the modelling of autonomous vehicles (AVs). These specific model additions restrict agent operation by implementing operational design domain (ODD) constraints in the West Midlands use case. These additions allow the modelling of new vehicle types and their impact on road risk.
- **A network-loading algorithm** - performs traffic assignment onto the modelled network that ensures a realistic usage of the different routes under changing scenarios[6]. These aspects have been configured in PHOEBE using project data to best simulate realistic traffic behaviour under changing urban scenarios. These aspects also support wider model integration within the PHOEBE framework to allow network assignment to be influenced by other behaviour models (mode shift or induced demand, for instance). **N**ew specific cost functions have been created to support modelling of the change of the network load and behaviours.
- **A social force model** – enables to control a simulation of pedestrian mobility in modelled regions[7]. This is configured using project related data on pedestrian mobility.

No data containing PII is used to configure, validate, and simulate the AIMSUN models. Model validity is tested in the use case regions before use by comparing simulation outputs with measured quantities. Ultimately, model outputs are reviewed by humans before any changes are made that could impact fundamental human safety or the built environment. This use of AI use is regarded as a low risk and is in the public benefit.

5.2.4 Use case and scenario microsimulation models

In addition to the general changes to the operation of the Aimsun software, the use case leaders configure the model for each region. Each use case model is detailed with specific review of data usage and changes to best represent mobility and the needs of the use case scenarios.

[3] The prior vehicle following model and its possible configuration is detailed in the following public documentation. https://docs.aimsun.com/next/24.0.2/UsersManual/MicrosimulationModellingVehicleMovement.html#car-following-model [LINK LAST ACCESSED 7 APR 2025].

[4] The prior lane changing model and its possible configuration is detailed in the following public documentation. https://docs.aimsun.com/next/24.0.2/UsersManual/MicrosimulationModellingVehicleMovement.html#lane-changing-model [LINK LAST ACCESSED 7 APR 2025].

[5] The route choice model is further documented in the following public documentation. https://docs.aimsun.com/next/24.0.2/UsersManual/StochasticRouteChoice.html [LAST ACCESSED 7 APR 2025].

[6] The network node assignment used in the microsimulation model is detailed in the following public documentation. https://docs.aimsun.com/next/24.0.2/UsersManual/StaticScenarios.html [LAST ACCESSED 7 APR 2025].

[7] The Pedestrian simulator is detailed in the following public documentation. https://docs.aimsun.com/next/24.0.2/UsersManual/PedestrianSimulator.html [LAST ACCESSED 7 APR 2025].

Athens (NTUA) – The model has been enhanced using Athens Traffic Management Centre data and manually analysed video footage to ensure predictive validity for scenarios being evaluated in the central Athens area.

Valencia (UPV) – The model has been configured and validated using a large variety of data sources. These include OD matrices, micromobility counts obtained through field observations, motorised vehicle counts provided by the Valencia City Council, and lane-specific speed records obtained from video analysis on the most representative road segments. Additionally, field measurements of traffic signal cycles at key intersections have been conducted and incorporated into the model. All of this has been undertaken to fine-tune and validate the model for the use case.

West Midlands (FLOOW) – The West Midlands model is comprised of two separate models covering an extensive contiguous region. The separate focus of two models allows for the fine tuning of each to support the study of different scenarios within each area. The models have been configured using extensive data including count and speed measurements across a large range of sources and modes. The Floow provided speeds, traffic volume estimates, turning frequencies, acceleration behaviours, excessive speeding behaviours, and OD matrices. Other data included infrastructure signal data, bus data (from the UK Bus Open Data Service), and mode counts from Vivacity[8] cameras, All Traffic Count (ATC) surveys to align and validate the model for the use case.

Each model was configured to generate realistic traffic simulations providing detailed data on mobility behaviours and network performance. It contributes to PHOEBE methodologies for evaluating urban mobility scenarios. Ultimately, the data derived from the model usage is used to help assess the potential impact of different interventions in each region and to calculate the KPIs.

This use of Aimson models uses only pre-anonymised or non-PII data and are regarded as LOW risk with outputs being in the public benefit without detrimental impacts to citizens.

5.2.5 Behavioural models

To provide dynamic understanding of urban risk changes the PHOEBE framework incorporates a set of behavioural models. These models inform the framework as to when behaviour change can impact risk. The incorporated behavioural models cover specific areas where influences are expected to alter risk in urban settings, each of these are detailed below.

Survey data collected from each of the three use cases have been used to develop and validate a mode choice model tailored for the specific regions. This model relies on a system of equations that process data specific to each use case, estimating how mode choice behaviours change under different circumstances. The survey datasets are used to estimate the parameters (or sensitivity) of specific predictors, e.g., Travel

[8] Vivacity cameras (https://vivacitylabs.com/products/smart-traffic-monitoring-solution/) data outputs have been made available for reuse in the project.

Mode choice

cost, time, and risk. These equations and the parameters are unique to each use case. Mode choice models are used to generate probability of travelling by a particular transport mode for a specific scenario and the potential for modal shift given a particular policy or infrastructural change. These use case specific models play crucial roles in estimating the mode choice and modal shift under regulatory interventions which can influence citizens' transportation mode choices and shift in choices of the given region, thereby affecting the overall risk profile associated with the transportation of that region. These models, developed by the Technical University of Munich, take the unsegmented demand matrices, origin-destination skim matrices and point estimates (predictor values, e.g., average speed, cost of travel of modes) and then segment the unsegmented demand matrices into the demand matrices associated with specific modes of transport. These segmented matrices are further used as inputs to the microsimulation models.

Induced demand

An induced demand model allows the project partners to understand where new demand may be generated and changed following other changes to the scenario. These induced demand models are informed from targeted surveys from which model parameters are determined per use case. As with the mode choice models, induced demand modelling utilises microsimulation OD matrices and point estimates to update mode-specific OD data as conditions change due to interventions.

Non Compliance

Particularly pertinent for the understanding of risk is the propensity for individuals to fail to comply with the local laws or expected traffic behaviours. Models of non-compliant behaviours in PHOEBE include a set of risk-related behavioural aspects that specifically impact risk in urban settings. These aspects are used to modify micro-simulation models and then feed into risk outputs using modelled insights a that align to surrogate safety measures (SSMs). These models have made extensive use of non-compliance related surveys and have validation of the modelling approaches derived using use case measurement and camera-based data. The non-compliance models focus on a range of different risky behaviours, including:

1) Vehicle speeding
2) Pedestrian red light running
3) Pedestrian crossing outside designated areas
4) Micromobility red-light running

Across these behavioural models used within the PHOEBE framework all data use is pre-anonymised or anonymous at source although in some validation aspects data may have originated from PII camera or mobility data although has already been anonymised. Overall, the models do not use personal or protected data. The AI is regarded as LOW risk with outputs being in the public benefit without detrimental impacts to citizens.

5.2.6 Telematics Data gathering and usage

Telematics data (floating car data) can be derived from a variety of means including directly from connected vehicles, aftermarket vehicle trackers (e.g. On-Board Diagnostics – OBDs), or smartphone sensors. Within the PHOEBE project the main source of data originates from smartphone sensors. Regardless of the device each of these technologies captures telemetry from GNSS, in the form of speed and position, and data from other sensors that might be present, such as accelerometers or magnetometers. Collectively, these data represent individualised naturalistic driving records. Positional data could contain PII and therefore are considered sensitive and are protected. Telematics data undergo anonymisation and aggregation to be used in the project, and with the consent of the data controller. These actions ensure that individual privacy

is maintained. Aggregated forms of such telematics data can provide privacy-safe insights into how roads are being used to help monitor and improve infrastructure.

For PHOEBE and road safety usage, the value of anonymised and aggregated telematics data lies in its ability to go beyond the limitations of traditional transport data, which often relies on fixed-point sensors. While traffic counters or speed cameras provide information only at isolated locations, telematics offers a continuous view of driver behaviour across the entire road network providing valuable new data for models, research, and transport stakeholders and validation. Thus, these data can provide new insights into behaviour over an entire road network, not just where monitoring infrastructure is already in place. For telematics data to be of value in PHOEBE modelling it remains essential to ensure representative sampling to reduce the potential for introducing biases.

When telematics data are obtained from fitted devices or connected vehicles the journey end points correspond well with the position of the vehicle, as the start- and endpoints of the journey can be defined from the ignition of the vehicle. However, in the case of data obtained from smartphones, the journey start and endpoints are not as clearly defined and must be inferred algorithmically. Telematics companies go to great lengths to ensure that in-car devices and smartphones produce similar results, so that their users receive the same quality of service regardless of the means of data collection. To support this task AI is utilised to identify the start and endpoints of journeys recorded by smartphone to ensure that high-quality data are available for the project. These systems operate on sensitive personal data requiring proprietary solutions to maintain data privacy, these systems are summarised below:

Within the Athens use case a proprietary trip recording mechanism[9] was configured and utilised that automatically identifies the start and end of trips. This mechanism uses the smartphone's AI-based activity recognition features to understand classified mobility events when a trip begins and when it ends and whether the user is a driver or a passenger. Trip recordings whose accuracy is low due to sensors noise are automatically rejected. Outputs enable trip records which are used for a variety of aggregated driving behaviour metrics supporting PHOEBE mobility models. The anonymised data resulting from processing this data are used in all modelling areas for the Athens use case.

Within the West Midlands use case a proprietary journey identification approach was used[10]. This process seeks to minimise the volume of transferred data. This approach used a collection of proprietary classifiers (phone signal, cell tower data, positional data, altimeter, gyroscope and other device sensor data as may be available) each configured for different device types and sensors to determine when a journey is taking place. This process takes place within the smartphone and helps to ensure quality secure data provision for the project usage without unnecessary data transfer and helps to produce a clean sample of motor vehicle behaviour from the eventual aggregated datasets. These data are aggregated to provide a statistical understanding of traffic properties by locations. These data support all models in the West Midlands use case, configurations, scientific investigative work, validations, and scenario investigations.

[9] This utilised a part of O7's proprietary technologies. To ensure data accuracy the project data was accredited via a validation process detailed in section 6.

[10] This utilised a part of The Floow's proprietary technologies, to ensure data minimisation and data accuracy to known modes. The outputs of aggregation are supported via accreditation as detailed in section (REF).

It should be noted that the Valencia use case does not use telematics data. The flexibility of the PHOEBE framework means that different data can be incorporated into the process depending on availability or requirements.

The methods outlined above involve the use of sensitive and protected data, necessitating specific handling procedures, checks, and considerations. The European Data Protection Board (EDPB) provides guidance on identifying high-risk data processing (WP248rev01), with particular emphasis on the processing of personal data in the context of connected vehicles and mobility-related applications (EDPB Guidance 01/2020). This guidance indicates that the processing of vehicle telemetry data may constitute high-risk processing, thereby requiring the implementation of specific controls and mitigation measures. The processing methods described above support these requirements by offering mechanisms to minimise extensive data collection, thereby contributing to compliance of GDPR[11].

Furthermore, appropriate controls stipulate that data processing must be conducted in a secure and protected environment, without any unauthorised transfer or exposure of personal information. To this end, calculations are carried out within protected proprietary systems with restricted and secure access. The processing also incorporates automated anonymisation and aggregation, enabling the generation of outputs suitable for sharing beyond the secure proprietary environment.

A final risk assessment concludes that the overall risk and potential impact are considered low, due to the implementation of strict controls and the removal of personally identifiable information (PII), alongside the fact that data is used strictly within the domain of road safety applications as a consented, permitted and expected use for mobility trajectory data.

5.2.7 Video automated analysis approaches

During PHOEBE project requirement gathering and data availability reviews, areas where the research needs of the project were not served by a specific data source were identified. Data gaps included statistics for non-compliance activities (see REF above). To support the needed model development in these areas new data collection was required. To observe some non-compliance activities the PHOEBE researchers chose to use camera-based installations to gather enough data over time for the project needs. For data protection and privacy concerns camera-based installations even in the public realm are considered high risk and require specific legal, privacy, ethical and authority checks and approvals. Dedicated ethical and data privacy reviews[12] were carried out for the Athens and Valencia use cases. For the West Midlands use case, access to anonymised data outputs from existing stereo camera installations was provided by the local stakeholders.

Within data privacy impact assessments and ethical reviews, it was essential to control automated processing activities upon gathered video feed data to minimise privacy risk. Restrictions included no processing of faces or vehicle number plates that may enable identification of recorded citizens within any processing. Ultimately, automated processing was targeted to understand locational behaviours without reference to individuals. Output data would be anonymous and with controlled processing and human review of outputs in secure handling of the video data. These controls mean that the risk was judged to be acceptable and in the public interest given its role in advancing our understanding of safety.

[11] General Data Protection Regulation article 5(1)(c) is supported by these efforts using technical approaches to support data minimisation.

[12] More about legal, data protection and ethical reviews is covered in wider deliverables *D7.4 Ethics Management plan*, *D7.3 Data Management Plan* and various registers and records related to these.

The AI processing of video feeds identified transport modes and the tracking of real-world positions. An object identification algorithm was used building on academic work from (Redmon et al, 2016) using the identification algorithm YOLO[13]. This algorithm is configured to identify transport modes within each frame. This base algorithm was combined with DeepSORT (Simple Online and Realtime Tracking) using approaches from (Wojke et al, 2017). This process identified the telemetry of objects in the observed scene with separate configuration per camera based upon its position and orientation. The separately configured approaches were used as follows in each use case.

For the Athens use case, the system was adapted to detect specific risk indicators, such as unsafe crossings made by pedestrians or cycles, near-miss events, and some metrics for indicative user behaviours. Resulting datasets were used in refining non-compliance models for the red-light violations and providing new insights on traffic conflicts in Athens.

A customized computer vision pipeline was deployed to monitor pedestrian and vehicle interactions at several key intersections with previously recorded pedestrian incidents and vehicle violations. Video data were collected from street-level perspectives using smartphones for two weeks, including peak and off-peak hours.

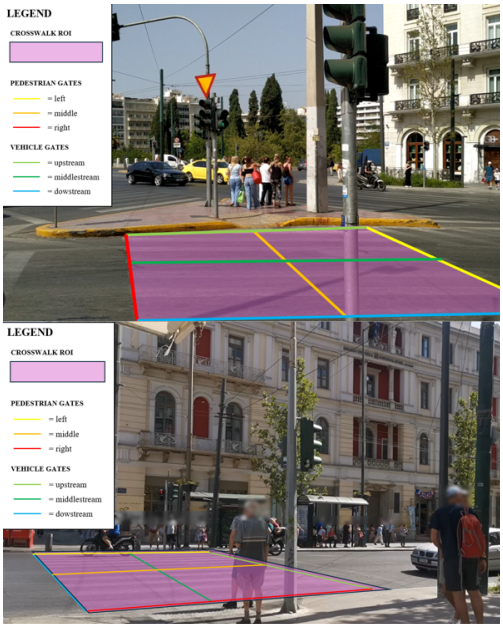


Figure 21 - The use of the custom computer vision pipeline using YOLO and DeepSORT as configured for two specific crossings in the Athens use case area

YOLOv8 was implemented in the system for real-time vehicle and pedestrian detection. A frame-by-frame identification and classification of objects involved in road interactions was achieved. Kalman filters, a conventional method applied in multi-object tracking, were employed by the system to make position predictions and resolve temporary occlusions of detected objects. For further object classification with greater precision, ResNet-50 was also implemented for feature extraction so that the system could maintain proper identity tracking despite difficult urban settings with intersecting movements.

As a result of employing ground-level cameras, homography transforms were employed to map perspective views to a top-down coordinate space. This facilitated the system's ability to produce reliable movement trajectories and distance-related metrics. To reduce noise within trajectory data due to camera shake or occlusions, Savitzky–Golay filters were employed. The filters preserved the important movement dynamics while improving the quality of trajectory data employed in behavioural analysis.

By defining Regions of Interest (ROIs) around crosswalks and intersections, the system monitored compliance with traffic signals.

[13] YOLO (You Only Look Once) used v10, specifically the nano ("n") model, this has the fewest parameters and is therefore the fastest enabling handling of higher resolution video.

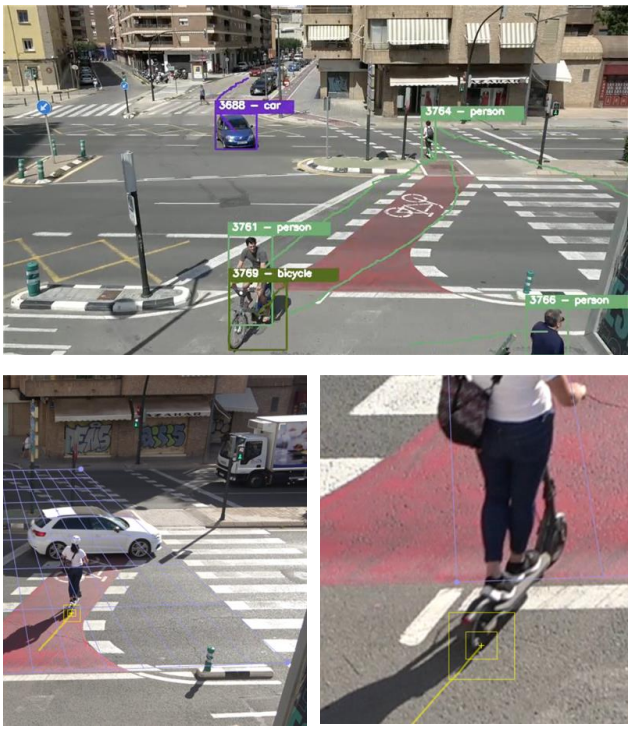


Figure 22 - The use of YOLO and KINOVA as configured for crossings in the Valencia use case area

In the Valencia use case, the approaches were extended further to provide more accurate data specifically for micromobility users. This added the use of the KINOVEA open-source software (Charmant, J. 2009) to track the movements of micromobility that was otherwise not well tracked. This approach has been used to identify potential conflicts at three intersections between those micromobility users who run red lights and motorised vehicles. Once identified, the mode-based objects their trajectories have been reconstructed with KINOVEA. These trajectories have been used in the development of behaviour models for compliance with red lights by micromobility users, as well as in the study of traffic conflicts in these situations.

6 Accreditation in PHOEBE

Within the PHOEBE framework a key component model is the use of a road safety assessment model that calculates Fatal and Serious Injury (FSI) estimates based upon locational characteristics and encoded conditions for parts of the mobility network. The models, which are developed by the International Road Assessment Programme (iRAP) are deployed globally using a singular risk methodology and data specification.

The PHOEBE project is pioneering the use of novel data sources as inputs for these models. To support usage of new data sources, and to ensure these sources are properly understood and checked for compliance with iRAP data specifications, a review process with accreditation is used called *AiRAP attribute accreditation*[14]

This process is being used in the PHOEBE project to ensure the quality of telemetry data being used for vehicle speeds and flows. The accreditation process verifies the data source is meets reliability standards, and ensures that the data is processed in such a way that meets iRAP's data definitions.

Accreditation maximises the potential impact from work done in the project to utilise this type of data in the iRAP models, meaning that the same data source can be readily used across Europe for the same purpose.

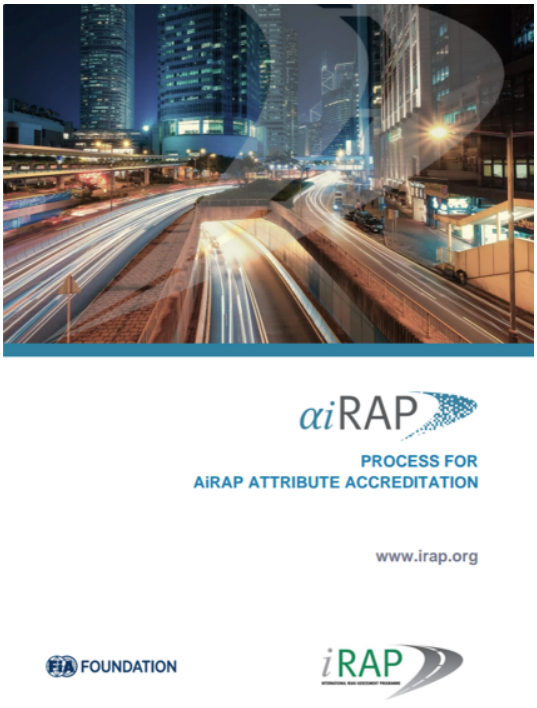


Figure 23 - Documented process for aiRAP attribute accreditation that was used within new data accreditations to help validate the accuracy and validity of new data in RAP model usage.

In PHOEBE, novel data sources includes telematics information sourced from two project partners: The Floow and OSeven).

Telematics data offers distinct advantages, providing granular insights into road user behaviour across the entire network. This data can be used to encode locational features relating to operational speeds and vehicle flow volumes, both of which are critical inputs to road crash risk calculations. As such, rigorous due diligence is required to ensure the integrity and representativeness of these inputs.

The accreditation process has several stages:

- **STAGE ONE – Application**. An application form is submitted outlining the attributes to be accredited, the data origin, control measures, and the responsible applicant.
- **STAGE TWO – Data conversion**. Data is converted into formats suitable for iRAP model inputs. Training is also undertaken to ensure the data meets defined quality standards.

[14] More details about AiRAP accreditation and related processes is detailed at https://irap.org/accreditation/ [LAST ACCESSED 13 APR 2025]

These four stages of the accreditation process are shown in more detail in the following process map, Figure 24.

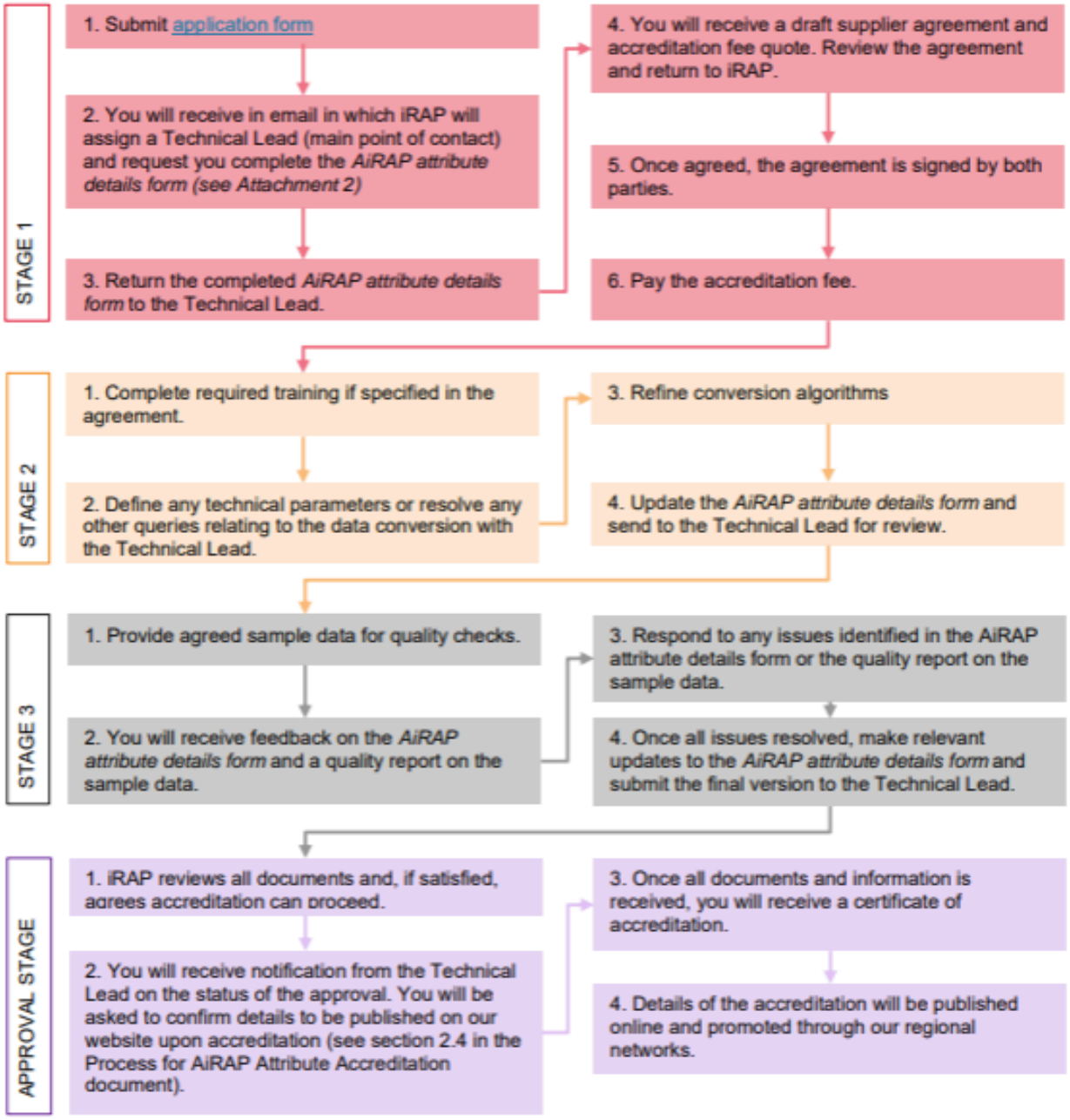
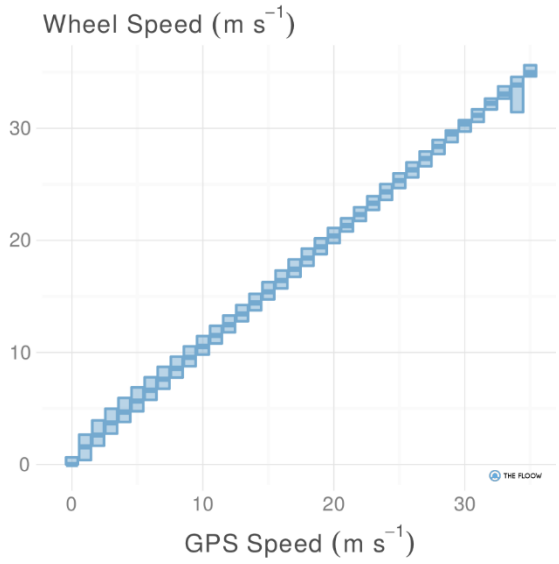


Figure 24 - the AiRAP attribute accreditation process map used to add new data into RAP usage in the PHOEBE project (please note stage 1 step 6 was not followed in project related work)



During the accreditation process, alternative data sources were used to validate both input attributes and risk model outputs. For example, GPS-derived speed data was compared with vehicle wheel speed data collected via dynamometers to confirm its accuracy. Figure 25 below demonstrates strong consistency between the telematics data and benchmark measurements.

Figure 25 - Part of data used in accreditation processes comparing GPS speeds to vehicle dynamometer wheel speed methodologies, this shows strong consistency of speed data to gold standard measured speeds. This utilised analysis from a prior autonomous research project MOVE_UK

As a result of this process, telematics data on operational speeds and traffic flows has successfully achieved AiRAP accreditation. This supports the robustness of the PHOEBE project outputs and increases the potential for future use and scaling of these data sources in other applications.

7 Bibliography

Charmont, J. 2009. KINOVEA open source software, https://kinovea.org

EDPB Guidance 01/2020. Guidelines 01/2020 on processing personal data in the context of connected vehicles and mobility related applications, https://www.edpb.europa.eu/system/files/2021-03/edpb_guidelines_202001_connected_vehicles_v2.0_adopted_en.pdf

EU 2016/679. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

EU 2024/1689. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Redmon, J. et al., 2016. You Only Look Once: Unified, Real-Time Object Detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788.

Wilkinson, M. D. 2016. The FAIR Guiding principles for scientific data management and stewardship. Scientific data 3.

Wojke, N., Bewley, A., & Paulus, D. 2017. Simple Online and Realtime Tracking with a Deep Association Metric. 2017 IEEE International Conference on Image Processing (ICIP), 3645-3649.

WP248rev01 2017. European Union Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is likely to result in high risk for the purposes of Regulation 2016/679 wp248rev0.1 https://ec.europa.eu/newsroom/document.cfm?doc_id=47711

8 Appendix

This document is supported by several combined live registers as previously detailed in sections REF and REF. These registers can be viewed publicly and are made available at:
https://docs.google.com/spreadsheets/d/1XFC36RPV7XWeDY-wwT7MdisFUU4w85aqqqJ8iaFpgBg/edit?usp=sharing

Please note this document will also be shared and linked on the project website https://phoebe-project.eu/ before the project's conclusion.

UK participants in Horizon Europe Project PHOEBE are supported by UKRI grant numbers 10038897 (The International Road Assessment Programme – iRAP) and 10056912

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement

The sole responsibility for the content of this document lies with the authors. It does not necessarily reflect the opinion of the European Union. Neither CINEA nor the European Commission are responsible for any use that may be made of the information contained therein.